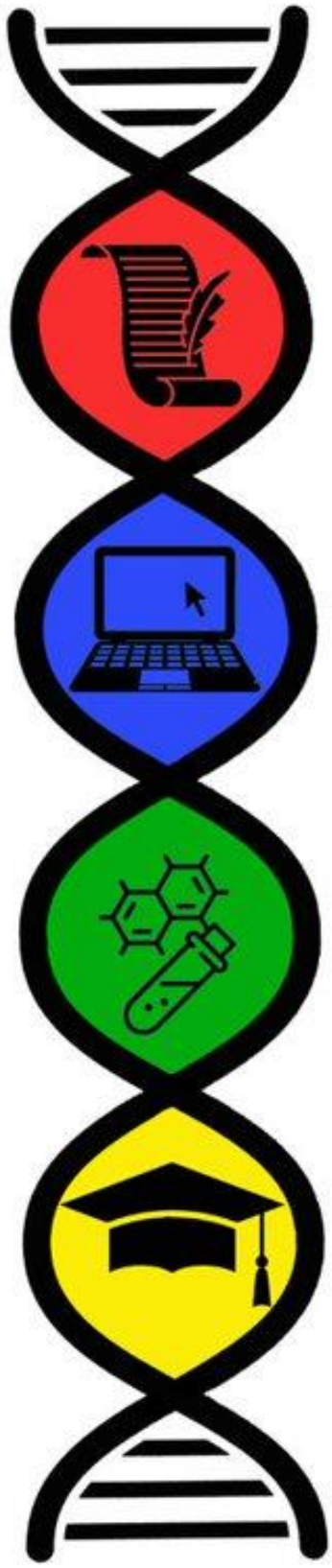

STANDARDS & GUIDELINES:

GENERATION & ANALYSIS OF HIGH THROUGHPUT SEQUENCING DATA

FOR APPLICATION TO AUSTRALIAN BIOSECURITY DIAGNOSTIC LABORATORIES

Version 1 March 2024

AUTHORS

Dr. Jane Oakey

Senior Principal Molecular Biologist
Biosecurity Queensland
Queensland Department of Agriculture and Fisheries
jane.oakey@daf.qld.gov.au

Dr. Daniel Bogema

Research Scientist
Biosecurity and Food Safety
New South Wales Department of Primary Industries
daniel.bogema@dpi.nsw.gov.au

Dr. Monica Kehoe

Plant Virologist and Molecular Plant Pathologist
Biosecurity and Sustainability
Western Australia Department of Primary Industries and Regional Development
monica.Kehoe@dpird.wa.gov.au

Sam Hair

Laboratory Scientist, Molecular Biology
Biosecurity and Sustainability
Western Australia Department of Primary Industries and Regional Development
sam.hair@dpird.wa.gov.au

The authors acknowledge contributions from **Dr. Mark Blacket**, Agriculture Victoria.

REVIEWING PANEL

Dr. Oliver Berry, CSIRO Environomics Future Science Platform

Ms Georgia Breckell, Ministry for Primary Industries, New Zealand

Dr. Barbara Brito Rodriguez, NSW Department of Primary Industries

Dr. Axel Colling, Australian Centre for Disease Preparedness, CSIRO

Dr. Andrew Daly, NSW Department of Primary Industries

Dr. Adrian Dinsdale, Australian Department of Agriculture, Fisheries and Forestry

Dr. Fiona Filardo, Queenland Department of Agriculture and Fisheries

Mrs Melinda Frost, NSW Department of Primary Industries

Dr. Wei Guo, Illumina Inc.

Dr. David Gopurenko, NSW Department of Primary Industries

Dr. Kim Halpin, Australian Centre for Disease Preparedness, CSIRO

Dr. Nichole Hammond, Pest Risk & Analytics, Plant Biosecurity, Department of Primary Industries & Regional Development (DPIRD), Western Australia

Dr. Tonny Kinene, Department of Primary Industries & Regional Development (DPIRD), Western Australia

Dr. Wycliff Kinoti, Agriculture Victoria

Dr. Tanya Laird, Department of Primary Industries & Regional Development (DPIRD), Western Australia

Dr. Dongmei Li, Ministry for Primary Industries, New Zealand

Dr. Lia Liefting, Ministry for Primary Industries, New Zealand

Dr. Stacey Lynch, Agriculture Victoria

Ms Sherralee Lukehurst. School Molecular and Life Sciences, Curtin University, Western Australia

Ms. Joanne Mackie, La Trobe University

Dr. Solomon Maina, NSW Department of Primary Industries

Dr. Matthew Neave, Australian Centre for Disease Preparedness, CSIRO

Ms Melanie O'keefe, Australian Genome Research Facility

Dr. Alexander M. Piper, Agriculture Victoria

Dr. Gareth Price, Head of Computational Biology, QCIF Bioinformatics

Dr. Asad Prodhan, Department of Primary Industries & Regional Development (DPIRD), Western Australia

Prof. Grant Rawling, Agriobio, Agriculture Victoria

Dr. Luciana Rigano, Ministry for Primary Industries, New Zealand

Dr. Katie Robinson, NSW Department of Primary Industries

Dr. Brendan Rodoni, Agribio, Agriculture Victoria

Ms. Bianca Rodrigues Jardim, La Trobe University

Neil Shepherd, National Association of Testing Authorities (NATA)

Dr. Craig Smith, Biosecurity Sciences Laboratory, Queensland Department of Agriculture and Fisheries

Dr. Jeremy Thompson, Ministry for Primary Industries, New Zealand

Dr. Lucy Tran-Nguyen, Plant Health Australia

Dr. Darren Underwood, Biosecurity Sciences Laboratory, Queensland Department of Agriculture and Fisheries

Dr. David Waite, Ministry for Primary Industries, New Zealand

Dr. John Webster, NSW Department of Primary Industries

Dr. Sarah Williamson, Birling Laboratory

Ms. Teresa Wilson, Biosecurity Tasmania

Dr. Johanna Wong, NSW Department of Primary Industries

Dr. William Wong, , Australian Department of Agriculture, Fisheries and Forestry

Australian Department of Agriculture, Fisheries and Forestry

Animal Health Laboratory, Ministry for Primary Industries, New Zealand

Subcommittee for Committee on Animal Health Laboratory Standards

CONTENTS

INTRODUCTION

High throughput sequencing (HTS) is a powerful technology that, in recent years, has revolutionised the study of biology, converting it from a data-poor to a data-rich science. The use of biological data to identify, characterise, and manage pests and diseases is a central goal of biosecurity, and has been greatly enhanced through use of HTS technology.

Judicious application of HTS in disease diagnostics allows the direct detection, investigation and characterisation of the genome sequences of various infectious agents, including elucidation of their role, relatedness and evolutionary aspects of infection biology and disease development. Furthermore, comparative sequencing studies on the genomes, exomes or transcriptomes of healthy and diseased cells and organisms provides more powerful diagnostic opportunities, as well as improved classification, forecasting and therapy selection for many infectious illnesses[2]. HTS has applications also in the examination of species' composition where massively parallel targeted loci can be sequenced simultaneously for detection, comparison of taxa or ecological studies. This can be applied equally to pest diagnostics either from specimens/parts of specimens or environmental samples. Early detection of many high priority pests and diseases is crucial to reducing the risk of entry into Australia and minimising the chance of spread if entry were to occur.

HTS is revolutionary for the Biosecurity arena, which requires sensitivity and specificity. However, HTS also is a highly disruptive technology and presents many challenges for routine adoption in biosecurity laboratories. The data-rich outputs produced using HTS present challenges to reporting and policy. HTS also requires significant upgrades to IT equipment and staff expertise. At the onset, laboratories must evaluate whether the adoption of HTS should include the substantial investment of infrastructure for both wet-lab and IT requirements, or whether to outsource to service providers for some, or all, of the process. The review article by Mintzer *et al* (2019)[11] may assist decision makers in this regard.

In 2018, the National Biosecurity Committee (NBC) recognised HTS as a priority to strategic surveillance while noting the requirement for policy and regulatory arrangements in its adoption. To address the smooth adoption of HTS by the Australian Biosecurity system the NBC commissioned a series of workshops to develop a HTS framework and implementation plan. Following the resulting recommendations, in February 2021 the HTS Implementation Plan Working Group (HTSIPWG), chaired by Dr. Brendan Rodoni, was formed and tasked with developing an implementation plan for the National Biosecurity HTS Framework. One of the recommendations of this plan was a working group of technical experts be established to draft the technical standards and guidelines that will support quality assurance for the generation and the analysis of HTS outputs for diagnostics relevant to

Australian biosecurity applications. To enable elements of accord between animal, plant and environmental biosecurity, the group was composed of highly skilled and experienced technical personnel covering wet-lab and data analysis fields, from multiple jurisdictions, and with animal and plant health experience. In May 2022, a draft of the technical standards and guidelines was widely reviewed by stakeholders. After consideration of reviewers comments, a final version 1.0 was endorsed by the Subcommittee on Animal Health Laboratory Standards (SCAHLS) and the Subcommittee on Plant Health Diagnostics (SPHD) and noted by the Animal Health Committee (AHC) and the Plant Health Committee (PHC) in February 2024. The Standards & Guidelines: generation & analysis of high throughput sequencing data for application to Australian biosecurity diagnostic laboratories, version 1, was endorsed by the NBC in March 2024 prior to publication.

Standardisation in the approach, application and interpretation of HTS is critical to ensure an element of commonality in the expectations, implementation, analysis, transparency, communication and understanding across Biosecurity jurisdictions. Many technical standards used for PCR tests are applicable to HTS, but some need to be revisited to account for variables that are not common to both techniques. In addition to the wet-lab processes, standard methods for analysis of the much larger and complicated datasets generated by HTS techniques are required for quality control and assurance of accurate, reproducible and consistent results.

The standards and guidelines have been collated using existing recommendations from multiple sources as baselines (either referenced directly or provided in the Resources section) from which a comprehensive single document is generated for application to the broader scope of Biosecurity laboratories. The document is directed at adoption of HTS techniques in all diagnostic aspects of Biosecurity and across all Australian jurisdictions. Moreover, and similar to the SCAHLS guidelines for nucleic acid testing[18] (PCR), this document may be used as assessment criteria for comparison with accreditation requirements. This document will form the baseline of future outputs from the HTSIPWG. Laboratories may wish to consider that forthcoming national policy regarding HTS usage in Biosecurity, national Biosecurity training programs, and the development of a high-quality Biosecurity database and bioinformatics platform, may all assume application of the technical standards described herein.

How to use this document

This document is not a methods manual. Laboratories must ensure they are familiar with the appropriate methodology to conduct HTS testing on their chosen platform. This document focusses predominantly upon Illumina and Oxford Nanopore Technologies (ONT) platforms as they are the more common platforms used by Australian Biosecurity laboratories (HTS National Capability in Biosecurity Survey conducted by Kim Halpin (ACDP -CSIRO), unpublished 2019). Laboratories using alternative HTS platforms must develop their own quality assurance to attain the same level of standards, and this must be fully documented with clear identification of any critical control points specific to their platform.

It is expected that this document to be used in conjunction with the laboratory's quality system.

To ensure commonality in elements of the process subject to local decisions and those which involve quality assurance critical control points, this document provides minimum standards that must be used. The document includes also a "guideline" level that exceeds the standard and is provided as a recommendation.

Laboratories may wish to prepare internal checklists or flow charts for HTS tests based upon the relevant parts of this extensive document.

This document is intended for ensuring confidence and commonality in *diagnostic applications*. It should be considered optional whether laboratories choose to apply these standards to their research and development activities. This document is not intended to limit the use of HTS in emergency scenarios such as disease/pest outbreaks where “off-label” or bespoke applications may need to be applied if standard applications are deemed unsuitable. Such use is considered research in the context of this document.

Glossary & Abbreviations used in this document

Adaptors	5' and 3' adapter sequences are ligated to fragments to be sequenced and include barcoding sequences, forward/reverse primers (for paired-end sequencing) and binding sequences/complexes for immobilizing the fragments for the sequencing platform to read
Amplicon sequencing	A method that targets specific genomic regions using PCR amplification. Amplicons from different samples or different primer sets can be multiplexed
Barcoding	DNA barcoding is a system for species identification focused on the short, standardised genetic regions acting as a "barcode". **Metabarcoding** is the technique of using HTS for barcoding multiple species in a sample simultaneously Some manufacturers use the term barcode as a synonym for index (see Index/Indexing below)
cDNA	Complementary DNA synthesised from single stranded RNA template by reverse transcription. RNA is required to be transcribed to cDNA for sequencing
Coding region	Also known as the coding DNA sequence (or CDS), this is the portion of a gene's DNA or RNA that codes for protein. Often, it is mutation in the CDS that confers phenotypic variation. Hence, the CDS is a common target for diagnoses and variant calling
Copy number	The number of times a particular locus/gene/organism is present
Coverage	Percentage of bases of the reference genome covered by the sequenced reads
Critical control point	A critical control point (CCP) is a step in the process where preventative measures can be applied to prevent, reduce, or eliminate a hazard or risk that could adversely affect the output
***De novo* assembly**	A method for constructing genomes (a.k.a. assemblies) from HTS reads, with no a priori knowledge of the correct sequence or order of those fragments
Depth/read depth	Describes the average number of aligned reads at any position of an assembly consensus. (This is often termed "coverage" in Illumina documents)
DNA	Deoxyribonucleic acid. Occurs as single stranded (ssDNA) or double stranded (dsDNA)
eDNA sequencing	Sequencing DNA that has been shed by organisms into their environments (environmental DNA).
FASTQ/FASTA	**FASTA** is a standardised text-based format for representing either nucleotide sequences or amino acid sequences, in which nucleotides or amino acids are represented using universally recognised single-letter codes. The format includes sequence names and comments to precede the

	sequences. **FASTQ** format is a text-based format for storing FASTA and its corresponding quality scores.
Finished genome	A rare and special situation where complete genomes from multiple individuals/isolates /cultures from a single population or outbreak situation are sequenced at high depth and combined to form a population reference that may be used to identify intrinsic polymorphisms.
Guideline	A recommended beneficial activity that exceeds the standard
Hi-C	A method used to analyse spatial genome organization and map higher-order chromosome folding and topological associated domains. May be used arrange contigs in correct order for *de-novo* assemblies.
Housekeeping loci	Housekeeping loci or genes are typically largely conserved constitutive genes required for the maintenance of basic cellular function. Hence, they are used often as targets for positive control reactions in molecular biology techniques
HTS	High Throughput Sequencing, also known as NGS (Next Generation Sequencing) or Deep Sequencing
HTSIPWG	High Throughput Sequencing Implementation Plan Working Group
Index/Indexing	When more than one sample or target is represented in a library, the individual DNA preparations are labelled with unique indexes that identifies them throughout the process and enables demultiplexing prior to sequence analysis. Some manufacturers (eg. ONT) call these barcodes.
Index switching	Index hopping or index switching is a known phenomenon that can affect multiplexed samples, where a small proportion of sequencing reads are incorrectly assigned from one sample to a different sample in a pool as a result of signal noise
LEADDR	Laboratories for Emergency Animal Disease Diagnosis and Response network
Library (HTS library)	Input nucleic acid modified into a form compatible with the HTS platform
Map to reference	Aligning and mapping sequence reads to a reference sequence/genome
Metadata	Data that describe other data, structured reference data that helps to sort and identify attributes of the information it describes. Metadata summarises basic information about data, which can make it easier to find, classify, use and reuse particular instances of data.
Metagenomics	The direct genetic analysis of genomes contained within an environmental sample or other sample that may contain multiple organisms. Similar to shot-gun sequencing, although directed at complex samples that contain multiple organisms.
Metabarcoding	see **Barcoding**

Multiplex	Simultaneous involvement of multiple elements, such as multiple primer sets or multiple samples
Paramagnetic bead	Micrometre-sized para-magnetic particles coated with DNA binding matrix and used to separate DNA from chemical or biological suspensions
PCR	Polymerase chain reaction. Can be conventional PCR, qPCR (quantitative PCR), nested PCR (multiple rounds of PCR using previous PCR as template for the next) or other variants of the amplicon generation premise
PhiX	PhiX is a fragment library derived from a well-characterised bacteriophage genome. Because of the known sequence and balanced GC%, it is used by the Illumina platform as a positive control and spiked into the library
Phred or Q-score	The most common metric used to assess the sequence quality
Pipeline	A bioinformatics pipeline is composed of a series of software algorithms to process raw sequencing data and generate analysed data. Bioinformatics pipelines are either designed and developed by a provider with or without customization by the laboratory or entirely developed by the laboratory
Platform	A particular technology that may require specific processes, reagents, consumables compared to other platforms that achieve comparable outputs
Pooling	Combining multiple libraries together, usually in specified concentrations to fit purpose
Presumptive result	An indication based upon reasonable opinion or belief, that will normally require confirmation prior to being used as a result in diagnoses
Primer dimer	Also known as primer polymers, the result of priming oligonucleotides binding to one another to form a longer molecule
Quality assurance	Part of quality management focused on providing confidence that quality requirements will be fulfilled. The confidence provided by quality assurance is twofold—internally to management and externally to customers, government agencies, regulators, certifiers, and third parties. (www.asq.org)
RNA	Ribonucleic acid
Reference genome	Verified, accurate, high quality, high depth assembly consensus used for comparison or mapping of generated sequence reads**.** Recognised reference genomes of viruses and many prokaryotes will be complete or, ideally, finished genomes, however many eukaryotic genomes are not.
Report	A diagnostic report is the set of information that is typically provided by a diagnostic service to the requesting or submitting party when investigations are complete. A ***report*** is distinct from a ***record or method documentation***, which describes the data, process and is information retained by the diagnostic service as evidence of the process.

Sample	In the context of this document, sample means any biological material, including but not limited to, bacteria and other microorganisms, plant, animal, environmental, or cell lines containing copies of the original sample components. A sample may refer also to derivatives from these materials.
SCAHLS	Subcommittee on Animal Health Laboratory Standards
Shotgun sequencing	Randomly fragmenting the genome into small DNA fragments that are sequenced individually then reassembled in their correct order
Standard	A repeatable, harmonised, agreed and documented way of doing something. Standards contain technical specifications or other precise criteria designed to be used consistently. Defined by Standards Australia as documents that set out specifications, procedures and guidelines that aim to ensure products, services, and systems are safe, consistent, and reliable
Trimming	Identification and removal of adaptors, primers and/or low-quality base-calls from the termini or raw sequence reads
Validation	Determination and confirmation of the performance characteristics of a process or test
Verification	Confirmation of the characteristics of a pre-validated process in a particular environment, such as a laboratory or for a particular sample or instrument
Wet-lab/Dry lab	Wet-lab refers to the processes or parts of processes that occur in a laboratory designed for manipulating liquids, biological matter, and chemicals. Dry labs are focused on computation, physics, and engineering

Glossary of computer terms used in this document

Command line interface A text-based method for controlling computer function. Text commands are used to run programs, manage computer files and interact with the computer

Graphical user interface (GUI) A method for controlling computer function through visual representations (graphics). The user interacts with a computer with icons, menus and other visual indicators or representations. Most common desktop computers are operated using a GUI.

Internet protocol A set of standards for transferring, addressing and routing data on the Internet. Examples include hypertext transfer protocol (HTTP) and file transfer protocol (FTP).

Linux A family of open-source operating systems based on the Unix operating system and linked by a common core (kernel) developed by Linus Torvalds.

Metagenome-assembled genome A single-taxon genome assembly based on one or more sorted metagenomes that has been asserted to be a close representation to an actual individual genome. One that could match an already existing isolate or represent a novel isolate.

Open-source software	Software with published source code that anyone can inspect, modify, and enhance.
Operating system	Software that manages computer hardware, software resources, and provides common services for computer programs. Examples include Microsoft Windows, macOS and Linux distributions such as Ubuntu, CentOS and OpenSUSE.
Programming language	Any set of rules that converts words and grammatical symbols into various kinds of machine code or binary output that can be processed by computers.
Quality metrics	A set of statistical measurements that gauge the quality of a high throughput sequencing output and help inform confidence in downstream results.
Ransomware	A criminal practice where hackers encrypt or take control of a computer or computer network and demand a ransom to regain access to files or systems
Software	The instructions, data or programs used to operate computers and execute specific tasks.
Software container	A standard unit of software that packages a program or analysis pipeline and all dependencies so it can run quickly and reliably from one computing environment to another.
Software dependency	A computer program or function that is reused in another piece of software or analysis pipeline. Dependencies are used commonly in computer programming as a means of avoiding repeating work already done. In biological computing dependencies are often software programs that perform a specific purpose, e.g. sequence assembly or sequence read alignment, within a larger analysis pipeline.
Software environment	A collection of programs, libraries, and utilities that allow users to perform specific tasks. A software environment for a biological analysis pipeline could include the operating system, the database system, specific computer programs or dependencies that allow successful execution.
Software repository	A centralised storage location for software. Software repositories enable developers to easily create, maintain, and track software packages. Examples include Github, GitLab, npm and the Python Package Index (PyPI).
Version control	The practice of tracking and managing changes to information over time. Version control systems are software tools that allow the management of changes computer programs, documents, large web sites, or other collections of information.

PLANNING/PREPARATION/EXPERIMENTAL DESIGN

There is no "one-size-fits-all" for application of HTS to biosecurity purposes. HTS can be used to query the entire length of an organism's genome or can be targeted to query a set of predefined genomic regions. It provides scope to study a genome through *de novo* shotgun assembly, to map and compare with a reference genome or part-genome, to explore the variation of targeted sites among many samples concurrently, or any combination of these and the expanding list of other applications.

It is important to consider what the aims are prior to starting the testing of a material or searching for a target-type. Different aims will affect the choice of sample preparation, multiplexing, library preparation process, platform and sequencing. Laboratories must establish that manufacturer's performance specifications for kits and platforms, and bioinformatics pipelines and databases are appropriate and relevant to the aims before commencing a diagnostic test and that a chosen method has reasonable discriminatory power for the intended use[3].

In methods that include comparison with, or mapping to, a reference genome, a laboratory should identify the reference genome as part of the preparatory step. Reference genomes must meet the criteria defined in Table 1, or be sourced from a reliable data repository. For example, a reference genome could be selected from the National Centre for Biotechnology Information (NCBI) RefSeq database which contains non-redundant, high-quality datasets that have been annotated and curated by NCBI[3]. The inadvertent use of poor-quality reference genomes can lead to erroneous results from the best sequencing run[4]. If no recognised reference genome exists, other sequences can be used as references so long as it is appropriate for the application and an appropriate caveat is documented within the results/report.

Where repeated or standardised testing on multiple or ongoing samples is required, then a standardised workflow should be established and verified which will require less decision making at the start of each test event. However, HTS using material that does not have a pre-verified workflow plan must undergo a documented experimental design process

There are metrics to consider such as coverage and depth that will influence the quality assurance and the utility of the data analysis substantially, and these need to be considered in the planning stage.

Sample preparation

Laboratories must consider the content of the starting material and whether nucleic acid from organisms other than the target is present, such as host or environmental matrix. Consideration should be given to methods of increasing the proportion of target nucleic acid or if the presence of host nucleic acid is acceptable. In the latter case, the host will essentially become the sequencing

target and experimental design should consider this when determining extent of multiplexing and estimating data output volume.

Platform and consumable selection

Laboratories must select the platform based upon the purpose of the test, differences in and tolerance of error rate, read depth and coverage required and total cost. Laboratories may also consider bioinformatics software required for analysis and that this software may be sequencing platform and/or operating system specific.

Specific applications are suited to different library preparation variables such as single/paired ends and insert/fragment size, and these can have profound impacts on the final result. Some applications benefit from using a combination of platforms, whilst others such as smaller genomes can be achieved from a single platform[22]. A technical example of this is presented in Wick's (2021) "Guide to bacterial genome assembly"[22].

Purpose

Laboratories must consider the purpose of their study and plan for appropriate depth and coverage. Broadly, genome sequences can be classified as incomplete drafts, high quality drafts, coding complete, complete genomes and finished genomes[10], and each classification has appropriate applications. Incomplete drafts may result from shotgun sequencing of pathogens within host tissue or pests from environmental samples and may require confirmation using other methods, whereas an identification may be made with confidence from a high-quality draft or complete sequence depending on the taxonomic level required for the particular case. A finished genome is a rare and special situation such as generating a universally accepted reference sequence that will not normally be conducted within diagnostic investigations by State or private Biosecurity laboratories. A finished genome represents complete genomes from several individuals/isolates/cultures from a single population or outbreak situation and identifies intrinsic polymorphisms. It is possible that only designated reference laboratories will have sufficient resources to generate the finished genomes, and enable jurisdictional laboratories to accurately analyse and compare samples. Alternatively, a finished genome may result from collaboration between several laboratories and multiple sequencing platforms.

Table 1 describes standard depth and coverage requirements for different scenarios of genome sequencing of relevance to Biosecurity, labelling categories according to the completeness of the data[10], and suggested applications of each category to ensure that data is not applied inappropriately. Standardising these classification categories according to read depth ensures that errors introduced through sequencing resolution technologies are recognised, and that users of the data are aware of

relevant limitations for application of the dataset. All sequencing platforms will produce some erroneous reads, and this information is included in the platform specifications as well as reviewed in published literature[19]. It should be noted that each category of genome assembly has its benefits and purposes, and a laboratory should not consider it necessary to achieve complete genomes in all instances where coding complete genomes, high-quality drafts or even incomplete drafts may be equally fit for purpose. Additionally, where the coverage does not meet the requirements for any genome classification in Table 1 (i.e., it is less than 50%), the data still may serve its purpose and be classified as a partial sequence.

For example, a reference sequence should be a finished genome if available or possible to attain[4]. On the other hand, the resources required to achieve a finished genome are not appropriate for detecting the presumptive presence of a pathogen or pest which could be achieved by an incomplete draft sequence and confirmed using alternative methods, or by a high-quality draft.

Laboratories may validate reduced representation genome sequencing approaches such as exome sequencing, RADseq, hybridization capture, and transcriptome sequencing. In this case, the laboratory must validate and standardise appropriate values for depth to achieve appropriate specificity and sensitivity and fitting the nature of the standards in this document.

Amplicon/barcoding approaches, targeting small sections of the genome sometimes from mixed samples, need to ensure that sufficient sequencing depth is achieved to reliably detect target taxa of interest after suitable quality control thresholds have been applied. In this case, the required depth must be established through validation processes to determine sensitivity and specificity of each test protocol as fit for purpose. Laboratories should be particularly aware of the increased risk of contamination in amplicon sequencing and the effect of low levels of contamination upon the sensitivity of the method to detect individual taxa as true positive results compared to contaminating false positive results. In the absence of validation, for example the necessity for testing of previously untested sample types and/or amplicon targets, laboratories should aim for minimum depth of 10 times the depth of the same sequence observed in negative controls (or minimum sequencing depth of 30x if control sample has three or less corresponding reads). For variant calling within amplicon sequences, laboratories must validate required depth of any variant to be identified as a true variant (as opposed to a sequencing artefact). In the absence of validation in new sample material or amplicon targets, variants should be called only if the amplicon has minimum sequencing depth of 50x and the variant is consistent in at least 10% of reads.

Anticipating depth and coverage

Depth and coverage can be anticipated using the estimated genome or amplicon size of the sample (including host and other components if appropriate) and the expected platform/kit output specifications. Different biosecurity applications of HTS require different minimum levels of depth and coverage.

A laboratory must determine which levels of depth and coverage are most appropriate to their aims and, together with estimated genome or amplicon size, make an informed decision regarding options for level of multiplexing, platform and consumable kit output. Typically, larger genomes should be obtained from a combination of short and long read technologies, to mitigate the risk of platform specific sequencing bias and artefacts.

Table 1 provides coverage and depth guidelines for consideration in planning HTS for genome sequencing.

Table 1: Genome sequencing classification criteria

	Incomplete draft	High quality draft	Coding sequence (CDS) complete	Complete genome	Finished genome
Number of contigs	Multiple discontinuous contigs/scaffolds	1 (per segment) for viruses. Prokaryote and eukaryote genomes may be represented by multiple discontinuous contigs/scaffolds with mean depth >30	Single contig per segment, chromosome, plasmid, or any other natural partitioning unit of the genome	Single contig per segment, chromosome, plasmid, or any other natural partitioning unit of the genome	Single contig per segment, chromosome, plasmid, or any other natural partitioning unit of the genome
Open reading frames	incomplete	incomplete	complete	complete	complete
Genome coverage	>50% genome represented	Approximately 90%	Approximately 99%	100%	100%
Mean read depth†	15x-30x	>30x	>100x	>100x	>400x
Description	Common result of shotgun sequencing where target species is low copy number or low titre	May be missing termini or have lower depth in some regions. May be achieved from incomplete draft using Sanger sequencing to fill gaps (those gaps may be <30x depth)	No gaps or regions of lower depth, all ORFs are complete, some non-coding regions may be missing. May be achieved with long read and/or Hi-C data for bacterial-sized genomes and above.	Fully resolved, including non-coding regions and segment termini. Circularised segments for bacteria. May be achieved from coding complete using rapid amplification of cDNA ends (RACE) or other methods.	Complete genome for multiple samples and/or haplotypes to provide population-level characterisation
Detection/ identification	Presumptive result, confirmation required	Minimum standard	Recommended standard		
Describing novel taxon	Does not meet minimum standard	Does not meet minimum standard, unless validation of specified loci can be determined	Minimum standard for description of a novel species/taxon	Recommended standard for description of a novel species/taxon	Recommended for establishment of reference genome
Variant calling	Does not meet minimum standard	Minimum standard for variant calling, only if using regions of >30X depth that show >90% consensus for the variant allele	Recommended standard		

†Mean read depth can be summed when sourced from more than one sequencing platform (e.g. 100X combined coverage from 25X Oxford Nanopore and 75X Illumina platforms)

STARTING MATERIAL: SPECIMENS AND SPECIMEN PREPARATION FOR HTS

GENERAL

In most situations, specimens deemed suitable for molecular techniques such as PCR will be suitable also for HTS. Laboratories must apply the same suitability assessment as those applied to other specimens according to their existing quality procedures. If a laboratory is conducting HTS on nucleic acids extracted and submitted by another party, then the minimum standards must be documented and made known to the submitter.

Specimens, subsamples and metadata must be traceable through unique identifiers according to a laboratory's existing quality management system and Laboratory Information Management System (LIMS).

HTS must be conducted according to the same principles as other nucleic acid amplification and/or manipulation techniques, with physical separation of phases and contamination control[18].

Sample Collection and Storage

To minimise the risk of contamination, samples for nucleic acid techniques have some special needs for collection and preparation, in addition to the usual requirements for pathology testing.

The requirements for sample collection, initial processing, transportation and storage depend on the specimen concerned and the nucleic acid target (DNA or RNA). Specimens are to be collected according to the following principles, as contamination of samples can occur at any stage of specimen collection and processing.

Samples that have been used for other tests prior to nucleic acid detection testing are at increased risk of contamination. This is particularly so where the previous test performed was processed with other samples containing the nucleic acid of interest. If a number of tests are to be performed for diagnostic purposes then dedicated subsamples should be collected for nucleic acid testing. Due to the sensitivity of the methods used, only small amounts of target DNA need to be present in a sample to give positive results. Contamination must be minimised to avoid false detections.

Standard

- **Specimens must be collected in accordance with written specimen collection protocols and by appropriately trained personnel. Protocols must be described in the laboratory quality management system**

- **Specimens must be stored in a manner appropriate to preserve nucleic acid**
- **When client-collected samples are used for diagnosis, clear instruction must be provided to the client to reduce the likelihood of sample contamination**
- **Staff must be aware that minor degrees of cross-contamination that would not be significant for other types of tests, may result in erroneous results by HTS techniques. In addition, these types of tests require samples to be such that degradation has not occurred at a level that the detection is inhibited. Methods for collection are therefore to employ techniques and reagents that minimise the risk of contamination or degradation, such as clean nuclease-free specimen containers and sampling tools and the separation of samples at all stages of the sampling process. Maintenance of an adequate cold chain and/or nucleic acid preservative during collection and transportation will substantially improve the quality of results obtained for many nucleic acid detection tests**

Guidelines

- Wherever possible, nucleic acid tests should be performed on dedicated samples or on sub-samples taken **before** other tests are performed. Where it is necessary to perform nucleic acid test on samples that have already been used for other purposes and there is a significant risk of cross-contamination, then the report must be annotated accordingly and, if possible, results of significant interest confirmed on a dedicated sample
- If samples are referred to another laboratory for testing, then it is the responsibility of the referring laboratory to ensure that the sample conditions outlined above have been met, and to inform that receiving laboratory if they have not been met
- Single use disposable equipment should be used wherever possible

Sample Integrity

Care needs to be taken to ensure that DNA and RNA remain intact during sample storage, transport and preparation. If the number of target molecules in the sample is very small and if degradation does occur, a false negative may be obtained. If the starting material for amplification is RNA, the sample should be processed as rapidly as possible after collection to minimise RNA degradation by ribonucleases. Specific instructions for handling samples to minimise nucleic acid degradation must be included in all relevant manuals and be available to staff in collection centers.

Standard

- **Laboratories should follow the HTS manufacturers guidelines for sample integrity according to purpose. Any differences from manufacturer recommended guidelines must be validated prior to use with diagnostic cases**
- **If a sample of likely compromised integrity cannot be resampled and laboratory is required to test it, then reduced integrity and the risks this may have must be documented and be stated as a caveat in the test report.**

Inhibitors

The presence of inhibitors of, or substances that interfere with, nucleic acid manipulation in some clinical, plant or environmental samples is a major concern in sample collection and preparation. Common inhibitors include EDTA and heparin anticoagulants used in blood collection, haem in blood, eyes, phenol used in isolation of nucleic acids, some cleaning agents such as shampoos and other atopic agents, and the inherent qualities of some samples such as those containing melanin, bile salts, urea, polysaccharides, or humic acids. Plant samples with high starch content (e.g. potato tubers) and bulk seed samples may require particular extraction buffers or additional steps at extraction (such as dilution or 2-mercaptoethanol). With small arthropods/insects stored in ethanol or removed from sticky traps there should be little carry through of ethanol/glue through to the extraction. Extraction processes must be validated to eliminate polysaccharides such as chitin exoskeleton[15].

Laboratories must be familiar with the possible inhibitors/inhibitor removal for their common sample types[17].

Standard

- **Procedures and methods applied directly to a sample must be designed to minimise the risk of false negative results due to the presence of inhibitors of required enzymatic activities**
- **In susceptible samples, nucleic acid extraction methods must be validated for their ability to remove inhibiting substances through use of control PCRs**

Guidelines

- If the extraction method cannot be demonstrated to remove all inhibitors reliably, then a control/housekeeping PCR must be used in all tests on that sample prior to processing the extract for amplification or direct HTS. This may be either by amplification of another target expected to be present or by spiking the sample with

control nucleic acid

- Spiking samples with control RNA is difficult because of its poor stability. If problematic, alternative strategies may be appropriate such as RT amplification of transcripts expected to be present in samples, or use of synthetic or commercial RNA template constructs
- Where commercial kits are used and include inhibition controls as an optional component it is recommended that these should be used

Sample Preparation

The quality of nucleic acid prepared from a specimen has a major effect on the subsequent probability of successfully performing the test.

Preparation of nucleic acids must use a method that is fit-for-purpose. The method must take into account the aim of the experimental design and HTS procedure, the requirements for the selected library preparation and the platform used for reading the sequence. For example, application of a nucleic extraction process suitable for tissue/environmental samples may not adequately release nucleic acids from some potential pathogens or other elements within that sample (such as seeds, fungi or Gram-positive bacteria), and if those elements are the potential targets of interest, this must be considered when selecting the nucleic acid extraction process. In this example, the laboratory might consider pretreatment with bead-beating and/or suitable enzymes. Additionally, potential downstream effects of components of the nucleic acid extraction methods used should be considered (e.g. carrier RNA).

For HTS, RNA usually requires transcription to cDNA and often it is necessary to perform second strand synthesis. Laboratories must use reagents that have been validated in-house as functional with their chosen sequencing platform, or verify a recommended process according to sequencing platform manufacturer.

Performing HTS upon material amplified prior to library preparation (such as targeted amplicons) requires consideration in each preparation step such as initial extraction, amplification set-up, amplicon purification and amplicon evaluation.

Standard

- **Nucleic acids must be extracted and purified using standard protocols suitable for nucleic acid manipulation. The procedures used for nucleic acid isolation from the full range of sample types, collection methods and the condition of specimens received by the laboratory must be validated as fit for purpose for intended HTS**

techniques and procedures detailed in the laboratory methods manual

- **To detect contamination from a reagent, the nucleic acid extraction protocol must include a no template control (with no specimen added). This must be checked after the extraction process is completed to ensure there is no nucleic acid detected. This negative control should be processed through HTS in parallel with the samples if practical to do so, but checked here as a critical control point for further processing of the DNA extracted from the samples**
- **Where samples of low quality, quantity or integrity have been received the laboratory must notify the referring party and seek recollection. If this is not possible, the laboratory must make an informed decision regarding the risk of using HTS technology before proceeding. Decision factors should include the purpose of the test, the risks of reduced read numbers, the risks of false negative detection and any other relevant factors.**
- **Where amplicons are prepared prior to constructing the HTS library, these must be purified from any residual amplification reagents (eg. commercial purification columns or paramagnetic beads) before proceeding with the library**

Guidelines

- Where degradation is suspected through, for example, sample history or type, there should be confirmation that extracted nucleic acid is of a suitable quality for testing. For example, this may involve gel electrophoresis (agarose for DNA or formaldehyde or glyoxyl gels for RNA) or applicable instrumentation such as Agilent TapeStation. Results of such an assessment must be recorded with the sample details/records

ADDITIONAL INFORMATION FOR SPECIFIC SPECIMEN TYPES

Animal and plant tissue, and cell culture

DNA and RNA can be extracted from tissue or cell culture using specific RNA or DNA extraction methods or using total nucleic acid (TNA) extraction. For pathogen sequence detection, identification and characterisation, methods that yield purity and concentration suitable for nucleic acid amplification must be used.

Standard

- **Nucleic acids must be extracted according to a validated method that yields purity and concentration suitable for nucleic acid amplification and/or manipulation techniques**

Guideline

- Laboratories should be mindful that nucleic acids from environmental contaminants will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing
- Such samples may require washing prior to extracting nucleic acids, or subsamples should be removed from the interior of the sample
- Laboratories may wish to consider the application of specific pathogen nucleic acid enrichment processes when a specific target is studied within a more complex sample matrix

Small invertebrates

DNA can be extracted destructively from whole invertebrates, or non-destructively retaining relatively complete specimens to be used as vouchers or for alternative diagnostic processes, such as microscopic examination.

Standard

- **Nucleic acids must be extracted according to a validated method that yields purity and concentration suitable for nucleic acid amplification and manipulation techniques**

Guideline

- Laboratories should be mindful that nucleic acids from environmental contaminants, including gut contents and endosymbionts, and will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing

Bacteria

Pure bacterial cultures can be picked from solid media and suspended in a diluent or grown in culture broth. Nucleic acid from intracellular or other unculturable bacteria should be extracted from the host sample following guidelines of that host sample extraction.

Standard

- **DNA must be extracted according to standard validated methods that yield purity and concentration suitable for nucleic acid amplification and manipulation techniques**

Guideline

- Laboratories should be mindful that nucleic acids from environmental contaminants will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing
- Laboratories may wish to consider application of a form of circular double stranded DNA enrichment to the extract from the host material

Fungi

Pure fungal cultures (e.g. following hyphal tipping) can be picked from solid media and suspended in a diluent or grown in culture broth. Nucleic acid from unculturable fungi should be extracted from the host sample following guidelines of that host sample extraction.

Laboratories must be mindful that nucleic acids from environmental contaminants will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing. Such samples may require washing prior to extracting nucleic acids. Additionally, antibiotics should be added to any fungal culture medium to reduce bacterial contaminants.

Standard

- **Nucleic acid must be extracted from fungi according to a standard validated method that yields purity and concentration suitable for nucleic acid amplification and manipulation techniques**

Guideline

- Laboratories should be mindful that nucleic acids from environmental contaminants will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing

Environmental Samples

DNA can be extracted destructively or non-destructively from mixed samples of whole organisms, or from samples taken from the environment containing trace amounts of DNA.

Standard

- **Nucleic acids must be extracted according to a validated method that yields purity and concentration suitable for nucleic acid amplification and manipulation techniques**

Guideline

- Laboratories should be mindful that nucleic acids from environmental contaminants, including gut contents and endosymbionts, and will be included in the extract and will prevail through any shotgun HTS approach such as whole genome sequencing
- Environmental and mixed-specimen samples contain lower amounts of DNA per specimen than single specimen samples, and subsequently may be more at risk of contamination

LIBRARY PREPARATION

GENERAL

Library preparation is the process that modifies the input DNA into a form compatible with resolution by the HTS platform. Different platforms work with different processes and principles, and therefore different library synthesis techniques.

Usually, library synthesis uses commercial kits and components supplied by the manufacturer of the HTS platform or their recommended supplier, that has been optimised for specified purposes on their respective platforms.

When more than one sample is represented in a library, the individual DNA preparations are labelled with unique indexes that identifies them throughout the process and enables demultiplexing prior to sequence analysis. Sample indexing should be performed at the earliest possible stage of library preparation to mitigate the effect of cross-contamination[12]. The number of indexed samples in a test needs to account for the maximum sequencing depth and coverage required, compared with the estimated genome or amplicon size of the sample and output from the platform used.

Libraries should be quality checked prior to loading on the sequencing instrument. Quality assessment includes fragment size and library quantification. Methods include qPCR, fluorometric quantification (Qubit or equivalent instrumentation), and/or microfluidics-based automated electrophoresis systems (Bioanalyzer, TapeStation or equivalent instrumentation). qPCR methods do not indicate library fragment length, but selectively amplifies full length libraries for quantification using adaptor sequences as priming sites. Fluorometric methods, such as Qubit, risk overestimating the library concentration because this method measures all dsDNA in the pool. This includes partially constructed fragments (incomplete library fragments) and residual primer dimers from PCR. Library quantification using spectrophotometry-based methods are subject to overestimation of library concentration and should be avoided. UV spec methods quantify single stranded nucleic acids and free nucleotides along with complete, dsDNA library fragments and are not appropriate for sequencing applications.

Indexed libraries are pooled prior to resolution. The pool will normally consist of equal concentrations from each library to achieve equal representation of sequence read numbers, although the control libraries (e.g. No template control, positive control, reference material etc.) can be intentionally decreased by proportion. Some library preparation methods include quantity normalisation techniques such as DNA-binding paramagnetic bead saturation. Alternatively, each library can be quantified and

normalised manually prior to pooling. The latter approach is required if the starting template was not of sufficient quantity to reach bead saturation. Manufacturer's instructions provide guidance.

Standard

- **The process used for library preparation must be documented as part of the laboratory's quality manual**
- **The kits and reagents used for HTS library preparation should be purchased from a commercial and reputable supplier or, if made in the laboratory, follow standard protocols and be assessed for fitness for purpose (with parallel testing against a commercial or previously tested consignment) before use**
- **The kits and reagents used for HTS library preparation must be documented in the laboratory's reagent register**
- **The kits and reagents used for HTS library preparation should be used before expiry date wherever possible. Extension of shelf life should occur only if the continued utility has been demonstrated as fit for purpose**
- **Laboratories that have modified kit components or manufacturers procedures must demonstrate equivalence or superiority of the modified procedure before putting the test into routine use. In this case, the modified procedure must be treated as an in-house test for validation purposes. Such modifications include adjustment made to amount of starting template**
- **Where a laboratory uses a commercial test kit in which the methodology and reagents are unchanged from the manufacturer's instructions, the kit does not need to be independently fully re-validated in the user's laboratory provided the kit has been shown to be fit for purpose. The laboratory must establish as fully as possible the reliability of the kit for their purposes and samples, through a reduced validation/verification procedure**
- **The number of indexed samples in a test should not exceed the maximum indicated from the depth and coverage required, compared with the estimated genome or amplicon size of the sample and output from the platform used**
- **Consecutive runs of the same sequencing instrument using the same index pairs should be avoided[11]**
- **An indexed negative control (no template or nuclease-free water) should be**

included in an indexed library unless it is impractical to do so.

- **Laboratories must perform quality assessment of libraries prior to resolution using qPCR, fluorometric quantification, and/or microfluidics-based automated electrophoresis systems while considering the points above. Commercial qPCR kits for this purpose must be verified as fit-for-purpose. In-house developed methods must be fully validated prior to use**
- **Pooling of libraries should follow manufacturer's instructions or be validated as fit-for-purpose**

Guidelines

- The integrity of the library kit should be maintained (alternative reagents should not be substituted)
- The kits and reagents used for HTS library preparation should be used before expiry date
- Laboratories should consider including a known fragment in the indexed library, as a positive control
- Quality assurance of the library should be performed using microfluidics-based automated electrophoresis systems that simultaneously will determine fragment size, quantity, and presence of incomplete libraries

SPECIFIC PLATFORMS

ILLUMINA

Standard

- **Dual indexes should be used for multiplexing samples. The index combination must be unique to each sample. Selection of indexes must conform with manufacturers compatibility guidelines**
- **A known percentage of diverse-sequence control must be added to the library pool prior to sequencing. Illumina recommend PhiX. This serves as a performance and quality control for the test. Laboratories should use Illumina's recommendations for the proportion of PhiX to add for the run type and instrument**
- **For amplicon sequencing, a PCR-positive should be included in the library and spiked into the pool at 1% or the limit of detection, whichever is larger**

Guidelines

- When dual indexes are combinatorial (the index combination is unique within the pool), there is a small risk of index switching. Index switching is a known phenomenon that has impacted HTS technologies and may result in assignment of sequencing reads to the wrong index during demultiplexing, leading to misalignment. Laboratories should consider using unique dual sequences (both indexes are unique within the pool) to eliminate this risk
- Where possible, rotate index combinations across successive runs and check for unexpected index combinations. This assists in detecting background contamination

OXFORD NANOPORE TECHNOLOGIES (ONT)

Standard

- **When a laboratory multiplexes samples beyond the standard barcoding options supplied by the manufacturer then dual indexes must be used**
- **A known positive and negative control should be included in the library where possible**

SEQUENCE GENERATION/RESOLUTION

GENERAL

Different sequencing platforms generate and resolve nucleotide sequences using different principles and strategies. Currently, the more common strategies are sequencing by synthesis (Illumina), single-molecule real-time sequencing (Pacific Biosciences), ion semiconductor (Ion Torrent sequencing), sequencing by ligation (SOLiD sequencing), and Nanopore single strand sequencing, however, new technologies are emerging all the time.

During the sequence generation and resolution, there may be "real-time" in machine metrics generated. Such metrics might include cluster density, signal intensity, comparison between tiles/lanes, active pore count, interim Q-scores and so on. These are useful for ascertaining the overall quality of the run and the performance of the instrument and consumables but will not necessarily translate to the quality of the data generated.

Standard

- **The instruments, kits and reagents used for sequence generation and resolution should be purchased from a commercial and reputable supplier**
- **Instruments must be regularly serviced and maintained**
- **If a laboratory chooses to make in-house reagents that are commercially available, these must be assessed for fitness with parallel testing against a commercial or previously tested consignment before use. (This does not apply to simple components that need to be made just prior to use such as 70% ethanol or sodium hydroxide solutions)**
- **Manufacturer's instructions should be followed when using instrumentation. Any deviations from manufacturers protocols must be validated as fit-for-purpose**
- **The kits and reagents used for sequence generation and resolution should be used before expiry date wherever possible. Extension of shelf life should occur only if the continued utility has been demonstrated as fit for purpose, and must be documented**
- **Instruments used for sequence generation and resolution should be maintained according to the manufacturer's recommendations. Maintenance records must be kept as part of the laboratory's quality assurance process**
- **Laboratories that outsource HTS should ensure the service provider meets these**

standards

Guidelines

- The kits and reagents used for sequence generation and resolution should be used before expiry date
- Laboratories should consider instrument service and maintenance contracts with manufacturers to ensure continued performance and support

RAW DATA QUALITY ANALYSES & TRIMMING

Following de-multiplexing, which may be part of the sequence resolution process conducted by the instrument, the first step in data assessment processing for any HTS run should be raw data quality control. This can be applied through predefined software such as FastQC[1] (for Illumina) or Porechop (for ONT). FastQC, for example, provides several external metrics for quality management such as the mean per-read base quality ratings, the allocation of GC information, and the detection of the most duplicated read (interpretation of the metrics is provided by the developers documentation[1]). Alternative similar *in silico* applications are available, and laboratories must determine the raw data QC techniques most suited to their working environment.

Quality control can be paired with removal of any adaptors used by the sequencing platform, trimming of poor-quality ends, and removal of poor-quality reads and any other artefacts from the data set.

Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the sequence quality. It indicates the probability that a given base is called correctly. Lower Q-scores indicate a higher risk of errors.

Q scores are logarithmically related to the base calling error probabilities, $Q = -10 \log_{10} P$ (Illumina.com). For example, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of 1 incorrect base call in 1000 base pairs. This means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%. A lower base call accuracy of 99% (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 bp sequencing read will likely contain an error.

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

The expected Q-score for a sequencing platform protocol should be indicated by the manufacturer. Q30 is considered a benchmark for Illumina sequence quality (Illumina.com: Quality Scores for Next-Generation Sequencing. Illumina technical note: sequencing. Quality Scores for Next-Generation Sequencing).

However, aggressive trimming of bases <Q30 may not always be required[21]. Different applications and purposes may accommodate lesser quality scores than others. For example, the presumptive detection of a pathogen within shotgun sequence data from host tissue may be achieved from lower

scores and be confirmed with further targeted testing, whereas variant calling will be more dependent upon consistent high quality base calls. With sufficient depth (Table 1), stochastic errors such as the above will be rare and able to be corrected through reference mapping or assembly.

Standard

- **Raw data quality control must be performed following demultiplexing and before any further data analysis**
- **Laboratories must determine the raw data QC metrics and techniques most suited to their working environment**
- **Laboratories must have documented acceptance criteria for raw data quality scores for their common applications. Data that does not meet or exceed those criteria must be used with caution and any diagnostic inferences must be confirmed using additional testing**
- **Laboratories must document quality control reports for data generation runs. If this is in the form of a FastQC (or other) output, then there must be acknowledgement of acceptance added and signed by the operator**

DATA ANALYSIS PIPELINES AND MANAGEMENT (BIOINFORMATICS)

GENERAL

The incorporation of HTS sequencing techniques into diagnostic workflows requires a data analysis step to convert raw data into test results. In contrast to other diagnostic technologies, this data analysis step usually takes place on separate systems than the instrument and often requires high performance computer hardware. Sequencing instruments generally provide outputs in standardised data formats (such as FASTQ or BAM), which are then processed with application and workflow-specific software.

HTS data analysis generally consists of several processing steps, each with a separate software tool, which are performed sequentially to provide a processed output. These outputs could include whole genome sequences, the presence of a sequence signature indicative of an exotic disease or pest, or the identification of Metagenome-Assembled Genomes (MAGs) from the sequencing of all purified DNA present in a sample. The chain of processing steps used to generate the final output is referred to as a software pipeline or workflow.

This section and the section focused on Computing Infrastructure is based on the Guidelines for Implementing Diagnostic Next Generation Sequencing for Animal Health Laboratories in Australia produced by the Laboratories for Emergency Animal Disease Diagnosis and Response (LEADDR) network.

Pipeline design

Bioinformatic analysis pipelines can be performed manually, with an operator performing each step and examining resultant outputs independently. However, a more efficient approach is automation of each step, generally through use of a programming language (for example, include python, R, bash). Automated pipelines can be provided as part of a commercial package, or be designed bespoke by external contractors, researchers or internal staff. Several design considerations must be considered for the design of automated bioinformatic analysis pipelines. The use of workflow managers such as Snakemake, Nextflow or CWL may be useful in this regard.

Standard

- **Internally or bespoke generated HTS analysis pipeline methods must be designed as open-source applications or at the least have the workflow open to audit for quality assurance. Where possible it is recommended in-house analysis pipelines and applications are made publicly available via version-controlled software repository services such as GitHub (https://github.com/)**

- **External biological sequence databases are often used as part of HTS analysis pipelines. These databases can be used to compare recently sequenced samples with samples sequenced previously and generally in the public domain. Administrators of these databases can often make changes to data format or location that can cause HTS analysis pipelines to fail**
- **External biological sequence databases and pipelines should be used with caution, as the laboratory will be dependent upon accuracy that cannot be verified**
- **HTS analysis pipelines that rely on external reference databases must be monitored and maintained to prevent failure caused by database structure or format changes. Laboratories that use internally generated pipelines must have a documented process for the monitoring of relevant databases**
- **The laboratory must use a form of version control to track software and database releases and updates to analysis methods[8]**

  The laboratory may consider use of dedicated version control software (such as Git, Concurrent Versions System (CVS), or Apache Subversion (SVN))
- **Software dependencies are software used within a larger pipeline. For example, a larger pipeline that identifies sequencing reads unique to an exotic pest may use a dependency (such as a read aligner) to align reads to a sequences of known exotic pests. When updated/upgraded, incompatibilities between software dependencies may cause pipeline failure or erroneous results**
- **Internally generated HTS analysis pipelines must control versions of software dependencies. Only software dependencies that have been validated as part of the HTS analysis pipeline must be included. Software dependency versions must be recorded or included in data outputs generated directly by HTS analysis pipelines**
- **Internally generated HTS analysis pipelines must have a documented method of dependency management either through packaging dependencies within pipelines, separate software environments (such as with Anaconda or Python virtual environments) or with software containers (such as Docker or Singularity)**

Guidelines

- HTS analysis pipelines should follow scientific coding best practices as described in Wilson *et al* (2014)[23]

- Due to the breadth of computer system architectures available, HTS analysis pipelines should be designed in a manner that maximises flexibility. Pipelines should be designed to run on numerous systems as described in the Computing Infrastructure section

Pipeline validation and verification

The general principles of validation of laboratory tests (ISO 17025) also apply for HTS assays. These include design, development, technical validation, and monitoring /improvement, documentation requirements and ultimately assessment of fitness for purpose. Here, we include only factors that are specific to HTS analytics validation, and should be considered *in addition to* the principles of ISO 17025.

Where a laboratory chooses to use commercial pipelines, these must be verified as fit for purpose. The verification process must be fully documented. Where a laboratory chooses to generate bespoke pipelines, these must be fully validated with documentation of the validation process. Additional assistance and guidelines to the validation of pipelines have been described by Roy *et al.* (2018)[16].

Standard

- **The validation (or verification) study must be designed to provide objective evidence that the bioinformatics pipeline is fit for the intended purpose**
- **During validation experiments appropriate values for key parameters of a bioinformatics pipeline must be established**
- **The laboratory must validate the entire bioinformatics pipeline as a whole, under the given operational environment. A laboratory may choose to put together its bioinformatics pipeline using any combination of commercial, open-source, or custom software. Regardless of whether an individual component has been validated, the laboratory is still required to validate the entire bioinformatics pipeline under their operational environment**
- **The laboratory must determine *performance metrics* of the pipeline. The laboratory must determine which metrics need to be studied for each application. Some commonly used performance metrics are:**
    - **Accuracy (combined diagnostic sensitivity and specificity)**
    - **Precision (repeatability and reproducibility)**
    - **Diagnostic sensitivity (e.g. number of test positive which are truly infected)**

- **Diagnostic specificity (e.g. number of test-negatives which are truly not infected)**
- **Limit of detection (e.g. analytical sensitivity). The LOD is the estimated amount of analyte in a specified matrix that would produce a positive result at least a specified percent of the time.**
- **Analytical specificity, that can be divided into:**

  a) selectivity, which refers to the extent to which the workflow/pipeline can accurately detect the targeted analyte in the presence of interference such as matrix components

  b) exclusivity (where relevant), which refers to the capacity of the workflow/pipeline to detect an analyte or genomic sequence that is unique to a targeted organism, and excludes all other known organisms that are potentially cross-reactive

  c) Inclusivity (where relevant), which is the capacity of an assay to detect several strains or serovars of a species, several species of a genus, or a similar grouping of closely related organisms

- **The validation study must define valid ranges for commonly assessed *quality metrics.* Based on the results of reference material and other previous experience, it is possible to establish what can be expected and/or accepted from a valid pipeline. Instances of deviation from these pre-defined ranges may not affect final results/conclusions for every application but should be documented**
- **Reference materials used for pipeline validation must be appropriate for assessing performance of the pipeline for its intended purpose. Validation of a bioinformatics pipeline generally involves executing it given some input data where the correct status of the variant is known (known as reference material, RM). RM may consist of well characterised data sets (e.g. FASTQ files), rather than physical materials such as DNA samples**
- **The validation study must establish appropriate hardware and operating system environments to allow successful execution of the pipeline. Validation must be conducted in a system that closely resembles the actual operational environment**

Guidelines

- The laboratory should compare the results from multiple pipelines, where possible, to allow identification of pipeline-specific artefacts

- A bioinformatics pipeline could fail due to the corruption of an input file generated by primary analysis or intermediate steps within the pipeline. It could also fail due to excessive load on the server or interrupted network connection. A laboratory should consider assessing whether the pipeline can detect corrupted files or interrupted execution, and generate appropriate error messages

Documentation

Standard

- **The laboratory must document all components of, changes to, and auditing of the bioinformatics pipeline. This includes the software packages, custom scripts and algorithms, reference sequences and databases. Any changes, patch releases or updates in processes or version numbers must be documented with the date of implementation such that the precise informatics pipeline and annotation sources used for each test and report is traceable**
- **The laboratory must document the pipeline quality metrics generated and assessed during a test**
- **The laboratory must document the results of the pipeline validation. The validation documentation must detail the performance of the pipeline such as the sensitivity, specificity and accuracy of the pipeline to detect variants and any limitations of the pipeline**
- **The laboratory must document the process of data management and storage. The laboratory needs to define the minimum set of data to store. Typically, this will involve storage of .fastq, .bam and .fasta files. It is recognised that long term storage, due to the large file sizes, may be problematic for some organisations**

Guidelines

- The laboratory must document all training and staff qualifications. Given the rapid advances in bioinformatics, when implementing HTS-based assays, the laboratory needs to consider appropriate staff training and ongoing professional development of staff in bioinformatics. Staff involved in the reporting of HTS results must have, as a minimum, an understanding of the bioinformatics analysis steps and resources used for annotation

Data management

Standard

- **The laboratory must ensure that data management meets documented requirements for data integrity and security including avoidance of tampering with primary data files and/or corruption of result files**
- **A clear data management strategy must be developed including which data to keep, minimum storage time and records of storage location, how to access, etc., in keeping with organisational and jurisdictional requirements**
- **A clear data backup strategy must be developed to ensure data integrity and protect against cybersecurity/data-loss risks**

Guidelines

- Data backup strategies should follow the 3-2-1 principle (3 total copies of data, 2 stored locally, 1 externally) or similar. This principle may not be relevant to data stored with cloud storage providers. In that instance, laboratories are encouraged to be aware of the data redundancy policies of their cloud storage provider. The laboratory must use structured databases wherever possible. Using an appropriate LIMS consistent with quality management principles is recommended
- Laboratories are encouraged to submit HTS data to appropriate externally shared databases, once it has undergone adequate QA and has been approved for release by the relevant local, state, territory and/or national management bodies (as appropriate)
- The laboratory should consider establishing an internal database of genomic findings. The curation of an internal laboratory database can assist with the local interpretation process in the future and may alleviate any confidentiality issues associated with early submission to public databases
- Cloud storage has the potential for reducing the loss of data due to hardware failure, and is readily scalable, but issues of bandwidth for access, security on non-approved servers and confidentiality of identifiable data remain major concerns. Laboratories considering Cloud storage must establish the physical sovereignty of the data with particular reference to inside/outside of national borders and if that is appropriate for

the data involved. Laboratories must ensure that this complies with jurisdictional legislative requirement.

- Laboratories should consider archiving raw HTS data for analyses with future software releases.

COMPUTING INFRASTRUCTURE

GENERAL

HTS technologies introduce complex analytical methods which often require substantial bioinformatics and IT infrastructure that are not the usual domain of regulatory and/or accreditation agencies. These systems may include stand-alone workstations, internal or external datacentres, or cloud computing providers, brief descriptions of these are provided in Grzesik *et al* (2021)[7]. Often Linux operating systems with a command-line interface are used for these systems, as this configuration allows for maximum flexibility and the use of cutting-edge software tools. However, these configurations require highly trained staff to operate. Windows-based computers are used for several applications, especially those that involve a graphical-user interface and usually require less staff training. However, for bioinformatics applications Windows-based computers are limited by a reduced number of available software tools. Cloud-computing service providers can provide computing resources without initial infrastructure costs. These can be sourced through cloud providers such as Amazon, Google and Microsoft Azure; or through cloud software platforms such as Galaxy. Several factors must be considered in the design and procurement of computing infrastructure for diagnostic HTS purposes.

Hardware & software

Standard

- **Computing hardware must at least meet the minimum specifications of the software**
- **The computing hardware should be capable of performing the required analyses efficiently and/or capable of running the chosen software using training/control datasets (i.e. datasets with characteristics consistent with samples to be analysed)**
- **Hardware must meet requirements for data storage, retention and security**

Guidelines

- Consideration should be given to equipment which exceeds the minimum specifications software in order to reduce processing time, and hence turnaround time.

Networking and data transfer

Standard

- **Confidentiality of data must be maintained during data transfer**

Guidelines

- During analysis, large datasets may need to be transferred between computing hardware (i.e. from sequencer to analytical computer or from sequencer to storage location). A speed of 1 gigabit/second is suggested as a minimum data transfer speed. This requirement will affect network cables as well as routers/switches. Infrastructure capable of faster transfers will reduce delays introduced by the transfer of large files
- Wherever possible, data should not be transferred using USB "memory sticks" or external hard drives with moving parts. Consideration should be given to the use of high-speed network connections between the various components of the computing hardware
- Appropriate steps should be taken to ensure that data corruption does not occur during transfer. Checksums for individual files or compressed files can be generated using a variety of software packages (see glossary). Laboratories should investigate data integrity checking software and implement strategies to identify file corruption

Data management and storage

Standard

- **The laboratory must develop a formal data management policy which minimizes the possibility of data loss**
- **The laboratory must ensure that data are stored in a manner that prevents loss in the event of hardware failure (i.e. data should have redundant backup)**

System integration and maintenance

Standard

- **The laboratory must show that the hardware and software used can be maintained appropriately, including installation, updates, security, and troubleshooting**

- **The laboratory must maintain a list of critical external biological sequence databases and other utilities and ensure that network access to these databases is maintained. Bioinformatic analysis pipelines can often rely on the comparison of data to external biological sequence databases and other utilities. Many of these use internet protocols (such as FTP) which can be blocked by information technology departments**
- **The laboratory must ensure that computing equipment used for HTS analysis is regularly updated and integrated with networking and cybersecurity policies of the wider organisation. The possession of large biological databases carries the potential for cybersecurity risk such as ransomware attack**

Guideline

- Laboratories may wish to consider Cloud storage[7]. Cloud storage has the potential for reducing the loss of data due to hardware failure, and is readily scalable, but issues of data sovereignty, bandwidth for access, security on non-approved servers and confidentiality of identifiable data remain major concerns

VALIDATION OF PROTOCOLS

Laboratories must validate all protocols within the HTS workflow that are developed in-house, and/or verify all protocols performed according to standardised processes such as manufacturer's instructions. Validation or verification is required to demonstrate a process is fit for purpose, providing confidence in the results.

Standard

- **Laboratories must follow validation/verification processes according to their jurisdiction and quality system. For example, veterinary laboratory test validation processes should be based upon those of the WOAH[13] and designed to suit a laboratory's HTS platform and usage**
- **Test validation is discussed elsewhere in detail [3,4,6,11,12,13,14,20], and will not be repeated here. Minimally, test verification must be undertaken to establish analytical sensitivity and specificity compared to existing gold standard techniques. In some cases, HTS may be more sensitive than the reference method and ways to arbitrate discrepant results should be applied[3] such as a confirmatory specific PCR or the application of latent class analysis[5]**
- **Laboratories must determine the optimum validation process for their application(s) and document fully that process as part of the quality management system**

Guideline

- Participation in proficiency testing or "ring-testing" should be considered between laboratories. It is acknowledged that few, if any, are available at the time of this document preparation, however laboratories should seek these out as and when they become available. Alternatively, laboratories may development relevant in-house proficiency tests.

REPORTING

One of the recommendations of the HTSIPWG is to define a cross-jurisdictional governance structure which would provide common policy and guidance to reporting of results and data from HTS. Whilst this remains in development, laboratories must have a documented clear process for the reporting of HTS used for diagnoses. Interpretation of results and integration with existing epidemiological knowledge is highly relevant. This process should be included within the quality system.

Standard

In the absence of a national guideline in this regard, laboratories should consider including the following in their reporting process as a minimum standard:

- **Unique specimen identity/laboratory number assigned, and the specimen type used or sub-sample if appropriate. For example, if the nucleic acid has been previously extracted for another purpose and later used for HTS**
- **Platform used (eg. Illumina MiSeq, Illumina NextSeq, ONT Minion etc)**
- **A statement that the data has met the laboratories expectations of data quality (e.g. Using FastQC, Q30 scores etc.)**
- **Brief description of process of analysis, ie. *De novo* assembly or mapping to a reference. For the latter, the identifier of the reference sequence (such as accession number) must be documented along with the percentage of identity and the percentage of coverage**
- **If applicable, a classification of the sequence generated, using the classification given in Table 1 of this document**
- **If using databases to form an identification, the name of the database, the accession number of the most likely taxon or taxa related to the specimen or aetiological agent found within**
- **Brief summary/interpretation in layman's terms that addresses the reason for conducting the test**

  For example, are these isolates indistinguishable or different? Here, a report should avoid ambiguous claims such as "the isolates are related" without substantiating the claim such as referring to valid relatedness/phylogenetic studies for that organism. For infectious agents, a report may include statements such as "consistent with a monoclonal infection/ consistent with a variation during an outbreak, further tests

needed/ indicates more than one infecting strain", or "these samples were consistent with epidemiologically related isolates of this outbreak/ consistent with variation within an outbreak but further tests needed/ represent epidemiologically unrelated strains" [3]

This section of the HTS quality standards will be superseded upon publication of specific policy guidelines pertaining to reporting.

INCIDENTAL FINDINGS AND INCOMPLETE SEQUENCES

One of the recommendations of the HTSIPWG is to define a cross-jurisdictional governance structure which would provide standardised policy and guidance for incidental findings. Whilst this remains in development, laboratories must ensure they have documented a clear and defined policy around notification of such incidental findings used for diagnoses.

Incidental findings may be defined as data of potential diagnostic or clinical relevance unrelated to the reason for performing the test. Incidental findings of significance would include, for example, the identification of pests or pathogens that were not targeted by the original sequencing activity.

Findings from HTS activities must be defendable and reproducible, and this includes findings using partial sequences or genomes that might indicate the presence of an exotic or reportable species.

Incomplete or incidental findings have been termed the "Incidentalome" and there is a large potential for over-interpretation[9]. Hence, incidentalomes must be confirmed as genuine using further testing prior to reporting results.

The local policy should consider factors such as the seriousness of the pest/pathogen found, whether it represents a known clinical situation or risk, whether the process used complied with validated standard procedures, whether the finding poses risk of further disease and suffering to the individual, co-habiting/close contact individuals, or potential disease outbreak[8] or spread of pest, and also should follow State and Federal regulations pertaining to suspected notifiable pests and diseases.

Standard

- **Notification of an incidental finding must follow the policies and procedures defined by each jurisdiction for notification of the finding of a new or incidental species, and these processes must be documented as part of the quality system. For example, the sample submission process may have a clause that identifies ownership of the sample and/or its derivatives to the laboratory and state the responsibilities of the laboratory in regard to legislative requirements to report significant findings.**
- **Limitations associated with the finding and options to improve on, or confirm, the result (e.g. additional sequencing, testing remaining extract by specific assays, further sampling) must be included in the process, so that the laboratory may make an informed decision on the significance of the finding and the subsequent steps**

- **Incidental findings or incomplete sequences should be considered in conjunction with symptoms (plants)/clinical presentation (animals)**
- **Incidental findings of significance must always be confirmed using an alternative method prior to reporting. If no suitable method exists, the laboratory should seek follow-up samples for confirmation. Laboratories may wish to consider describing presumptive results until confirmation can be established.**
- **Incomplete sequences that do not reach the minimum classification description of incomplete draft (see Table 1) should undergo additional HTS or an alternative confirmatory method prior to reporting**

This section of the HTS quality standards will be superseded upon publication of specific policy guidelines pertaining to reporting incidental findings.

STAFFING SKILLS, TRAINING

The importance of maintaining technical capacity and expertise in diagnostic services cannot be underestimated. The employment of skilled personnel in HTS (both wet lab and bioinformatics) and of upskilling existing staff is strongly encouraged. HTS is a rapidly evolving tool and staff conducting HTS should undergo continual education and upskilling.

Standard

- **Laboratory staff must only perform HTS with appropriate training. This may be limited to specified wet-lab or dry-lab processes if required by the organisational structure of the laboratory**
- **Laboratories conducting HTS must maintain a staff skills register, competency records and/or training records pertaining to HTS**
- **Activities to upskill, maintain currency and accreditation should be undertaken frequently, and as often as possible. Frequency must be determined according to the needs and scope of the laboratory.**
- **Appropriate skilling activities can include, but is not limited to:**
    - **Peer-to-peer training within an organisation**
    - **On-line resources such as those listed in the resources section of this document**
    - **Proficiency testing/ring testing (wet-lab and dry lab applications) when available**
    - **Laboratory residentials, or sabbaticals with laboratories proficient in the appropriate skills**
    - **Attendance at workshops, conferences, symposia and webinars hosted by third parties such as:**
        - **Sequencing providers (e.g. AGRF)**
        - **Manufacturers (e.g. Illumina, Nanopore)**
        - **Computational resources (e.g. Galaxy, Pawsey Centre, NCI and QCIF)**
        - **Training providers (e.g. BioCommons, BioPlatforms, Melbourne Bioinformatics, Monash University, QCIF Bioinformatics and ABACBS)**
- **Staff training must be documented as part of the laboratory's quality system**

RESOURCES

Galaxy Australia online web-based platform of bioinformatics tools and tutorials
usegalaxy.org.au

GitHub Internet hosting service for software development and version control using Git. Popular within the bioinformatics community

Guide to bacterial genome assembly GitHub hosted "choose your own adventure guide" to bacterial genome assembly. An easy to follow document that outlines different assembly techniques depending on read type (long and/or short)
https://github.com/rrwick/Trycycler/wiki/Guide-to-bacterial-genome-assembly

Illumina self-learning resources
sapac.support.illumina.com/training.html
sapac.support.illumina.com/bulletins.html

ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories
www.iso.org/standard/66912.html

Laboratories for Emergency Animal Disease Diagnosis and Response (LEADDR) network
www.awe.gov.au/agriculture-land/animal/health/laboratories/guidelines-next-gen-sequencing

WOAH/OIE/World Organisation for Animal Health
www.woah.org/en/home/
Standards for high throughput sequencing, bioinformatics & computational genomics
www.woah.org/fileadmin/Home/eng/Health_standards/tahm/1.01.07_HTS_BGC.pdf
Principles and methods of validation of diagnostic assays for infectious diseases
www.woah.int/fileadmin/Home/eng/Health_standards/tahm/1.01.06_VALIDATION.pdf
Biotechnology advances in the diagnosis of infectious diseases
www.woah.int/fileadmin/Home/eng/Health_standards/tahm/2.01.02_BIOTECH_DIAG_INF_DIS.pdf

Oxford Nanopore self-learning resources
nanoporetech.com

SCAHLS guidelines for nucleic acid detection techniques
www.awe.gov.au/agriculture-land/animal/health/laboratories/procedures/other/guidelines-for-nucleic-acid-detection-%28nad%29-techniques

Validation
Further information about experiments to assess performance of molecular assays and data presentation are available at www.agriculture.gov.au/agriculture-land/animal/health/laboratories/tests/test-development/validation-nucleic-acid-detection

REFERENCES

1. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Belak, S., Karlsson, O., Leijon, M., Granberg, F. (2013) High-throughput sequencing in veterinary infection biology and diagnostics. *Rev. sci. tech. Off. int. Epiz.*, 2013, 32 (3)
3. CLSI (2021) *Molecular methods for genotyping and strain typing of infectious organisms*. 1st ed. CLSI guideline MM24. Clinical and Laboratory Standards Institute.
4. Chain, P. *et al.* (2009) Genome Project Standards in a New Era of Sequencing. *Science* 326: 236-237
5. Cheung, A. *et al*. (2021) Bayesian latent class analysis when the reference test is imperfect. *Rev. sci. tech. Off. int. Epiz.*, 2021, 40 (1)
6. Gardner, I. *et al*., (2020) Validation of tests for OIE-listed diseases as fit-for-purpose in a world of evolving diagnostic technologies. *Rev. Sci. Tech. Off. Int. Epiz*., 40 (1)
7. Grzesik, P., D. R. Augustyn, T. Wyciślik and D. Mrozek (2021). Serverless computing in omics data analysis and integration. *Briefings in Bioinformatics* 23(1).
8. Hall, A., Hallowell, N., Zimmern, R. (2013) Managing incidental and pertinent findings from WGS in the 100,000 genomes project. *A discussion paper from the PHG Foundation.* ISBN 978-1-907198-12-0. www.phgfoundation.org
9. Krier, J. and Green, R. (2014) Management of incidental findings in clinical genomic sequencing. *Current Protocols in Human Genetics*. DOI: 10.1002/0471142905.hg0923s77
10. Ladner *et al.* (2014) Standards for sequencing viral genomes in the era of High Throughput Sequencing. *mBio* 5(3): e01360-14
11. Mintzer, V., Moran-Gilad, J., Simon-Tuval, T. (2019) Operational models and criteria for incorporating microbial whole genome sequencing in hospital microbiology e A systematic literature review. *Clinical Microbiology and Infection* 25: 1086-1095
12. NPAAC (2017) *Requirements for human medical genome testing utilising massively parallel sequencing technologies*. National Pathology Accreditation Advisory Council
13. Newberry, K and Colling, A. (2019) Quality standards and guidelines for test validation for infectious diseases in veterinary laboratories. *Rev. Sci. Tech. Off. Int. Epiz*., 2019, 40 (2)
14. OIE World Organisation for Animal Health. Principles and methods of validation of diagnostic assays for infectious disease. Chapter 1.1.6 of *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2021*
15. Paszkiewicz, K., Farbos, A., O'Neill, P., Moore, K. (2014) Quality control on the frontier. *Frontiers in Genetics* 5: 157
16. Roy, S. *et al.* (2018) Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines - A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *Journal of Molecular Diagnostics* 20: 4-27
17. 16, Schrader, S., Schielke, A., Ellerbroek, L., Johne, R. (2012) PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology* 113: 1014-1026
18. Subcommittee of Animal Health Laboratory Standards (SCAHLS): Veterinary Laboratory Guidelines for Nucleic Acid Detection Techniques. March 2008
19. Stoler, N. and Nekrutenko, A. (2021) Sequence error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* 3: 1-9
20. Van Borm, S., Wang J., Granberg, F., Colling, A. (2016) Next-generation sequencing workflows in veterinary infection biology: towards validation and quality assurance. *Rev. Sci. Tech. Int. Epiz.* 35:67-81
21. Watson, M. (2014) Quality assessment and control of high-throughput sequencing data. *Frontiers in Genetics* 5:235

22. Wick, R. (2021) Guide to bacterial genome assembly. GitHub repository, https://github.com/rrwick/Trycycler/wiki/Guide-to-bacterial-genome-assembly

23. Wilson G*, et al*. (2014) Best Practices for Scientific Computing. *PLoS Biol* 12(1): e1001745.