Implementation of next generation sequencing for use in animal disease investigations in Australia

GUIDELINES FOR IMPLEMENTING DIAGNOSTIC

NEXT GENERATION SEQUENCING

FOR ANIMAL HEALTH LABORATORIES IN AUSTRALIA

Version 4, 08 March 2019

Guidelines for Implementing Diagnostic Next Generation Sequencing for Animal Health Laboratories in Australia

Background

This document is aimed at the Laboratories for Emergency Animal Disease Diagnosis and Response (LEADDR) network and other laboratories preparing for implementation of next generation sequencing (NGS) for use in animal disease investigations in Australia. At the time of completing this first version, there were no national standard publications specifically aimed at the use of next generation sequencing in animal health diagnostic laboratories in Australia.

Animal health diagnostic laboratories in Australia should adhere to these guidelines to ensure that the high quality of diagnostic testing across Australia is maintained. It is hoped that this document may provide the basis for a national standard in the future.

The first version was drafted at LEADDR's Annual Face to Face Meeting on 22 November 2017 in Sydney, New South Wales and finalised out-of-session by the network on 17 May 2018. It was based on the Royal College of Pathologists of Australasia (RCPA) 3/2014 “Massively Parallel Sequencing Implementation Guidelines”. The fourth version incorporates suggestions from a number of external reviewers. This document forms part of the deliverables of a LEADDR project funded by the Australian Government Department of Agriculture and Water Resources through the Agricultural Competitiveness White Paper.

Outline

Chapter	Topic	Page
	Definitions	3
	Abbreviations	5
CHAPTER ONE	Ethical and legal Issues	6
CHAPTER TWO	Wet lab	8
CHAPTER THREE	Bioinformatics	13
CHAPTER FOUR	Reporting	20
CHAPTER FIVE	IT Infrastructure	22
	Recommendations	24
	References	24
	Version Control	28

Definitions

Term	Definition
Base call	The nucleotide, or base, (guanine (G), cytosine (C), adenine (A), or thymine (T)) assigned to a sequencing signal. These are platform derived. A series of base calls comprises a sequence read.
Bioinformatics pipeline	the primary, secondary and tertiary analysis of massively parallel sequencing (MPS) data that is performed computationally by individual computer algorithms; the pipeline is the specific combination and order of these algorithms used to analyse the data
Exome	the protein-coding region of the genome; it is made up of exons
GC-content	guanine cytosine content of a gene or region of DNA
Genome	the basic information-encoding, replicable part of an organism, in the form of DNA or RNA (in various forms), and might include organelle or extra-chromosomal (e.g. plasmids) components
LEADDR	The LEADDR network reports to the Animal Health Committee in Australia, and coordinates a national laboratory network to harmonise or standardise the testing of targeted emergency animal diseases and support the management of emergency animal disease incidents as needed
Incidental findings	any finding of significance that was identified by the test but is not related to the request for testing; these may also be interchangeably referred to as secondary findings or unsolicited findings
Indel	a mutation class where there has been either an insertion or a deletion of nucleotides, or a combination of both, compared to the reference sequence
Next generation sequencing (NGS), also called massively parallel sequencing (MPS)	the collection of technologies used to enable the sequencing of many, usually short fragments of DNA, at the same time to provide greatly increased sequencing coverage of either individual samples or multiple samples that can be distinguished by the use of introduced sample indexing; this may also be referred to as MPS; these terms are used interchangeably throughout this document
Orthogonal testing	the utilisation of different validation or confirmation techniques that are functionally and statistically independent from the original testing
Primary analysis	the analysis of hardware generated data, machine statistics, production of sequence reads and quality scores

Term	Definition
Quality assurance (QA)	A system which comprises of a set of procedures intended to ensure that a performed service (e.g. diagnostic testing) adheres to a defined set of quality criteria and meets the requirements of the customer
Quality control (QC)	Quality control is the set of measures and procedures to follow in order to ensure that the quality of a service or product is maintained and improved against a set of benchmarks and that any errors encountered are either eliminated or reduced.
Quality scores	These are platform derived reflections of the signal to noise ratio and reflect the probability that the base call was correct. An acceptable raw base call quality threshold should be established during validation, and incorporated in bioinformatics filters to eliminate poor quality data during analysis.
Reference materials	a set of materials that have predetermined and known properties for use in experimentation to provide control and comparison; these are typically resourced commercially or generated by a laboratory network
Sample indexing	the embedding of sequence markers directly onto sample molecules, which enables the identification of NGS samples; this technique is important in modern NGS systems where multiple samples are multiplexed into the same sequencing reaction
Sequence reads	a series of base calls
Secondary analysis	QA filtering of reads, assembly and alignment of reads, QA and variant calling on aligned reads
Sequencing	sequencing is the process to determine the sequence of nucleotides
Tertiary analysis	QA/QC of variant calls, annotation and filtering of variants, assessment of pathogenicity and clinical significance, genome browser driven assessment, and other 'sense making analyses' such as population frequency and structure assessment, treatment/prognostic/classification associations
Variant calling	Identifying when a base call is different to the reference genome; includes indels
Wet laboratory	a laboratory that utilises chemicals or biological matter that are handled in liquid solutions or phases for analysis

Abbreviations

Term	Description
AS	Australian Standard
CVO	Chief Veterinary Officer
DNA	Deoxyribonucleic Acid
FASTA	Text-based nucleotide sequence, using single-letter codes
FASTQ	Text base nucleotide sequence with attached quality scores
ISO	International Organization for Standardization
IVD	*In Vitro* Diagnostic device
LEADDR	Laboratories for Emergency Animal Disease Diagnosis and Response
LIMS	Laboratory Information Management System
MPS	Massively Parallel Sequencing
NATA	National Association of Testing Authorities, Australia
NGS	Next Generation Sequencing
OIE	World Organisation for Animal Health
PCR	Polymerase Chain Reaction
QA	Quality Assurance
QC	Quality Control
qPCR	Quantitative PCR
RCPA	Royal College of Pathologists of Australasia
RM	Reference Materials
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
VCF	Variant Call Format

CHAPTER ONE: Ethical and legal Issues

1. Introduction

Diagnostic testing by next generation sequencing (NGS) share many ethical, legal and socioeconomic issues with other forms of veterinary investigation. While there has been substantial progress in implementing diagnostic approaches in inherited diseases and cancer in human health, the application of NGS to microbiology diagnostics is still in its infancy. Being a new technology, NGS methods present technical challenges. In addition, multiple new issues are presented with NGS. For example, incidental findings, particularly those which could impact trade and Australia's disease free status, are potentially magnified due to the volume of information that these tests may yield.

NGS (e.g. whole genome sequencing or exome sequencing) can generate information pertinent to the identification of diseases other than the targeted animal health condition being investigated. NGS can, therefore, be viewed as comprising both a diagnostic and a screening function. The scale of this overlap in test function is unprecedented. The implications of this complex testing scenario for submitters requires clear explanation.

NGS should not be performed without careful consideration of these broader issues. For this reason, this chapter on ethical and legal implications of NGS precedes the chapters detailing the analytical, interpretive, reporting, and resource requirements for such testing.

2. Responsibilities

2.1 Explicit consultation framework

There should be an explicit consultation framework between the laboratory and the submitter. While all diagnostic testing requests imply a consultation between the submitter and the laboratory running the test, this should be an explicit requirement of referrals for NGS. The Chief Veterinary Officer (CVO) for the state or territory in question, must be informed of significant results.

The laboratory should clearly state their policy on reporting incidental findings and pathogens of unknown significance. The submitter should know what analytical approach will be undertaken, the policies of the laboratory service with respect to reporting of findings, incidental findings, storage of data, and links with research bodies, so that this information can be conveyed to the submitter during the informed consent process.

2.2 Consent process

A consent process should be in place and submitters need to consent to NGS. Details of the request and scope of testing need to be clarified prior to testing as part of the test request. This can be accomplished via a consent form.

2.3 Contributing to public databases

On-line sequence databases have been a fundamental resource for infectious disease researchers and practitioners alike. Access to pathogen sequence facilitates an understanding of the pathogen genetic sequence and the evolution of the pathogen as well as the development of new diagnostic testing. The utility of existing tests can also be assessed, and an evidence base for the use of a new test can be built. All submitters should be asked for permission during the consent process to allow the contribution of their sample's data to public databases for the common good as appropriate. Consent from the jurisdiction's CVO may also be required.

2.4 Incidental finding policy

Submitters should receive a clear written record of the policy regarding the reporting of incidental findings. Whatever approach the laboratory chooses, clear verbal and written communication of the policy about what findings will and will not be disclosed should be provided.

2.5 Record and sample retention

Samples should be retained in accordance with existing National Association of Testing Authorities (NATA) requirements. The existing NATA standard for samples submitted for diagnostic testing specifies the retention of diagnostic material for "three months from the date of issue of the report". It is reasonable to apply this to samples submitted for next generation sequencing.

3. Reporting Responsibilities

3.1 Reporting rules

Data generated by NGS are subject to the usual reporting rules and obligations operating in each jurisdiction. As with any other test result from an animal health laboratory in Australia, the laboratory is responsible for the security and privacy of submitter's information. Results that are of national and/or international importance should be reported to the CVO in the relevant jurisdiction. Refer to the jurisdiction's list of notifiable aquatic and terrestrial animal diseases for more information.

3.2 Data storage

Data generated by NGS should be stored in accordance with NATA requirements and the standards operating in each jurisdiction. Standards Australia AS/NZS ISO/IEC 17025 incorporate electronic storage of laboratory information management system (LIMS). For human medical testing, NPAAC advocates storage of .bam files for 4 years, .vcf files for 10 years and reports for ~100 years.

4. Purpose and Scope of Testing

4.1 Purpose and scope of testing

Throughout the process of testing, there should be explicit distinctions between the two types of NGS:

a. targeted diagnostic testing (i.e. of specific pathogens, selected gene(s), or exomes) and
b. discovery studies which include searching for currently unknown agents.

An understanding of the role of NGS in investigation of disease is rapidly evolving. Provided a suitable ethical framework is in place, the diagnostic laboratory may provide samples and data to researchers for assistance, but should always retain sufficient sample for the minimum retention period and laboratory records to meet NATA requirements for the retention of data.

For more information see references: 1-10

CHAPTER TWO: Wet lab

1. Introduction

Despite a range of available platforms and workflow options, NGS has been adopted in all areas of molecular diagnosis. There are critical differences between NGS and traditional Sanger sequencing. The wet laboratory process is one such area of critical difference. Robust quality assurance (QA) and quality control (QC) procedures are essential to ensuring the reliability of NGS results.

This chapter will focus on the wet laboratory issues including laboratory environment, sample and library preparation, template generation, sequencing and QA.

1.1. Wet lab processes

Many of the guidelines in this document are common to all forms of nucleic acid testing. These guidelines should be read in conjunction with ISO 17025 and all relevant NATA documents.

It is not our intention to generate a user guide and provide all the solutions. Instead, we try to include some relevant resources for reference. For example, some relevant wet laboratory issues can be found from the website of the Division of Laboratory Programs, Standards, and Services (DLPSS) of the American Centers for Disease Control and Prevention (CDC) and the World Organisation for Animal Health (OIE).

For more information see references: 1-10

2. Measures to Control Contamination

2.1 Physical design

As with all molecular testing, it is advised that there are separate areas for preparation of reaction mixes, and a unidirectional workflow from the preparation area to the areas with samples and finally amplified product. The laboratory should be designed to minimise the contamination of samples at different stages of the workflow with other samples and/or amplified products. The laboratory should ensure the physical design can accommodate separate areas for animal derived samples and amplified material. Possible cross contamination between these areas including by movement of equipment, staff, or aerosols should be assessed and minimised. Measures should be available to both detect cross contamination between clinical samples, and to eliminate it. Elimination may include the use of hypochlorite or other appropriate decontamination measures.

2.2 Minimising cross contamination between samples

The laboratory should ensure recommended and appropriate maintenance and cleaning processes are performed to eliminate carryover contamination. No template controls should be included on all sample preparation batch runs to check that reagents are free from nucleic acid contamination. The laboratory should include a monitoring process for carryover contamination as part of regular internal QC. Sample indexes (barcodes) used to identify unique reads in pooled libraries can be used to detect carryover contamination. These should be re-used as infrequently as possible. Consecutive runs of the same sequencing instrument using the same barcode indexes should be avoided. Frequent reuse of the same set of barcode indexes will compromise the laboratory's ability to detect cross-contamination at any stage of the sequencing procedure. It is recommended that a large collection of

barcodes can be drawn upon and are frequently turned-over. Indexes need to be carefully selected. Index switching can occur through homologous recombination. Workflows using the latest plated, single use unique dual index (UDI) designs, where the index barcodes are completely different for every sample, are highly recommended as UDI's minimise cross-contamination and eliminate index switching.

2.3 Sample Indexing

Sample indexing should be performed at the earliest possible stage of library preparation to allow subsequent detection of cross-contamination. The laboratory should avoid workflows that offer the potential for undetectable sample cross-contamination. Workflows that call for multiple manipulations, additions, and incubations of samples prior to index ligation or amplification increase the risk of undetectable sample to sample cross-contamination whereas workflows which add unique indexes to each sample early in the library preparation process provide a means to make cross-contamination detectable. If stocks of index adapters/primers are used for multiple rounds of sample preparation, great care should be taken to prevent cross-contamination of these reagents during handling. Single-use, plated indexing reagents are preferable to prevent cross-contamination.
The use of No Template Controls (NTC) is recommended to allow detection of contamination in reagents. The NTC should be carried forward to the completion of library preparation and should be included in sequencing.

2.4 Orthogonal testing

It is recognised that carry-over contamination can occur, both within and between NGS runs. At this stage it is very hard to avoid cross contamination, particularly with RNA targets. To have confidence in testing results, it is recommended to conduct orthogonal testing using a conventional diagnostic test methods to confirm the identity of the pathogen detected in the original material. This may take the form of agent culture, a real time PCR on original material, Sanger sequencing on original material, immunohistochemistry (IHC) on fixed tissue, and electron microscopy (EM) on appropriately prepared samples. There should also be strong clinical and epidemiological support for the diagnosis. If any of the above do not support the NGS result, careful consideration should be given to the interpretation of the sequence data.

3. Wet Workflow Validation

Validation of the wet workflow is very important, and indeed a massive undertaking for NATA accreditation. Quality metric decisions should be based on a laboratory's validation data.

3.1 Genomic platform

The genomic platform used must meet the specifications required for the diagnostic purpose and be operated in accordance with best practice as determined by the manufacturer. Where possible, bleach washes of the instrument fluidics should be employed between NGS runs to eliminate carry-over. Consideration should be given to biases inherent in the platform of choice. There is value in using a combination of platforms. Particular attention should be given to ensuring that any systematic weaknesses or errors of the sequencing system do not limit the diagnostic sensitivity and specificity of the assay, or that if such flaws exist, that orthogonal testing is employed to detect variants in regions of bias. Examples include regions of high guanine cytosine (GC) content or repetitive regions.

3.2 Validation of wet laboratory workflows

The laboratory should validate the operational performance of the wet laboratory workflow used in molecular diagnosis for a particular purpose. For example NGS assays have been used to study multifactorial diseases and co-infections, new and emerging diseases of viruses (virus discovery) and

other pathogens and for testing of adventitious agents of vaccine seeds. Diagnostic applications, e.g. for disease diagnosis and surveillance and other applications are on the way.

Expansion of genomic methods for diagnostic applications makes it increasingly important to demonstrate data quality, reliability and reproducibility. The laboratory should empirically determine their minimum requirements for data quality.

Analytical sensitivity and specificity are important performance characteristics for genomic diagnostic applications. Diagnostic laboratories should document these aspects of the laboratory workflow by comparison of test results obtained under conditions defined above, to those obtained from a gold standard method (e.g. Sanger sequencing or real time PCR LOD (limit of detection) determination). Reference materials with a known truth set are useful during validation and ongoing QA/QC.

For more information see references: 2, 7

3.3 Standard controls for NGS

The laboratory should regularly monitor the performance of the wet laboratory workflow used in molecular diagnosis. Inclusion of known DNA control/standard samples at <10% of the pooled libraries at regular intervals would allow ongoing monitoring of assay performance and data analysis processes. There is a warning against using plant viruses as controls as this could be misleading. Commercially prepared spike in controls are available.
See comment about NTC (negative control) above.

3.4 Outsourcing NGS services

The use of outsourced platforms and services for diagnostic services should meet all of the standards outlined in this document. If part of the next generation sequencing process is to be outsourced, NATA accredited providers or providers showing full compliance with ISO and/or NATA standards should be used. It remains the responsibility of the clinical laboratory to review, retain and make available for audit all documentation related to clinical testing.

4. Sample selection, preparation and storage

4.1 DNA quality and quantity

The laboratory should assess the quantity and quality of DNA samples before proceeding with diagnostic application. Failure to exclude samples of poor quality or insufficient quantity of amplifiable DNA can significantly affect the sensitivity and specificity of genomic diagnosis and lead to the possibility of false negative results. Fluorometric quantitation with dyes specific for DNA/RNA is highly recommended over spectrophotometric measurement. This is of particular significance where the sample type may be associated with limiting amounts of DNA. This information should be disclosed in the report sent to the submitter. However it should be remembered that even in the event of sufficient total nucleic acid, testing may still result in a false negative.

4.2 Optimising DNA sample concentration

The laboratory should determine an appropriate range of DNA sample concentration and types to be included for an efficient test using genomic methods. Where appropriate, consideration should be given to including related affected and unaffected samples in the analysis. For example sequencing healthy animals from the same affected herd may confirm a normal commensal and thereby exclude a novel agent as a new pathogen of interest. It may be possible to have different standards for targeted and de novo approaches.

5. Library Preparation

5.1 Sample tracking

The laboratory should have an effective system to track the samples during the multiple step process of library preparation. A LIMS capable of tracking a multistep workflow, with multiple samples, and QC steps should be considered. Critical steps, such as the addition of molecular barcodes, must be identified and managed appropriately to ensure sample integrity and tracking is preserved.

5.2 Adequacy of DNA fragmentation

For those laboratories that use protocols making use of DNA fragmentation, quality assessment of DNA fragmentation procedure is essential to ensure the right size distribution and accurate amount of fragmented DNA samples. The latter is critical for equal molar representation if multiple barcoded samples are to be subsequently pooled for library preparation.

5.3 Quality assurance (QA) measures

The laboratory should determine the optimal conditions for library preparation. Documented metrics of performance of library preparation should be generated and used to QC library preparation steps on all clinical samples. For example, effect of input mass of DNA, fragmentation conditions, PCR cycles, etc. should be assessed. QC metrics in the form of Bioanalyser traces, spectrophotometric or fluorometric readings, or real-time PCR results should be produced and routinely collected and compared to those of an optimal validated run where appropriate.

6. Template Generation

6.1 DNA library

The laboratory should have a quality assessment procedure to assess the quality and quantity of a prepared DNA library used for template generation. An accurate estimation of DNA library quantity is essential for optimal clonal amplification. Quantification should be based either on fluorometry, or on amplifiable templates (i.e. DNA fragments with proper ligated adaptors). For example, quantitative PCR (qPCR) has high levels of sensitivity and specificity and can accurately measure quantities of DNA.

6.2 PCR amplification

The laboratory should have a quality assessment procedure to assess the adequacy of PCR amplification used for template generation. Quality assessment of the clonal amplification procedure is essential to ensure an adequate representation of DNA samples in the template. This is critical for equal representation if multiple barcoded samples have been pooled during library preparation.

7. Data Generation

7.1 Coverage

For known targets, the laboratory should establish empirically the coverage necessary for accurate detection of sequence variants and copy number changes, and provide the best estimation of false positive and negative rates.

The laboratory should employ QC measures that specify the quantity and quality of DNA sequence data to accurately differentiate all targeted sequence variants. This is especially critical when a multiplexed target enrichment procedure has been used to generate libraries.

7.2 Barcoding

If multiple samples are to be sequenced simultaneously, the laboratory should have QA measures to demonstrate that DNA sequence data generated cannot be attributed to the wrong sample. Consideration should be given to the use of barcoded DNA samples, particularly using plated unique dual index tags, and the possibility of sequence data being misdirected to the wrong specimen (i.e. index mis-assignment). The use of dual indices can be useful to reduce index mis-assignment, while

the use of no template controls (NTC) can assist in identifying the level of index mis-assignment, if NTC samples are carried forward to sequencing.

For more information see reference: 11

7.3 Data storage

Data should be stored as required for diagnostic DNA studies. Consideration should be given what would be the suitable data format to keep (see further discussion in Bioinformatics Section). The raw reads and quality scores should be kept as a minimal requirement. Data storage should also comply with overarching regulatory and legislative requirements (see section in Ethical and Legal Issues.)

7.4 Exception log

Any exception should be recorded for samples where steps used in the analytical process deviate from laboratory standard operating procedures. This exception log should be kept with the reason(s) for deviation and should retain links to the sample.

For more information see references: 12, 13

8. Quality Control (QC) and Quality Assurance (QA)

8.1 Quality metrics

The technical manager should be able to identify the appropriate quality metrics that are suitable for their genomic tests. Consideration should be given to cross platform confirmation. Sanger sequencing should be considered to reduce false positive and/or negative rates, particularly with small indel variants. The limitation of next generation sequencing should be presented in the final report (See the details in the Reporting section).

QC of sequencing data may include:

- Base call quality scores
- Read depth
- Uniformity of read coverage
- Read enrichment (for capture-based methods)
- Percentage PCR duplicates (for capture-based methods)
- GC bias
- Decline in signal intensity along a read
- Minimum data requirement per sample

The laboratory should implement QA measures that evaluate the entire process. Well-characterised DNA samples should be used as internal QC samples. Consideration should also be given to obtaining reference materials.

The laboratory should base their quality metrics decisions on their validation data.

8.2. Monitoring of quality over time

Acceptable intra-and inter-run variability should be established during validation and monitored in diagnostic laboratories. It is important to determine assay precision, i.e., the degree to which repeated measurements give the same result – both repeatability (within-run and between run and operator variation) and reproducibility (between-laboratory, and/or between platform variation).

Monitoring of quality metrics over time should be performed to identify any trends which may affect the performance of the assay. Upgrades to instruments, sequencing chemistries and reagents or kits

used to generate genomic data must be verified or validated prior to implementation and performance monitored.

Genomic technologies are rapidly evolving. Consideration should be given whether positive findings in genomic analysis should be confirmed by a different chemistry or a second method, particularly when new agents, or notifiable agents have been identified.

For more information see references: 11, 14

8.3 External QA

When and where possible, the laboratory performing diagnostic NGS should participate in suitable genomic proficiency testing or inter-laboratory sample exchange programs to meet the requirements for external quality assessment measures. It is recognised that at this stage, it may be difficult to find such programs. However as time progresses, it is anticipated that such programs will become available.

For more information see references: 15, 16

CHAPTER THREE: Bioinformatics

1. Introduction

1.1 Scope

Diagnostic applications of NGS in animal health laboratories span a wide range of approaches. These may include resequencing of single genes, gene panels, whole exomes, and whole genomes of known and unknown pathogens.

The scope of this chapter is restricted to consideration of NGS applied to diagnostic DNA analysis, analyses of RNA, transcriptomes, and epigenetics. Issues addressed cover the range of NGS for genes, panels of genes, exomes and whole genomes. As the size and complexity of the analysis increases, additional procedures and safeguards may need to be included to ensure robustness and reliability of the analysis.

1.2 The Bioinformatics Pipeline

A bioinformatics pipeline refers to a number of computational tasks, generally applied sequentially. The pipeline begins with the output of an NGS instrument such as an image or FASTQ files, and progressively analyses these data through key steps, ending up with a variant call format (VCF) file, or even further with an annotated spreadsheet (comma separated value, tab separated value) or Text file.

While there is no one standard pipeline, most bioinformatics pipelines convert the data through a series of fairly standardised steps. A bioinformatics pipeline can be provided by the NGS instrument vendor, using proprietary software, or using open-source software. None of these approaches has been shown to be innately superior to the others, provided they are selected, tuned, validated or verified (as appropriate) and applied correctly.

Primary analysis:

This phase receives raw electronic information from the NGS instrument, and converts it using the vendor's proprietary algorithms into genomic signals such as nucleotide sequence and quality of the individual base calls ("base calling"). The laboratory usually has relatively little control of this phase as it is under the instrument manufacturer's control. Where multiplexing strategies have been applied, de-multiplexing is performed at this analysis stage; de-multiplexing re-identifies the sample from which individual sequence reads were derived.

For amplicon sequencing strategies, primers have to be trimmed from the reads. The outputs of the primary analysis phase are usually FASTQ files. QC (including machine metrics) and acceptance criteria should be applied at this stage.

Secondary analysis:

This phase receives the FASTQ files from the primary analysis, and maps (or aligns) it to the reference sequence (if this exists) and identifies changes from the reference sequence (variant calling). De-novo assembly can also be undertaken, especially when there is no reference sequences (see 'De-novo assembly').

The secondary analysis pipeline must be tailored to the NGS technical platform used. For example, duplicates arising from PCR strategies are typically marked for capture-/enrichment-based approaches where this strategy helps identify clonally-derived sequences which, if not accounted for, can contribute to erroneous variant calls. In contrast, PCR duplicates are not marked in amplicon-based sequencing strategies as clonally-derived sequences cannot be distinguished from independently derived PCR duplicates, unless unique molecular barcodes (also called UMIDs) are used. Unique

molecular barcodes label the individual genomic templates / DNA molecules from a single sample used as input into the library preparation and thus allow differentiation between PCR duplicates derived from independent DNA molecules vs those derived from the same clonal population.

Local realignment can optimise alignment in regions where sequence variation is present to increase accuracy and minimise false-positive variant calls.

Variant calling is then performed to identify sequence variations from the reference such as SNVs and small insertions/deletions, copy number alterations and structural changes.

The outputs of the secondary analysis phase are usually BAM and VCF files. There are a large number of commercial, academic and in-house tools in use for the secondary analysis of MPS data. Further QC should be applied at this stage; for example, assurance that the bioinformatics pipeline has run to completion, review of any error logs, monitoring of the total number and type of variant calls obtained, and review of any gaps in sequence coverage

Tertiary analysis:
Tertiary analysis concerns the annotation of the identified sequence variants and may involve a combination of the following: comparison of the identified sequence variants to those reported in databases (e.g. Genbank), and research of the sequence variant/gene published in peer-reviewed literature.

For large-scale genomic investigations, such as whole-exome or whole genome analysis, tertiary analysis further involves a process of variant filtering and prioritization, by removal of findings of lesser interest. The aim of filtering and prioritization is to reduce the number of candidates to those most-likely associated with disease. For genome-scale investigations, variant filtering and prioritization is typically performed in a (semi-)automated fashion. The resulting pre-filtered set of candidate sequences is then manually reviewed in further detail to allow interpretation and classification of the sequence, and to take into account the current limitations of annotation databases. Clinical interpretation and reporting of findings are discussed in chapter 5.

The outputs of annotation and filtering phases commonly are annotated VCF or CSV/TSV (spreadsheet) files. Further QC system standards can be applied at this stage; for example, review of gene symbols and aliases in use by different components of the pipeline which may result in failure to map identifiers between tools, review of expected number of known vs novel variants.

2. Documentation

The laboratory has a choice of using vendor-supplied pipelines, open-source pipelines, or some combination of both. In general, less documentation is required for vendor-supplied pipelines, but more customisation and fine-tuning is possible for in-house developed or applied software. The requirements described in this section apply regardless of the source of the bioinformatics pipeline.

2.1 Documenting the informatics pipeline

The laboratory must document all components of, changes to, and auditing of the informatics 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. If information from public websites is used, the date of access should be documented.

2.2 Version control

The laboratory must use version control to track software releases and updates to analysis methods. The laboratory may consider use of dedicated version control software to assist with this requirement for managing software code, such as Concurrent Versions System (CVS), Apache Subversion (SVN), or Git. There are also dedicated software tools for management and control of laboratory method documents and validation records.

2.3 Quality metrics

The laboratory must document the quality metrics assessed during a test. For the informatics pipeline, relevant quality metrics include but are not limited to: the total number of reads passing quality filters, the percentage of reads aligned, the number of single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) called.

2.4 Pipeline validation

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 validation document must be readily available to staff involved in MPS based genetic testing.

2.5 Records

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

2.6 Data handling and storage

The laboratory must document the process of data handling and storage. The laboratory needs to define the minimum set of data to store. Typically, this will involve storage of .bam, .vcf files but not image files. Alternatively, the laboratory may store .fastq files to allow reanalysis of the primary data. Interpreted variant call files, such as those after review of the initial calls must also be stored. It is recognised that long term storage, due to the large file sizes, may be problematic for some organisations.

2.7 Conditions for data reanalysis

The laboratory must define and document the conditions for data reanalysis. As our understanding of sequence variation expands and our bioinformatics tool set improves, it may be necessary to re-evaluate the annotation of a variant or to re-analyse the sequence data. The laboratory must specify under which circumstances, if any, such reanalysis is to be performed.

3. Validation

The general principles of validation of laboratory tests (IVDs) (see NATA Requirements) also apply for NGS assays. These include design, development, technical validation, and monitoring /improvement, documentation requirements and ultimately assessment of fitness for purpose. However, that document does not address aspects specific to genomics and NGS, which is covered in greater depth in resource documents such as Clinical Laboratory Standards Institute (see reference 25), and reference 2.

Risk of errors in bioinformatics pipeline: In an analysis pipeline for identification of sequence variants, one must have high confidence that the resulting variant calls have high sensitivity and specificity. Although true positives (TP) can be distinguished from false positives (FP) easily through external validation, it is almost impossible to systematically distinguish false negatives (FN) from the vast

number of true negatives (TN). Different pipelines may vary widely in their degree of concordance of classification of findings with the risk of false negative rate being particularly difficult to address, especially with indels compared to SNVs. The majority of differences between variant calling pipelines appear, however, in 'problem regions' of the genomes, such as repeat sequences, regions of sequence homology elsewhere, low complexity regions and regions with errors in the reference assembly; the concordance between calls can often be further improved by applying post-variant calling filters to remove artefactual calls.

Besides variant calling, the use of different variant annotation software programs and transcript annotation files can also make a substantial difference in annotation results that are not commonly appreciated (Reference 35). This report highlights the need to ensure bioinformatics pipelines are subjected to rigorous validation and QC, especially for clinical diagnostic applications, and that any limitations are clearly documented and commented on, where required, in the report.

3.1 Design of validation study

The validation study must be designed to provide objective evidence that the bioinformatics pipeline is fit for the intended purpose (see also 3.2). During validation experiments minimum values for key parameters of a bioinformatics pipeline will be established.

The validation study must identify and rectify common sources of errors that may challenge the accuracy of the bioinformatics pipeline. As part of the validation study, it is important to gain an understanding of common error sources that may compromise the validity of the pipeline, such as:

- Inherent limitation of individual programs
- Inadequate optimisation of parameters of individual programs
- Problems with data flow between individual programs
- Use of incorrect auxiliary files (e.g. wrong genome reference)
- Hardware or operating system failure

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 (i.e., same hardware specification, same operating system, same parameter setting, and same input load).

3.2 Validation process

The laboratory must determine standardised performance metrics of the pipeline. The use of standardised performance metrics ensure that validation results could be communicated and compared unambiguously. Some commonly used performance metrics are:

- The frequency of True Positive, True Negative, False Positive, and False Negative results (dependent on prevalence)
- 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 negative which are truly not infected)
- Limit of detection (e.g. analytical sensitivity)

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 some general statistics that we could expect from a valid pipeline. Deviation from these pre-defined ranges may indicate a necessity for closer examination, but does not automatically imply a validity problem.

Acceptability criteria must be defined to describe clearly the minimum quality metrics required to demonstrate the bioinformatics pipeline is fit for purpose. One way to demonstrate acceptability and fitness for purpose is to undertake proficiency testing carried out by a NATA accredited (or international equivalent) third party. However, currently there are not many proficiency testing providers for NGS. Sample exchange can be a useful in the absence of formal QAP programs.

3.3 Benchmark

The laboratory must benchmark the bioinformatics pipeline using reference material, where available. The reference materials chosen 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. These input data are called Reference Material (RM). The usefulness of a RM depends on obtaining a large variety of input, from sequence containing only simple SNV to sequences containing complex indels. RM can be generated entirely by in silico simulation, or sequencing real oligonucleotides of known sequences. Note that for the purposes of specific bioinformatics QA, this RM may consist of well characterised data sets (e.g. FASTQ files), rather than physical materials such as DNA samples. It is possible to obtain a large variety of RM from in silico simulation. Nonetheless, RM from real sequences should also be employed as they capture characteristics of real data.

3.4 Multiple pipelines

The laboratory should compare the results from multiple pipelines, where possible, to allow identification of pipeline-specific artefacts. Multiple pipelines could generate quite different variant calling results from the same input FASTQ file. One strategy to validate a pipeline is to measure the concordance between the results of a given pipeline against several other widely used pipelines. High concordance does not necessarily guarantee correctness, but low concordance indicates problems. Poor concordance commonly overlaps with ‘problem regions’ of the genome, e.g. low complexity regions, as discussed above. Any limitations of the chosen pipeline must be defined as part of the validation study.

3.5 Data corruption

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. As part of the validation procedure, it is important to assess whether the pipeline can detect corrupted files or interrupted execution, and generate appropriate error messages.

3.6 Hardware and operating system

The validation study must establish appropriate hardware and operating system environments to allow successful execution of the pipeline. The bioinformatics pipeline can be executed in a dedicated computer server, a shared high performance computing (HPC) environment, or the cloud. The successful execution of these programs also depends on the use of appropriate operating system, appropriate auxiliary software program, and supporting reference files (e.g., the human reference genome file, and gene annotation file). Validation should be conducted in a system that closely resembles the actual operational environment. See also issues raised in section 5 of this chapter.

3.7 Acceptable performance specifications

When changes are made to the test system, the laboratory must demonstrate that acceptable performance specifications have been met before using the changed test system for clinical purposes.

3.8 Limitations

The laboratory must define the limitations of the informatics pipeline. Common limitations of the bioinformatics pipeline include but are not limited to: the maximum size of indels detectable, regions of poor mapping and/or excessive read depth, regions of poor sequence coverage, repeat regions and homopolymer sequence regions that may affect variant calling. There may also be specific limitations of individual specimens that can affect the capability of a given bioinformatics pipeline.

4. Quality Control (QC) and Quality Assurance (QA)

It is important to distinguish QC for checking the quality of sequencing data, and QC for ensuring the correct execution of the bioinformatics pipeline. Data QC is important for checking whether the sequencing data is of sufficient good quality to ensure variant calling can be performed to the required standard. On the other hand, pipeline QC is concerned about whether the bioinformatics pipeline has been correctly executed according to the predefined quality metrics for a given sequencing data input.

QC of bioinformatics pipeline may include the following metrics:

- Mapping quality
- Presence of duplicate reads
- Expected number of variants
- Expected percentage of known variants

4.1 Metrics

The laboratory must monitor quality metrics and acceptability criteria of the informatics pipeline established during pipeline validation. Quality metrics are to be recorded for each test performed and interpreted in the context of the acceptability criteria that were defined during pipeline validation.

Deviation of achieved quality metrics from defined acceptability criteria must be investigated and mitigated. Significant deviations may require repeat of the test.

Quality metrics and acceptability criteria must be reviewed regularly to ensure relevance to current test performance. Revalidation must be performed where ongoing deviations are observed and/or substantial changes to the informatics pipelines have been made. Choice of appropriate quality metrics can be of significant help in troubleshooting the source of the problem in an underperforming test. Trend analysis of bioinformatics quality metrics may also prove to be useful. The appropriateness of the chosen quality metrics to monitor test performance needs to be reviewed regularly, at least annually.

4.2 External QA

The laboratory must participate in QA programs for the analysis and interpretation of DNA sequence variants, where such programs are available. Proficiency testing may involve an external QA program, sample exchange, use of electronic sequence files, reference materials and other approaches.

Examples of QA program include those organised by the Royal College of Pathologists of Australasia (RCPA) and the European Molecular Genetics Quality Network (EMQN). Currently, their QA programs for NGS analysis are in pilot phases.

The laboratory should consider the use of reference materials for ongoing monitoring of test performance.

5. General Informatics Aspects

This section refers to general issues that are applicable in all circumstances and environments. Where a laboratory uses off-site or hosted facilities (including “cloud” facilities), these requirements must be met for all stages of the process, including those not physically co-located or under the direct control of the laboratory.

5.1 Data security and privacy: Data management

The laboratory must ensure that data management meets requirements for data integrity and security including avoidance of tampering with primary data files and/or corruption of result files.

NGS data may involve the management of very large data files (in excess of hundreds of gigabytes) on shared computing resources. Strategies need to be put in place to ensure the integrity of data files is maintained (e.g. use of checksum tools during file transfer, management of data permissions and ‘write’ access rights) and that a secure copy of the primary data files (FASTQ) is maintained elsewhere from ‘working copies’ which allows regeneration of results files (BAM, VCF, annotations), if this should be required.

5.2 Structured databases

The laboratory must use structured databases wherever possible. The use of spreadsheets or text files to store information is discouraged as these typically disallow satisfactory traceability or auditing of changes made. Using an appropriate LIMS consistent with quality management principles is strongly recommended.

5.3 Data storage and backup

The laboratory must establish a procedure for the storage and backup of data with particular reference to the management of raw sequence data, primary, secondary, and tertiary analysis files. The data files to be stored long-term must be identified.

The laboratory must ensure adequate data storage and backup capacity is available.
For NGS data this may require terabytes of storage to accommodate primary and secondary analyses files. Network speed to manage data transfer and access also needs to be considered.

For more information see references: 2, 3, 5, 6, 10, 11, 12, 17-37

CHAPTER FOUR: Reporting

1. Introduction

This chapter aims to establish principles and provide guidelines that should assist in the preparation of a genomics report that provides valid information relevant to the submitted sample and the disease event under investigation.

A report can only capture what is known about the sample. Therefore laboratories should establish a definition of what is their acceptable minimal metadata. Having this standardised and consistent is helpful for reporting, as well as *ad hoc* sequence sharing, database uploading, and publishing etc.

Approaches to genomic analysis vary in terms of the technology and methodology used as well as the breadth of genetic variation that is interrogated; the analysis may yield information about a single pathogen or may extend to encompass a mixture of pathogens. This presents a number of challenges for the animal health laboratory when preparing a report based on NGS data.

A key issue in reporting genomic tests is that pathogens of known or possible pathogenicity may be identified which may be unrelated to the primary clinical indication for the test. Such incidental findings are inevitable in high-resolution genomic studies utilising NGS. Orthogonal testing recommended in Chapter 2 section 4.2 should be undertaken wherever possible.

It is essential that results are reported clearly, consistently and unambiguously, using established nomenclature.

2. Key requirements of a Genomics Report

The minimum suggested content for a report is described below. The following list is not a recommendation for the structure of the report, but rather points that should be addressed. Some could be ordered to provide a one page summary of important results relevant to the disease investigation, followed by additional pages detailing the test and any further recommended tests or results that may be appropriate.

Report Details
- Reporting laboratory details
- Title of report
- Report status and report authorisation
- Issue date and time of report
- Submitter's name and address
- Unique laboratory identifier
- Animal or sample identification
- Sample type (blood, tissue and site, fluid)
- Secondary specimen identifier (referring laboratory identifier)

Test description
- Test category
- Purpose of test (e.g to assist in the diagnosis of … or the exclusion of…)
- Methodology used including confirmation by an orthogonal method (such as Sanger sequencing or real time PCR) if performed
- Limitations to test including any remaining uncertainty where it exists; this would include a narrative around any QC failures (i.e. outside acceptance criteria). It may be possible to use a confidence score.

Result summary
• Reference sequences including genome build or reference sequence version
• Variant reporting policy for the reporting laboratory that complies with relevant guidelines
Interpretive comment
• Narrative comment indicating the identified sequence
• If applicable the need for follow up or confirmatory testing should be indicated on the report.
• If this is the first detection of a disease agent, interpret cautiously seeking input and directive from the relevant state, territory and/or national veterinary authority (including CVO) initially.
• Notification may include taking results to the relevant veterinary authority (including CVO).

The laboratory should include recommendations for appropriate follow-up in reports. In situations in which further studies may be warranted (e.g. testing of other tissues, or other animals), these recommendations should be included in the test report.

3. Internal Laboratory Databases

The laboratory should establish an internal database of genomic findings. The curation of an internal laboratory database can assist with the interpretation process in the future.

4. Sharing Genomic Data

The laboratory is encouraged to submit sequence data from NGS to appropriate databases, once it has undergone adequate QA and has been approved for release by the relevant state, territory and/or national CVO(s).

It is proposed that this could be trialled focusing on specific pathogens of interest; for example avian influenza H5 and H7. This would enable the monitoring of drift, and the assessment of primer and probe relevance.

5. Summary Comment

The utility of a genomics report can be increased by the preparation of a standardised report that adheres to established guidelines. However, of equal importance is the need to ensure that those who read these reports have the necessary training to interpret these results.

CHAPTER FIVE: Information Technology (IT) Infrastructure

1. Introduction

Genomic technologies introduce complex analytical methods which require substantial bioinformatics and IT infrastructure which are not the usual domain of regulatory and/or accreditation agencies. This chapter will discuss the specific IT infrastructure issues that should be addressed by laboratories considering genomic methods.

1.1 IT process overview

Following sequencing, primary analysis (base-calling) usually occurs on the sequencing instrument and is beyond the control of the user. Secondary analysis (alignment and variant-calling) can occur on- or off-instrument. Tertiary analysis usually occurs off-instrument. Most of the sequencing manufacturers provide appropriate computing power and storage on-instrument. Where analyses occur off-instrument, it is necessary for the laboratory to consider the following issues:

- The level of processing power required to perform timely analyses
- The need to ensure data integrity during transfer across a network
- Data management and storage

Given the vast potential of genomic methods to generate genome-wide data, laboratories will need to actively consider precisely which data they will store and the retention time of that data. In some cases it may be that institutional IT departments and policies may be able to accommodate data within centralized storage facilities. However, there may be many cases where this is not possible and the problems will need to be addressed locally.

2. Data processing infrastructure and capacity

Specific requirements will vary according to the platform and style of analysis (i.e. the requirements for small-scale targeted sequencing will be different to those for whole-exome or whole-genome sequencing). The choice of computing hardware specification (i.e. type and number of CPUs or GPUs, amount of RAM, type and amount of storage platform and operating system) will be governed by the chosen software/analytical pipelines (see Bioinformatics chapter).

2.1 Computing hardware

Computing hardware should at least meet the minimum specifications of the software. Further consideration should be given to equipment which exceeds the minimum specifications in order to reduce processing time, and hence turnaround time.

The laboratory should show that the choice of hardware and software can be maintained appropriately, including installation, updates and troubleshooting.

The choice of operating system will also be largely determined by the specific software and analytical programs being used. At a minimum, a 64-bit operating system should be installed (memory allocation can be severely restricted in some/all 32-bit operating systems).

The chosen computing hardware should be shown capable of performing the required analyses and/or capable of running the chosen software using training/control datasets (i.e. datasets with characteristics consistent with clinical samples to be analysed). Datasets may be supplied by software providers, or may be obtained externally.

3. Data Transfer

Wherever possible, data should not be transferred using USB "memory sticks" or external hard drives. Consideration should be given to the use of high-speed network connections between the various components of the computing hardware.

Genomic methods have the capacity to generate very large data files. During analysis data may need to be transferred to different computing hardware (i.e. from sequencer to analytical computer or from sequencer to storage location). A speed of 1 gigabit/second (i.e. Gigabit Ethernet) 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.

Confidentiality of data should be maintained during data transfer. Appropriate steps should be taken to ensure that data corruption does not occur during transfer.

This is a significant issue, especially as files increase in size. Laboratories should implement a system to show that data transferred between different elements of their computing hardware have not been corrupted during the transfer. Consideration should also be given to similar mechanisms for data transferred to external organisations for analysis. Checksums for individual files or compressed files can be generated using a variety of software packages.

4. Data management and storage

During data generation and analysis, a series of files of varying sizes are created. In Sanger sequencing, the stored data includes unedited chromatograms ("raw" data), edited chromatograms, sequence alignments and summarized results/reports. Equivalent components can be identified within NGS pipelines, although the amount of storage required will be significantly larger.

Some genomic data may need to be repeatedly accessed and analysed over a greater period than expected in typical data retention policies (e.g. whole genome or whole exome data). Where possible, the laboratory should determine the feasibility of very long term data retention. The laboratory should develop a formal data management policy which minimizes the possibility of data loss. During analysis, genomic data will be transferred to a number of different computers for analysis and/or storage.

The laboratory should ensure that data are stored in a manner that prevents loss in the event of hardware failure (i.e. data should have redundant backup). The specific choice of computing hardware for storage purposes will vary between laboratories. The specifications of storage devices will be substantially different from the specifications of processing devices (see above). The important characteristics of storage devices will be quantity, speed and redundancy. It is suggested that "solid state" devices are inappropriate for long-term data storage as their life-span has not been empirically determined.

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.

For more information see references: 38, 39

Additional references which address quality issues in genomic sequencing generally: 40-42

Recommendations

Laying down a foundation of best practices for a new diagnostic technology is important, both at the laboratory level, and at a national level. These guidelines will evolve, as the technology improves and stakeholder expectations change. In parallel we will see more control of the process through accreditation requirements.

Unlike previous technologies, there is an emphasis on sharing testing output for the common good. Earlier in the guidelines it was proposed that the sharing of sequencing data could be trialled, with a focus on specific pathogens of interest; such as Hendra virus, *Brucella suis*, avian influenza H5 and H7. This would enable the monitoring of sequence drift, and the assessment of primer and probe relevance. It is recommended that a pilot project be initiated to trial this process.

The lack of NGS proficiency testing is recognised as a barrier to implementing external quality control in a QA system. Another recommendation is to create a resource listing proficiency testing providers, acknowledging that most of these will currently be in the human health space and possibly overseas.

Through the continued sharing of NGS experiences and pitfalls across the LEADDR network, it is hoped that this network will continue to influence the NGS landscape in animal health laboratories in Australia.

References

[Reference Number, Citation and Link to publication]

1 NPAAC standard: Requirements for the Information Communication. http://www.health.gov.au/internet/main/publishing.nsf/Content/92DFF5519A8D199ECA257BF000199DD6/$File/Reqmts%20Info%20Coms%202013.pdf

2 Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012 30(11):1033-6. PubMed PMID: 23138292. http://www.nature.com/articles/nbt.2403

3 Royal College of Pathologists of Australasia (RCPA) 3/2014 Massively Parallel Sequencing Implementation Guidelines, 2014

https://www.rcpa.edu.au/Library/College-Policies/Guidelines/Massively-Parallel-Sequencing-Implementation

4 Ellard S, Lindsay H, Camm N, et al. Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation. UK Association for Clinical Genetic Science; 2014. http://www.acgs.uk.com/media/774807/bpg_for_targeted_next_generation_sequencing_may_2014_final.pdf

5 Linderman MD, Brandt T, Edelmann L, et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Medical Genomics 2014; 7: 20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022392/

6 Standards for high throughput sequencing, bioinformatics and computational genomics. In OIE Manual of diagnostic tests and vaccines for terrestrial animals, 2017. Chapter 1.1.7. http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.07_HTS_BGC.pdf

7 Aziz N, et al. (2014) College of American Pathologists' Laboratory Standards for Next-Generation Sequencing Clinical Tests. Arch Pathol Lab Med.
http://www.archivesofpathology.org/doi/10.5858/arpa.2014-0250-CP?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed&code=coap-site

8 Rehm HL et al (2013) ACMG clinical laboratory standards for next-generation sequencing. Genet. Med. 15 733-747

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4098820/

9 Allcock RJN, Jennison AV, Warrilow D (2017) Towards a Universal Molecular Microbiological Test. J Clin Microbiol. 55(11):3175-3182.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5654900/

10 Zook JM, et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32:246-251.
https://dash.harvard.edu/handle/1/12561353

11 College of American Pathologists: Molecular Pathology Checklist 2012. Includes massively parallel sequencing. *Part of a suite of checklists available for purchase online*

https://www.news-medical.net/news/20120802/CAP-publishes-revised-version-of-molecular-pathology-checklist-with-dedicated-section-on-NGS.aspx

12 College of American Pathologists' Laboratory Standards for Next-Generation Sequencing Clinical Tests. doi: 10.5858/arpa.2014-0250-CP
http://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2014-0250-CP

13 NATA: Requirements for the Retention of Laboratory Records and Diagnostic Material (Fifth
http://www.health.gov.au/internet/main/publishing.nsf/Content/B8562E2C3D131ED8CA257BF00019153C/$File/V0.24%20Retention.pdf

14 FDA: MicroArray Quality Control project
https://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/ucm119532.htm

15 NPAAC: Requirements for Participation in External Quality Assessment (Fourth Edition 2009).
https://www.legislation.gov.au/Details/F2009L04586/Explanatory%20Statement/Text

16 Muellner et al. 'Next-Generation' Surveillance: An Epidemiologists' Perspective on the Use of Molecular Information in Food Safety and Animal Health Decision-Making. Zoonoses Public Health. 2016 Aug;63(5):351-7

https://onlinelibrary.wiley.com/doi/abs/10.1111/zph.12230

17 PHG Foundation (2011): Next steps in the sequence. The implications of whole genome sequencing for health in the UK.

http://www.phgfoundation.org/report/next-steps-in-the-sequence

18 Schrijver et al. Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology. J Mol Diagn 2012 Nov;14(6):525-40.
https://www.sciencedirect.com/science/article/pii/S1525157812001754?via%3Dihub

19 Vihinen, Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis. Hum Mutat 34:275–277, 2013.

https://onlinelibrary.wiley.com/doi/full/10.1002/humu.22253

20 Pabinger et al, A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2013 Jan 21. PubMed PMID: 23341494.
https://academic.oup.com/bib/article/15/2/256/210976

21 Analysis of in silico tools for evaluating missense variants, A summary report. National Genetics Reference Laboratory, Manchester. 2012.
http://www.ngrl.org.uk/Manchester/sites/default/files/publications/Informatics/Tool%20Analyses/Missense_Prediction_Tool_Report.pdf

22 EuroGentest, Guidelines for diagnostic next generation sequencing. 2014.
http://www.eurogentest.org/fileadmin/templates/eugt/pdf/NGS_Guidelines/EuroGentest_NGS_guidelines_2014_-_final_draft_02-12-2014_v2.pdf

23 NPAAC standard: Requirements for the Retention of Laboratory Records and Diagnostic Material.
https://www.health.gov.au/internet/main/publishing.nsf/Content/B8562E2C3D131ED8CA257BF00019153C/$File/V0.24%20Retention.pdf

24 NPAAC standard: Requirements for Medical Pathology Services.
https://www.health.gov.au/internet/main/publishing.nsf/Content/21B7F24866EF1DEECA257C2A001EC403/$File/Reqmts%20for%20Medical%20Path%20Services.pdf

25 Clinical Laboratory Standards Institute. MM09-A2: Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Approved Guideline - Second Edition (February 2014) Validation

https://www.clsi.org/standards/products/molecular-methods/documents/mm09/

26 Jennings L et al. Recommended Principles and Practices for Validating Clinical Molecular
http://www.archivesofpathology.org/doi/10.1043/1543-2165-133.5.743?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed

27 Mattocks CJ et al. A standardized framework for the validation and verification of clinical molecular genetic tests. EJHG 2010.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3002854/

28 NPAAC: Requirements for the Development and Use of In-House In Vitro Diagnostic Medical Devices (Third Edition 2014)
https://www.health.gov.au/internet/main/publishing.nsf/Content/8838AD5DB81477D5CA257BF00019166E/$File/v0.21%20Reqs%20for%20the%20Devt%20&%20Use%20of%20IVDS%2015%20Jan%202014.pdf

29 Cornish A and Guda C (2015) BioMed Research International. A comparison of variant calling pipelines using genome in a bottle as a reference.

https://www.hindawi.com/journals/bmri/2015/456479/

30 Heinrich et al. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic acids research.
https://academic.oup.com/nar/article/40/6/2426/2408990

30 Meynert et al. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 2014.
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-247

31 Meynert et al. Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 2013.
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-195

32 Chin et al. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genetics 2013.
https://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-14-6

33 Pirooznia et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics. 2014. practice. Nat Biotechnol 2012 Nov;30(11):1033-6. doi: 10.1038/nbt.2403. PubMed PMID: 23138292
https://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-8-14

34 O'Rawe, J. et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5, 28 (2013).
https://genomemedicine.biomedcentral.com/articles/10.1186/gm432

35 McCarthy, D. J. et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6, 26 (2014).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4062061/

36 ACMG Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine.
http://www.nature.com/articles/gim201530

37 CMGS Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation.
http://www.acgs.uk.com/media/774807/bpg_for_targeted_next_generation_sequencing_may_2014_final.pdf

38 The potential for cloud computing services in Australia. A Lateral Economics report to Macquarie Telecom. October 2011.
http://archive.industry.gov.au/ministerarchive2011/carr/MediaReleases/Documents/CLOUDCOMPUTINGANAUSTRALIANOPPORTUNITY.pdf

39 Financial Considerations for Government use of Cloud Computing. Australian Dept of Finance & Deregulation. Nov 2011.

https://www.finance.gov.au/files/2011/11/Cloud-Financial-Draft-Better-Practice-Guide-AGIMO-Blog.pdf

39 Weiss et al. Best Practice Guidelines for the Use of Next Generation Sequencing (NGS)
https://onlinelibrary.wiley.com/doi/full/10.1002/humu.22368

40 Next Generation Sequencing (NGS) guidelines for somatic genetic variant detection. New York State Department of Health, 2015 update.

https://www.horizondiscovery.com/reference-standards/what-are-reference-standards/quality-controlled/new-york-state-guidelines

41 Rehm et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med 2013 Jul 25. doi: 10.1038/gim.2013.92. [Epub ahead of print PubMed PMID: 23887774.]
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4098820/

42 NPAAC: Requirements for the Supervision of Medical Pathology Laboratories.
http://www.health.gov.au/internet/main/Publishing.nsf/Content/74F0211140E493B9CA257BF0001FEAB2/$File/NPAAC%20Supervision%20Document%202007%20-%20FINAL.pdf

Document Development History

Version	Author	Date	Details/Amendments
Next Generation Sequencing Guidelines -Draft	LEADDR Network	22/11/2017	Document creation
Next Generation Sequencing Guidelines - Draft	Kim Halpin	26/04/2018	Circulated to LEADDR network for comments
Next Generation Sequencing Guidelines – V1	Kim Halpin	16/05/2018	Submitted to Department of Agriculture and Water Resources
Next Generation Sequencing Guidelines – V2	Kim Halpin	25/05/2018	Incorporating Department of Agriculture and Water Resources suggestions
Next Generation Sequencing Guidelines – V3	Kim Halpin	26/02/2019	Incorporated suggestions from external reviewers (Gavin Wilke, Illumina; Karin Kassahn, SA Pathology; David Warrilow, Qld Health)
Next Generation Sequencing Guidelines – V4	Kim Halpin	06/03/2019	Incorporated suggestions from external reviewer Amy Jennison (Qld Health & Communicable Diseases Genomics Network)