PHLN
Public Health Laboratory Network

Ensuring national capacity in genomics-guided public health laboratory surveillance

A report of the Public Health Laboratory Network (PHLN) expert advisory group on whole genome sequencing

The report was prepared by the WGS Technical Advisory Group of PHLN and endorsed by PHLN in September 2014.

Version: 1.0

Authorisation: PHLN

Endorsed date: September 2014

PHLN recommendations

Background

Microbial whole genome sequencing (WGS) has the capacity to revolutionise the characterisation of pathogens in clinical and public health laboratories. High throughput WGS is rapid, and relatively cheap for the amount of information that can be extracted from the data. There has been an explosion in the number of microbial genomes being sequenced, and publicly available, providing an excellent opportunity to utilise this data for clinical and public health purposes. A key limitation to the widespread utilisation of microbial genomics in clinical and public health laboratories in Australia is the lack of robust, streamlined, simple to use bioinformatic pipelines to generate clinically meaningful data from raw WGS data. These bioinformatic pipelines are required to allow genome assembly and annotation, as well as comparative genomics to extract common typing information, and other genetic features associated with relevant clinical phenotypes.

A large array of commercially produced and in-house generated (in many cases open-source and no cost) sequence analysis tools are now available, making it difficult for laboratories commencing WGS to know which tools to use for their analyses. Many of these pipelines will adequately perform draft genome assembly, annotation and basic comparative genomics. In addition, these partially assembled genomes can then be used to determine the multilocus sequence type of the strain, and other “typing” characteristics, as well as searching for specific resistance or virulence related elements.

While some clinical and public health laboratories may have close links with research laboratories performing WGS analysis, in general the in-house tools used in the research environment may

not be appropriate for use in the clinical or public health setting. As the number of microbial genomes sequences produced in clinical and public health settings is likely to expand rapidly, key issues in the analysis of this data will be the large amount of data generated (eg. about 1 gigabyte of compressed sequence data), and the computing power required to analyse the data. In addition, some analysis tools require manual “one genome at a time” analysis, which will become impractical as the number of sequences being compared becomes large. Also some classes of analysis scale non-linearly (commonly quadratically N^2) with the number of genomes, so alternative approaches will be needed.

Main challenges in WGS harmonisation

There are four main issues in applying WGS nationally to the PHLN: data generation (sequencing), data storage (including backups), data analysis (bioinformatics) and data sharing (or distribution).

Data generation

Most laboratories within the PHLN are not routinely performing WGS of isolates. It is assumed that all labs will eventually do WGS, and do it in-house, rather than use a third party sequencing service provider; the expected volume of sequencing and security/privacy issues probably rules that out.

Data storage

In the short term, the computer that controls the sequencing instrument will have sufficient storage to hold a certain amount of data. But in the long term that data will need to be stored elsewhere, to enable analysis, and for archival purposes. The data must be backed up properly, either by replication to other PHLN nodes, or to an online backup provider.

Data analysis

Processing of data will need a reasonably powerful computer. Each PHLN could use their own system, or a single central system could be shared. These systems could be physical machines on-site, or virtual machines running on a cloud service. Cloud computing can be purchased through Amazon or Google, or an allocation applied for from Nectar the Australian Research Cloud (nectar.org.au). The software pipelines would need to be consistent across the PHLN, including exact versions of software within the overall pipeline.

Data sharing

Even on a shared system, each PHLN member can, for each sequenced isolate, choose to share their data with other members, or keep it private. The advantages of sharing are great, but not always applicable in the first instance. Online storage systems encrypt all data, and security from other users is built in.

Table 1. Suggested minimal meta-data schema for contextual WGS data sharing

Field	Question	Examples
Sample name Micro-organism Strain	What?	Blood culture Listeria monocytogenes MLST-3
Category	What?	Clinical/Environmental/Food source
Collection date	When?	September 2014
Geographic location	Where?	Postcode of patient
Characterised by	Who is the data custodian?	Public Health Microbiology, Communicable Disease, Forensic and Scientific Services, Queensland

Minimum requirements to quality control and proficiency testing

Quality control processes used in the routine library preparation and WGS are beyond the scope of this report. Whilst these steps in the next generation sequencing of pathogen genomes are important, they are largely defined by instructions from manufacturers. During library making the users are encouraged to monitor the accuracy of sequencing, to control for contamination and to assess the quality of fragmentation using established metrics. They include but are not limited to Phred quality scores per cycle and per read, proportion of unique reads, N50, pair-end distance distribution, over-represented sequences, percentage of reads mapped to reference and bias. Complete concordance of results is unlikely for next-generation sequencing technologies; however, 95-98% reproducibility is expected. This review intends to focus attention on the key QC issues relevant to the optimisation of sequencing data analysis and data comparability.

There is a growing number of commercial and 'open source' programs which are available for the mapping and assembly of short reads. However, international community remains uncertain about whether public health microbiologists will converge towards a few 'validated' pipelines. The accuracy of identifying variants depends on the depth of sequence coverage. Increased coverage improves variant calling, while low coverage increases the risk of missing variants (i.e. false negatives) and assigning incorrect allelic states (i.e. false positives). Higher coverage is required for the reliable detection of genomic sequences from potentially mixed cultures.

Table 2: Suggested minimum requirements for bacterial genome sequencing experiments

	Genome assembly	Depth of coverage	% of genome covered
Identification	Standard draft	Variable	50 ("core genome")
Characterisation	High quality draft	20-50	80-90
Microbial forensics	Coding complete	50--> >50	90-99

PHLN recommendations

PHLN should recommend that public health and microbiology laboratories work towards a coordinated approach for WGS analysis. Consideration should be given to adopting analysis pipelines and tools/naming conventions that will make WGS data analysis from Australia internationally compatible.

1. Minimum testing requirements for wet bench experiments

Laboratories should establish a minimum coverage threshold necessary to detect variants based on their diagnostic and public health questions and to ensure backward compatibility and reassessment. At least 50-fold coverage is expected for the WGS of priority bacterial pathogens of public health concern when these experiments are part of ongoing laboratory surveillance or outbreak investigation.

2. Minimum bioinformatics requirements

Laboratories investigating multijurisdictional outbreaks should nominate reference genomes for sequence analysis.

Software packages for the assembly of microbial genomes from sequencing data must be objectively reviewed and compared.

3. WGS proficiency testing

Routinely include reference material on every NGS run for the purpose of public health surveillance. PHLN encourage regular (at least twice per year) quality assurance exercises for each microbial pathogen DNA/RNA sequencing project including both spiked in (technical wet laboratory challenge) and simulated electronic (informatics challenge) sequences. PHLN should

promote the exchange of electronic reference materials between jurisdictional laboratories to ensure interoperability of WGS results.

4. Data harmonisation, sharing and exchange

PHLN facilitates the exchange of electronic reference materials between jurisdictional laboratories to ensure interoperability of WGS results. PHLN support the development of secure and sharable WGS data storage and analysis platform based on cloud computation infrastructure (RDSI and NECTAR). As much as possible metadata and WGS data should be made available for data analysis (minimum GMI metadata example) without compromising patient confidentiality and privacy. Laboratory WGS reports should be meaningful to clinicians and public health physicians, and should be back compatible with historic typing methods where possible. Laboratories should report sequencing data in a structured format, according to evolving information technology standards, to enable the deposition of genotypes into electronic public health information management systems as well as pathogen tracking over time and distance.

5. WGS competencies and training

PHLN supports development of the curriculum for training in WGS and bioinformatics for laboratory scientists and medical microbiologists.

The report was prepared by the WGS Technical Advisory Group of PHLN and endorsed by PHLN in September 2014.

PHLN Expert Advisory Group on Whole Genome Sequencing

Member	Organisation
A/Prof Vitali Sintchenko (Chair)	Centre for Infectious Diseases and Microbiology-Public Health, ICPMR-Pathology West and The University of Sydney, NSW
Prof Ben Howden	MDU, The University of Melbourne, Victoria
Dr Torsten Seemann	MDU and Monash University, Victoria
Mr John Bates	Public Health Microbiology, Communicable Disease, Forensic and Scientific Services, Queensland
Dr Amy Jennison	Public Health Microbiology, Communicable Disease, Forensic and Scientific Services, Queensland
Dr Rikki Graham	Public Health Microbiology, Communicable Disease, Forensic and Scientific Services, Queensland
A/Prof David Wiley	Pathology Queensland
Dr Ivan Bastian	SA Pathology
Dr Rodney Ratcliff	SA Pathology
Dr Adam Merritt	Pathology West, Western Australia
Dr Nadine Holmes	Centre for Infectious Diseases and Microbiology-Public Health, ICPMR-Pathology West and The University of Sydney, NSW
Dr Rosie Sadsad	Centre for Infectious Diseases and Microbiology-Public Health, ICPMR-Pathology West and The University of Sydney, NSW
Dr Grant Hill-Cawthorne	Marie Bashir Institute for Infectious Diseases and Biosecurity and School of Public Health, The University of Sydney, NSW
Professor Dominic Dwyer	Centre for Infectious Diseases and Microbiology Laboratory Services, ICPMR-Pathology West and The University of Sydney, NSW
Dr Karina Kennedy	Infectious Diseases and Microbiology, Canberra Hospital, ACT
A/Prof Barry Gatus	Microbiology Department, SEALS, NSW
Dr Louise Cooley	Infectious Diseases and Microbiology, Royal Hobart Hospital, Tasmania
Dr Debbie Williamson	Institute of Environmental Science & Research, New Zealand
Dr Richard Hall	Institute of Environmental Science & Research, New Zealand
Dr Gary Lum	Commonwealth Department of Health