Report on the GISAID Initiative Workshop on Bioinformatics held at the National Institute of Health Research and Development (NIHRD) Jakarta, Indonesia 4th - 8th November 2019

by Naomi Komadina and Dimitar Kenanov

1. PURPOSE

The purpose of the workshop held at NIHRD was to provide technical support and build the bioinformatics capacity of the laboratory staff at NIHRD, thereby enhancing influenza surveillance capabilities. Specifically, to provide knowledge and training on Next Generation Sequencing (NGS) generated sequence assembly, analysis and interpretation of influenza virus sequences.

2. BACKGROUND

The Centre for Research and Development of Biomedical and Basic Health Technology at the NIHRD, is the WHO designated National Influenza Centre (NIC) for Indonesia charged with the surveillance of influenza like illness (ILI) across the Indonesian archipelago. The main objective of influenza surveillance in Indonesia is the collection of circulating influenza virus specimens, the identification and characterisation of the viruses in house and the provision of clinical specimens and influenza viruses to the WHO Influenza Collaborating Centre in Melbourne, Australia for advanced genetic and antigenic analysis which form the basis for WHO recommendations on the composition of influenza vaccines each year.

Influenza viruses from outbreaks are identified at the NIHRD by real-time RT-PCR with some viruses undergoing full genome sequencing using Next Generation Sequencing technology. Assembly of the sequence data for analysis requires specific programs which require specialist skills in data assembly. Further-more, to understand how the influenza viruses from Indonesia fit into the world-wide influenza picture, knowledge of several specialised bioinformatics programs is required. Training in these programs is a necessity for staff at NIHRD to become skilled in these genetic analysis programs thereby providing a greater understanding of influenza characteristics within Indonesia.

3. ACTIVITIES & FINDINGS

During this 5-day hands-on training workshop, the NGS sequence assembly pipeline developed at the Singapore based Bioinformatics Institute A*Star was installed on 3 laptops. Although all participants had laptops available, a decision was made not to install the NGS pipeline on other laptops due to the amount of time, a full day, that installation took on the 3 laptops and it was felt that it was more beneficial to spend more time training participants in data assembly so that there was more hands-on time dedicated to learning how to use the program. A lecture was provided by the developer of the pipeline on the steps required to successfully assembly raw sequence data into contig sequences. This was followed by a hands-on demonstration with the program projected onto a screen so that participants could follow each required step. Participants were instructed which programs to use for preparation of reference libraries for use in sequence assembly and data analysis, checking newly assembled data, analysis of data, creating phylogenetic trees with bootstrap analysis, displaying and editing trees in FigTree and exporting trees for annotating in Power Point for use in both presentations and publications.

Apart from the NGS pipeline each attendee had access to a computer with the pre-loaded programs enabling the students to follow the tutorials in in real time gaining hands on experience. All the presentation materials, tutorials as well as FigTree, a freeware program, was made available to all the participants.

The following are the bioinformatics programs the trainees received extensive hands-on training during the workshop:

*Bioinformatics Institute A*Star NGS pipeline*: Detects both Influenza types A and B and in the case of Type A, subtypes in NGS generated data. Detection is per influenza segment. The pipeline is based on mapping reads to an index of reference genomes thus circumventing: de-novo assemblies of the reads and BLAST comparisons of assembled contigs. This method saves time by reducing the complexity of the task at hand. The program undertakes several steps to achieve its goal:

- raw reads quality check
- trimming the reads
- mapping trimmed reads to References Index
- specifies statistics for mapping – mapped reads, coverage
- choosing templates based on the statistics
- remapping the reads to the chosen templates only
- calling variants
- specifies statistics for re-mapping – mapped reads, coverage
- creates consensus sequences based on variants and coverage
- aligns consensus sequences to templates
- provides final statistics

The outcome provides several files containing results of the analysis, a summary table including statistics for the detected types/subtypes and consensus sequences per segment. Participants received training in all stages required to successfully output consensus sequences from their NGS raw data.

The GISAID Initiative's EpiFlu™: An influenza database, which formed the basis for the training workshop, as it provides the largest collection of influenza isolates available for use from both human and animal sources. Participants were taught how to use EpiFlu™ to search for genetic sequence data deposited by NIHRD and WHO CCs, generated from viruses originating from Indonesia and from neighbouring regions. Instructions were provided on how to create worksets, such as reference data sets for use in NGS sequence assembly, use of worksets for inter-institutional research collaborations, how to search for viruses of interest using available parameters such as specimen dates, type, subtypes, region & countries. Training also included instructions on visualization of data made available in GISAID using the following analysis tools:

FluSurver: A mutation analysis program which provides information & statistics on the prevalence of these mutations. Information on drug resistance is also provided along with visualisation tools of where the mutations appear in the HA & NA protein molecules of influenza.

NextFlu: A real time visualisation tool of influenza phylogeny depicting influenza evolution. Phylogenetic trees are annotated with branch amino acid substitutions as well as Clade designations. Used to select clade reference viruses to aid in country data analysis of influenza.

GENEiOUS: A multifaceted sequence data analysis program. This program was used to prepare reference data sets for use in NGS sequence assembly. Participants also received instruction on how to check the quality of the assembled sequence data, importing & exporting data, how to align sequence data, translate to proteins and phylogenetic analysis using maximum likelihood with bootstrapping analysis.

FigTree: A program which visualises phylogenetic trees. Instruction on how to import & export trees, root the trees, visualise bootstrapping values as well as preparing the trees for use in publications or presentations. Instruction on how to export trees in a format for use in Power Point was provided.

Power Point: Used for generating presentations. Instruction was provided on how to import phylogenetic trees into power point. How to annotate the trees as a final product for use in presentations and publications.

4. CONCLUSIONS & RECOMMENDATIONS

As this group consisted of seven participants, all had similar knowledge of the processes involved in sequence data assembly and that of sequence data analysis with most gaining a good grasp of the workflow patterns required to be able to produce meaningful results as well as a good understanding of the analysis programs used. Not all participants were able to attend all sessions due to other commitments. Some of the participants had better quality sequence data available and could successfully assemble their sequence data and take that data through the whole analysis process. Two participants from NIHRD were identified as having gained the strongest grasp of the material presented. These would benefit from further in-depth training which would re-inforce the complex theory & methodology presented and in turn these trainees could further act as local trainers thereby further enhancing the sequence assembly and influenza data analysis skills of the staff at NIHRD.

5. ACKNOWLEDGEMENTS

Dr. Siswanto MHP, DTM, Director General, NIHRD for hosting the workshop.
Dr. Vivi Setiawati for facilitating the workshop at the NIHRD.
The GISAID Initiative for the provision of workshop funding.
Dr. Sebastian Maurer-Stroh, Deputy Executive Director (Research) Bioinformatics Institute, A*STAR, Singapore.
Professor Ian Barr, Deputy Director, WHO Centre for Reference and Research on Influenza, Melbourne, Australia.

Instructors:
Dimitar Kenanov, Bioinformatics Institute, A*STAR, Singapore.
Naomi Komadina, WHO Centre for Reference and Research on Influenza, Melbourne, Australia.