Public Health England

Bacterial and viral NGS analysis in a public health agency using Galaxy

Ulf Schaefer, Anthony Underwood, and Jonathan Green

Advanced Laboratory- and Bio-Informatics, Microbiology Services, Public Health England, 61 Colindale Ave, London NW9 5EQ, United Kingdom

INTRODUCTION

Public Health England (PHE) is home to the United Kingdom's national microbiology reference laboratories and deals with the surveillance and control of infectious disease. Assays for the investigation of selected pathogenic bacteria and viruses are being migrated from traditional wet lab based methodologies such as Multiple Loci VNTR Analysis to methods based on Next Generation Sequencing (NGS) data. This development in part of a wider paradigm shift in public health and clinical microbiology to employ Whole Genome Sequencing (WGS) on pathogens for diagnostics and disease control (Didelot et al., 2012; Köser et al., 2012; Török et al., 2012). Organisms primarily sequenced at PHE are *Salmonella enterica*, *Staphylococcus aureus* (Fig. 1), *Escherichia coli*, *Streptococcus pyogenes* (group A strep) and *Streptococcus pneumoniae* as bacterial pathogens and HIV, Hepatitis B, C, and E, Measles, Influenza, and MERS-CoV as viral pathogens.

Apart from the set up of an NGS service and automated analysis of these priority organisms, one of the key challenges in the management of this paradigm shift in public health is to enable microbiologists and epidemiologists with little to no bioinformatics knowledge and training to interact with and derive scientific value from NGS data. We maintain a local installation of Galaxy in an attempt to address this challenge. This local installation houses all specialized software required for public health microbiology and phylogenetics. Furthermore it provides bespoke workflows for standard analyses regularly employed in outbreak investigations, such the creation of a SNP tree from multiple viral or bacterial NGS samples. In addition to an overview of our hardware and software setup, this presentation will highlight

i) An example of a public health specific workflow that can be used for routine reference microbiology services.

ii) Some of the soft issues around employing Galaxy in this context, such as user acceptance, training, and support.

METHODS

Clinical samples for all cases of a notifiable infectious disease are submitted to the national reference laboratory at PHE. Samples from the priority organisms are routinely submitted for WGS after DNA isolation. Automated computational pipelines perform routine analyses such as quality control, trimming, Multilocus sequence typing (MLST) (Maiden et al., 1998) or serotyping on these samples.

In order to facilitate the generation of further scientific or public health value from these samples we run a local Galaxy installation. Our Galaxy server is an 8 core, 32GB virtual Centos 6.5 server running 3 handler and 3 web Galaxy processes. Apache proxies between the webservices. The Galaxy PostgreSQL database is installed on a separate server. Jobs for most tools are dispatched to a HPC compute cluster with 256 cores available, running Univa Grid Engine 8.1.5. The shared file system is Lustre, a parallel distributed high performance file system.

We maintain 3 user groups, Administration, Bioinformatics and Regular users and enforce a disc quota for the latter two of 250GB and 100GB respectively. Also, jobs dispatched to the HPC cluster through Galaxy are submitted to a dedicated queue, which has a below average priority and is limited to execution on 25% of the available cores. These limitations are necessary, because the available computing hardware is to be used for routine pipeline tasks and the availability of sufficient resources for these tasks must be ensured at all times.

We also maintain a ProFTPd server to aid users with data uploads.

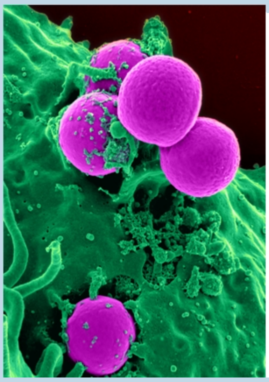

Figure 1: Human neutrophil ingesting Methicillin-resistant *S. aureus* (MRSA) [source: Wikipedia.org]

RESULTS

One of the most common questions in public health microbiology is about the phylogenetic relationships between the samples in an outbreak. This type of information advises decisions about public health interventions and clinical management (e.g. Brooks et al., 2013). While for a trained Bioinformatician the associated work can easily be done with a number of command line tools, the challenge is to enable wet lab biologists to do this work independently. We developed a standard procedure for doing this type of analysis using reference mapping with bwa-mem, three instances of GATK and in-house software to annotate, combine and filter VCF files (Fig. 2). We have published this workflow and due to the minimal parameters that are required users are able to use it on their own.

In a next step, it is fairly easy to create a multiple sequence alignment file from the resulting VCF compendium files and from that a character based phylogenetic tree. For ongoing outbreaks we have stored the VCF compendium files from previously sequenced cases in shared data libraries, so that lab biologists can quickly import the data and ascertain the evolutionary relationship between past cases and a new case. This has proven to be especially useful for scientists investigating hospital acquired infections.

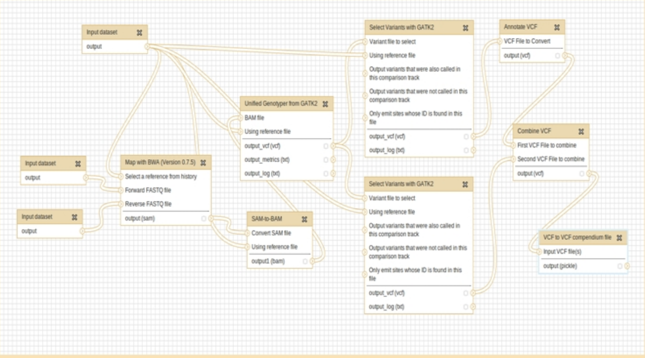
Input dataset
output
Input dataset
output
Input dataset
output
Map with BWA (Version 0.7.5)
Select a reference from history
Forward FASTQ file
Reverse FASTQ file
output (sam)
Unified Genotyper from GATK2
BAM file
Using reference file
output_vcf (vcf)
output_metrics (txt)
output_log (txt)
SAM-to-BAM
Convert SAM file
Using reference file
output1 (bam)
Select Variants with GATK2
Variant file to select
Using reference file
Output variants that were also called in this comparison track
Output variants that were not called in this comparison track
Only emit sites whose ID is found in this file
output_vcf (vcf)
output_log (txt)
Select Variants with GATK2
Variant file to select
Using reference file
Output variants that were also called in this comparison track
Output variants that were not called in this comparison track
Only emit sites whose ID is found in this file
output_vcf (vcf)
output_log (txt)
Annotate VCF
VCF File to Convert
output (vcf)
Combine VCF
First VCF File to combine
Second VCF File to combine
output (vcf)
VCF to VCF compendium file
Input VCF file(s)
output (pickle)

Figure 2: Standard workflow for SNP calling in bacterial WGS reads

Another workflow that exemplifies the use of Galaxy within our organisation deals with the subtyping of HIV samples. Viral sequencing reads for the Reverse transcriptase (RT) region of the HIV genome for a number of patients are uploaded and mapped against the complete genome of the Human immunodeficiency virus type 1 (HXB2). Subsequently an in house tool is applied that was specifically developed to determine the consensus sequence and minority variants in viral samples. The consensus sequences are concatenated with a range of RT reference sequences. These are then aligned with mafft and a phylogenetic tree is derived using FastTree (Fig. 3). Depending on how the patient samples group with the references, a subtype of HIV can be derived for the patient. In Fig.3 patients 3, 4, and 5 are clearly subtype B, while we can be less sure about patient 2 being subtype C. The result for patient 1 is ambiguous.

Another question that also occurs very commonly in our specific context is about the presence or absence of one or more specific genes in a set of NGS reads. An example is the determination of phase variation in *Salmonella enterica*. *S. enterica* assemble different types of Flagellin in this way, which has important consequences for virulence and immune response evasion of the pathogen. This type of question can easily be answered using a workflow containing a de-novo assembly of a bacterial genome followed by a BLAST step against a given set of reference genes.

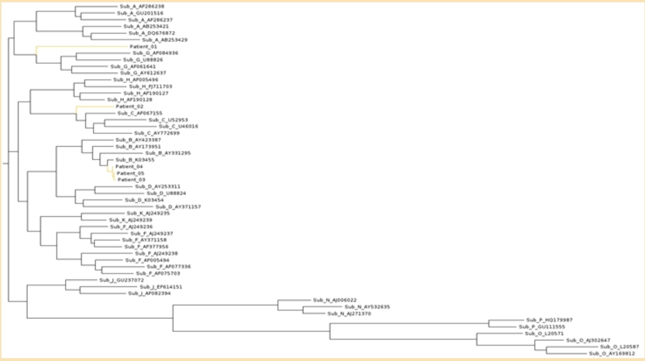

Figure 3: Subtyping tree for 5 HIV NGS samples

DISCUSSION

The challenges associated with this type of application of Galaxy are less of a technical and more of a human resource nature. Users with computing skills that are at the level of using email and office applications in Microsoft Windows need to have the capacity to conduct Bioinformatics analyses. Without the use of Galaxy this goal would not be achievable, since the training required to teach this category of users the use of a command line environment would be insurmountable. Even with the use of Galaxy, challenges remain. A substantial training effort remains and users need to invest time and be motivated to learn the system. Experienced staff are sometimes hesitant to embrace new technologies.

We have found it most useful to arrange a series of weekly seminars where users are mentored on an individual basis by a Bioinformatician in running projects through Galaxy. These projects need to be chosen in correspondence with the user to highlight the usefulness of Galaxy in particular and the usefulness of NGS data in general to their specific field of interest. Approaches that require less man power have unfortunately not been successful.

Conveying an understanding of the underlying complexities especially in terms of scalability is an open issue. Users often demand that processes that they have successfully applied to a small number of samples, will automatically work in the same way with sample numbers several orders of magnitude higher.

CONCLUSIONS

- Galaxy has proven to be a valuable facilitator for laboratory staff to use microbial genomics in a public health setting.
- Workflows prepared by Bioinformaticians that require minimal parameterisation and that can be applied to answer routine public health questions a proving especially useful.
- Common analyses in public health, such as the creation of a SNP tree from multiple NGS samples can be conveniently answered using our local Galaxy instance.
- The biggest concerns revolve around issues of education, training and user acceptance.
- The investment in terms of Bioinformatics man-hours that are required to enable a single user to be productive using Galaxy independently prove to be significant.

REFERENCES

Brooks JI, Niznick H, Ofner M, Merks H, Angel JB. Local phylogenetic analysis identifies distinct trends in transmitted HIV drug resistance: implications for public health interventions. BMC Infect Dis. 2013 Oct 30;13:509. doi: 10.1186/1471-2334-13-509.

Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012 Sep;13(9):601-12. doi: 10.1038/nrg3226.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998 Mar 17;95(6):3140-5.

Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012;8(8):e1002824. doi: 10.1371/journal.ppat.1002824.

Török ME, Peacock SJ. Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory--pipe dream or reality? J Antimicrob Chemother. 2012 Oct;67(10):2307-8. doi: 10.1093/jac/dks247.