Experiences in implementing large-scale biomedical workflows on the cloud: Challenges in transitioning to the clinical domain

Sehrish KANWAL[a,1], Andrew LONIE[a], Richard O. SINNOTT[a]
Charlotte ANDERSON[b]
[a] Department of Computing and Information Systems
[b] Victorian Life Sciences Computation Initiative
The University of Melbourne

Abstract. The sequencing of the human genome has brought about many opportunities and challenges for the realisation of personalised health. Whilst researchers are able to analyse and derive results that can be published in journals, the rigor required in moving from a research setting to a clinical setting increases dramatically. Workflows represent one way in which analysis can be defined reflecting the many steps involved in analysing genomics data that in principle can be repeated by others. The Cloud also provides ways to re-establish the software environment for enactment of workflows required for data-intensive genomic analysis. However the challenge of what is the best analytical workflow remains. This paper explores this issue through systematic exploration of a range of biomedical workflows on the NeCTAR Research Cloud and the resultant evidence in diversity of possible workflows and their results. The challenges for the future acceptance of genomics workflows in the clinical domain are discussed.

Keywords. Workflows, genomics, data-intensive, Cloud, diversity, clinical domain

1. Introduction

Since the completion of the Human Genome Project, genomics has emerged as a key focus for the biomedical and clinical community to help realise the vision of personalised health and establish the basic biological understanding of a multitude of diseases [1]. Improvements in DNA sequencing technologies [2] regarding cost, accuracy and speed, have aided in the identification a range of differences (variants) in the genetic makeup of individuals and populations. Through efforts such as the 1000 Genomes project [3] and approaches for analysis of NGS data [4], a progression of approaches for variant discovery are now being used by researchers. Nekrutenko and Taylor [5] discuss important issues such as accessibility, interpretation and reproducibility for the analysis of next generation sequence (NGS) data including whole genomes and exomes, and propose solutions for future developments. A large number of computational tools and workflow platforms have been developed to support analysis of NGS data e.g. Galaxy [6], Taverna [7], Omics Pipe [8] and Mercury [9]. However adapting and extending already built pipelines requires considerable computational knowledge and expertise.

The major challenge that lies ahead is the many ways in which increasingly large and diverse biological sequence datasets can be analysed and interpreted [10]. As the amount of data from these technologies piles up, considerable informatics expertise and tools are needed to store, analyse and interpret these data to attain accurate knowledge that can be translated into clinical practices. It is very important to understand the critical aspects that ensure workflow implementations that are consistent enough to be reproduced by others and ultimately be translated into clinical settings. The sustainability of clinical genomics research requires the plausibility of reproducibility of results to be as easy as data production. We need to fill this gap by proposing and implementing practices that can ensure repeatability, reproducibility, confirmation and ultimately extension of others work. This research aims to answer these questions by demonstrating end-to-end reproducible clinical genomics analysis workflows on the National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) Research Cloud. A workflow is generally defined as a reproducible process composed of a set of coordinated tasks executed using software. Workflows have different purposes including collecting data from various data sources, processing/transforming data to compute results and enabling interoperability between different tasks. We identify that the different workflows can indeed be re-established and re-enacted on the Cloud, however the choices in the workflows that are selected impacts directly upon the repeatability of scientific evidence. We provide illustrations of this diversity.

2. Experimental Case studies

The experimental case studies conducted can be divided into two main categories: workflows created as part of the NeCTAR funded endocrine genomics virtual laboratory (endoVL – www.endovl.org.au) [11] using the Galaxy workflow environment and workflows [12] developed through the Melbourne Genomics Health Alliance (MGHA- www.melbournegenomics.org.au) project using the Bpipe environment [13].

2.1. EndoVL Project

The endoVL project was an initiative to establish an Australia-wide endocrine genomics virtual laboratory [11]. A major motivator and use case for developing endoVL was to identify, store and search for genetic variants in patients with endocrine-based disorders. A range of case studies was conducted as part of endoVL focused on analysis of exome data from patients with a rare disorder: disorder of sex development (DSD). Sequencing was undertaken at the Australian Genome Research Facility (AGRF) sequencing facility following the Illumina TruSeq exome capture using the Illumina HiSeq2000 platform to generate 100bp paired-end reads. Three well established bioinformatics groups in Australia (Group A, Group B and Group C) participated in this study. The identity of these groups is anonymised here deliberately due to the differences of the results found and the potential for misinterpretation of the results (e.g. which group was better than another). It was essential to note that all of these groups *independently* undertook their own analysis of the data.

2.1.1. Workflows created by the three groups

The endoVL project explored the different approaches taken for the bioinformatics analysis of NGS data by the different groups. The three groups initially used their own in-house bioinformatics data processing pipelines. This resulted in a diversity of the independent approaches and radically different interpretation of the data – specifically the numbers of variants found. The diversity of results was presented in [14]. The three groups were subsequently requested to use a common bioinformatics analysis environment to analyse single exome data on six patients with DSD. The analysis environment was made accessible through the Genomics Virtual Laboratory (www.genome.edu.au) running on the NeCTAR Research Cloud. For the second case study all groups had to use this resource, which was based around the Galaxy workflow environment [15]. Galaxy allows saving analysis histories as documentable entities that can be used as data objects to run on the same Galaxy instance or even on different machines. The groups came up with three different workflows and results despite using this common analysis platform as discussed in the next section.

2.1.2. Results

To identify and interpret variants in a specific set of genes known to be involved in DSD, the groups were given a defined list of genes used to identify variants in these specific genes (Table 1). The results of this analysis had differences with >50% concordance for single nucleotide variants (SNVs) among the three groups (Table 1). Also, the transition/transversion (Ti/Tv) ratio was quite accurate for the variants called by the three groups (~2.0). In this case, however, every detail about the workflow was recorded and the results were different but far more overlapping than the original independent "in-house" approaches that were taken [14].

Table 1. Total number of variants called/Variants called based on subset genes list

Sample	**Total Variants/Variants in subset gene region**			**Common (%age concordance) - Ti/Tv ratio of SNVs only**
	Group-A	**Group-B**	**Group-C**	
BELS1	44306/705	64766/778	80748/1035	524 (68%) – 2.03
BELS2	51800/657	53298/609	81144/1005	483 (75%) – 1.91
BELS3	57556/755	54915/662	83263/1074	536 (76%) – 2.08
NLDS1	51993/653	50164/587	75079/917	484 (78%) – 2.18
NLDS2	55929/738	53682/648	79756/1037	550 (79%) – 2.11
NLDS3	54980/692	53108/604	80827/1018	499 (75%) – 2.02

As there was no truth set available for the DSD patient data under analysis, it was not possible to determine which workflow identified the "correct" variants. This also shows the current heterogeneity of computational genomics analysis with the absence of agreed and acceptable approaches for data analysis and discovery. Systematic approaches for workflow definition, evaluation and re-use are essential when moving into the area of clinical diagnostics and treatment.

2.2. Cpipe Project

The heterogeneity in the previous analysis process motivated us to work towards an enhanced workflow, which is now used by clinicians at the MGHA. The MGHA aims to integrate clinical research and genomic medicine for the betterment of patients. Currently MGHA are using a targeted bioinformatics pipeline: Cpipe. Cpipe is the clinical version of Bpipe [13] and is used to carry out exome sequence analysis of human samples on the Victorian Life Sciences Computational Initiative (VLSCI) HPC cluster (https://www.vlsci.org.au). Cpipe is an automated and flexible pipeline that can help produce reproducible and precise results at individual or population-wide scale.

2.2.1. Cpipe on the Cloud

The setting up of Cpipe on a HPC Cluster is a complex process that can only be performed with the help of people involved in developing and running the pipeline or by an experienced bioinformatician that is aware of the set-up of the VLSCI cluster. However, it is also essential that the results of a genomic analysis and also the steps involved in an analysis can be independently repeated by others, especially when moving into clinical settings. This is a challenge with Cpipe on a HPC system.

To tackle this, Cpipe was provisioned on the NeCTAR Research Cloud using snapshot technology to make this pipeline easily accessible and usable for other researchers. New users can use this snapshot to launch new GVL instances that can communicate with the Object Store to download the Cpipe tar file and reproduce (also, if desired, extend) the environment used.

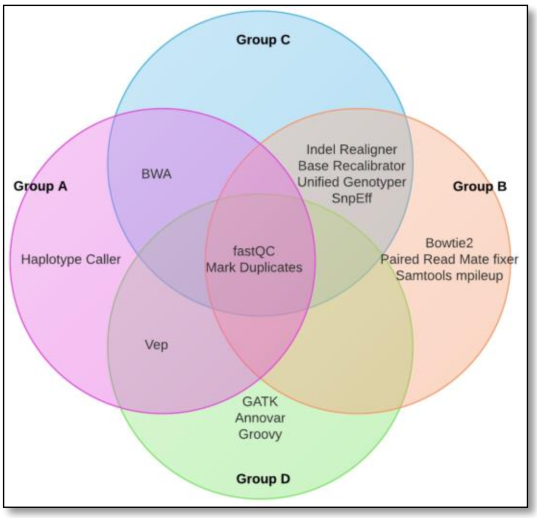
This complexity of installation and configuration of complex workflows will always be required when dealing with complex genomics datasets that comprise multiple tools that need to be coupled together. However the Cloud provides the capability to easily repeat the exact environment and have others use this immediately through the Software as a Service (SaaS) paradigm.

2.2.2. Comparison of analysis using Cpipe (Group D) and the three pipelines (based on galaxy) from endoVL project

To compare and contrast between the four pipelines (three from endoVL project – section 2.1.1 and fourth from Cpipe project –section 2.2), the Genome in a Bottle dataset NA12878 [16] was used to analyse and validate pipelines on Cloud because it has been extensively studied and analysed to establish a validated truthset. The truthset contains the variants that are known to be present within NA12878 dataset. Hence workflows should ideally identify these variants that are *known* to occur. The NA12878 dataset was used with the four workflows on Cloud and the results were compared with the truthset for NA12878. The Venn diagram of tools used by the four groups is shown in Figure 1. The diagram demonstrates the differences in the preference for tools between the four pipelines. Group-D used most of the analysis steps recommended by GATK [17], whereas the other three pipelines used an edited version of the same recommendation based on their personal experience and choices. For example Group-A and Group-C used BWA as an alignment tool whereas Group-B used Bowtie2. This difference in the preference for tools resulted in variable results (explained in the next section). This experiment actually helped to systematically explore a range of biomedical workflows on the NeCTAR Research Cloud and the resultant evidence in diversity of possible workflows and their results

3. Results

The highest percentage (95%) of overlap with truthset was detected for Group-A, followed by Group-D (94%) as shown in the Table 2. Table 3 summarises the sensitivity, specificity and false discovery rate for variants produced by the four groups. The sensitivity value signifies the percentage of correctly identified variants (actual positives); the specificity value signifies the percentage of correctly rejected variants (negatives) and the false discovery rate signifies the incorrectly identified variants. The highest values for sensitivity and specificity (95% and 66% respectively) are observed for Group-A. The sensitivity value for variants predicted by Group- C and Group-D is same (95%), whereas specificity value for Group-D (59%) is better than Group-C (50%). The preference of workflow with the high sensitivity or specificity value will depend on the clinicians and final use of workflow. However, the systematic evaluation of workflows to gain an insight into these values (i.e. sensitivity, specificity and false discovery rate) is important to be considered if these workflows are being finally deployed to analyse patient's data



.

Figure 1. Comparison of tools used for alignment, variant calling and quality control by the three groups

Table 2. The total number of variants found by each group and the percentage overlap with the truth set

Group	Total number of variants	Overlap with truthset	Percentage
A	26124	24937	95
B	22949	21261	93
C	26615	24874	93
D	26256	24807	94
E (truthset)	26159		

Table 3. The sensitivity, specificity and false discovery rate for variants from each group

Group	TP	TN	FP	FN	Sensitivity (%age)	Specificity (%age)	False Discovery Rate (FDR) (%age)
A	24937	2312	1187	1222	95	66	5
B	21261	1821	1688	4898	81	52	7
C	24873	1757	1742	1286	95	50	7
D	24807	2050	1449	1352	95	59	6

4. Conclusion

The literature on studies involving use of NGS and other technologies such as microarrays shows that there is absence of over-all agreement on how data should be analysed and presented. This research demonstrates that as there is (and continues to be!) an enormous number of tools and data processing workflow systems being developed, however there is a little detailed assessment of the application of these to establish best practice and specifically, recommendations and practices that ensure that they meet the rigorous requirements demanded when applied (translated) into clinical settings. Moreover, this research also shows that the results vary across different workflows and these need to be verified either by wet labs or clinicians in order to be successfully translated into clinical settings. Can Cloud help tackle significant issues imposed by the ever increasing genomics datasets? Virtualisation and Cloud technologies can certainly help with many of the issues of data-intensive experiments, without imposing significant overheads.

As a next step, our research aims at designing strategies to explore the workflows with other disease datasets e.g. diabetes and then analysing the results. The quality assurance and importantly the use and translation into clinical settings have major implications for personalised health more generally. The systematic analysis needed to aid evaluation and comparison of workflows is an essential activity to validate any conclusions and this is especially so in the clinical (as opposed to the research) domain since the application of results in clinical/hospital settings will require clinical validation (and have consequences for the patients).

References

1. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nature Reviews Genetics, 2011. **12**(11): p. 745-755.
2. Metzker, M.L., *Sequencing technologies—the next generation.* Nature Reviews Genetics, 2009. **11**(1): p. 31-46.
3. Siva, N., *1000 Genomes project.* Nature biotechnology, 2008. **26**(3): p. 256-256.
4. Nielsen, R., et al., *Genotype and SNP calling from next-generation sequencing data.* Nature Reviews Genetics, 2011. **12**(6): p. 443-451.
5. Nekrutenko, A., ; Taylor, J., *Next generation Sequencing data interpretation: enhancing reproducibility and accessibility.* Nature, 2012. **13**.
6. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
7. Wolstencroft, K., et al., *The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.* Nucleic acids research, 2013: p. gkt328.
8. Fisch, K.M., et al., *Omics Pipe: a community-based framework for reproducible multi-omics data analysis.* Bioinformatics, 2015: p. btv061.
9. Reid, J.G., et al., *Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline.* BMC bioinformatics, 2014. **15**(1): p. 30.
10. Baker, M., *Next-generation sequencing: adjusting to data overload.* nature methods, 2010. **7**(7): p. 495-499.
11. Sinnott, R.O., et al., *Development of an endocrine genomics virtual research environment for Australia: building on success*, in *Computational Science and Its Applications–ICCSA 2013*. 2013, Springer. p. 364-379.
12. Sadedin, S.P., et al., *Cpipe: a shared variant detection pipeline designed for diagnostic settings.* Submitted to: Genome Medicine, 2015.
13. Sadedin, S.P., B. Pope, and A. Oshlack, *Bpipe: a tool for running and managing bioinformatics pipelines.* Bioinformatics, 2012. **28**(11): p. 1525-1526.
14. Kanwal, S., et al., *Challenges of Large-scale Biomedical Workflows on the Cloud – A Case Study on the Need for Reproducibility of Results*, in *28th IEEE International Conference on Computer Based Medical Systems*. 2015: Sao Paulo, Brazil.
15. Afgan, E., et al., *Harnessing cloud computing with Galaxy Cloud.* Nature biotechnology, 2011. **29**(11): p. 972-974.
16. Zook, J. *New high-confidence na12878 genotypes integrating phased pedigree calls*. 2014; Available from: https://sites.stanford.edu/abms/content/new-high-confidence-na12878-genotypes-integrating-phased-pedigree-calls.
17. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome research, 2010. **20**(9): p. 1297-1303.