Experiences in Implementing Large-scale Biomedical Workflows on the Cloud: Challenges in Transitioning to the Medical Domain

Sehrish Kanwal, Andrew Lonie, Richard O. Sinnott and Charlotte Anderson

Department of Computing and Information Systems,
The University of Melbourne, Australia
Email: skanwal@student.unimelb.edu.au

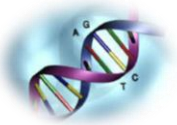
Content

- Introduction
- Current Research
- Case Study
- Details of the Case Study
- Results
- Conclusion

Introduction

- Generation of tera/petabytes of genomic data at an unprecedented rate and at increasingly reduced cost
  - Bioinformatics workflows (a set of coordinated tasks)

    → analysis of genomics data

    
  - The ultimate objective

    → personalised/precision medicine

- Variant discovery process

Sample DNA → computational tools → Reference sequence

Introduction (cont.)

- Computational tools and workflow platforms such as Galaxy, Taverna, Omics Pipe and Mercury

- Computational knowledge and expertise

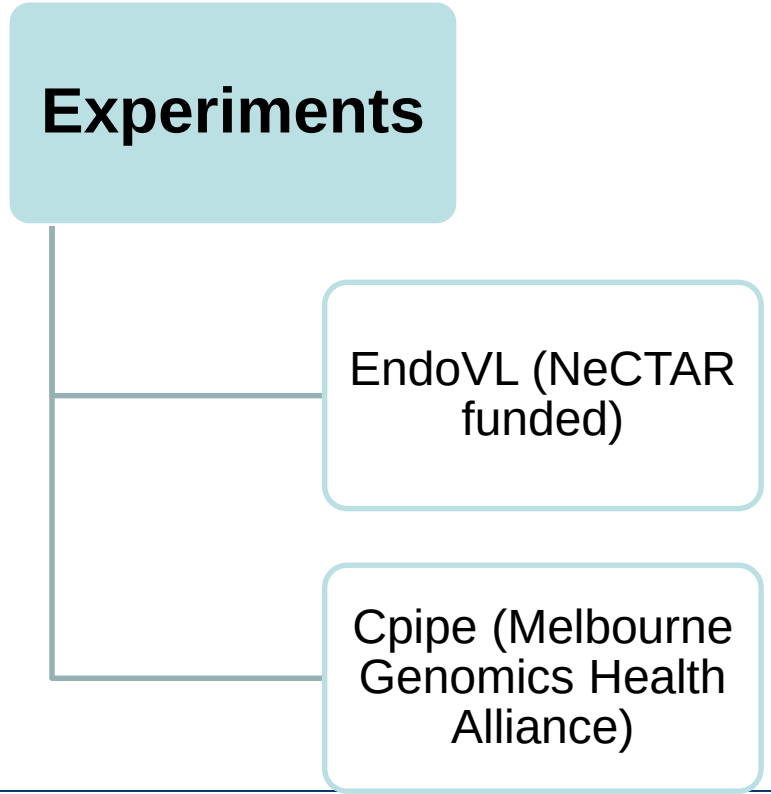
  Store, Analyse and interpret data

- Sustainability of clinical genomics research requires the plausibility of reproducibility of results to be as easy as data production

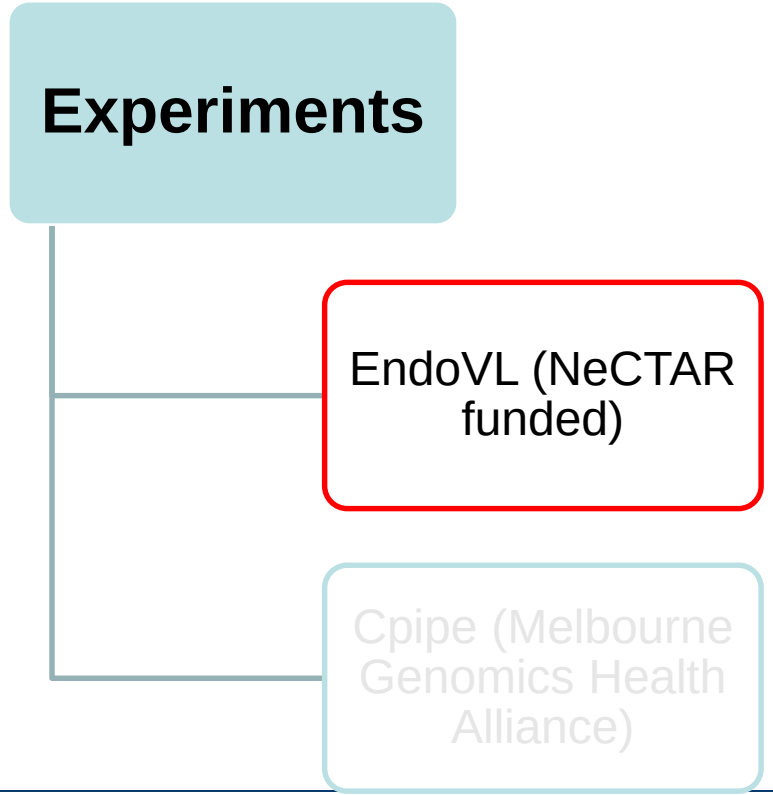
Current Research

- We need to fill this gap by proposing and implementing practices
- Aim of this research is to demonstrate end-to-end reproducible clinical genomics analysis workflows on the research Cloud
- 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
- Illustrations of this diversity

Case Study (Overview)



Case Study (Overview)



Case study - EndoVL

- Initiative - Endocrine Genomics Virtual laboratory (Sinnott et al., 2013)
- Three well established bioinformatics analysis groups across Australia participated in the study

- Analyse the exome data from 15 patients
- Used completely different approaches and cameup with widely varying results

- Focused on improving bioinformatics processes
- Use a common platform on NecTAR Cloud i.e. Galaxy (Goeks et al., 2010)

Results

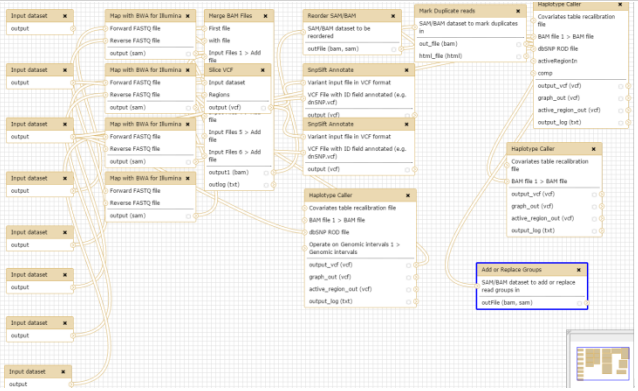
Table 1. Total number of variants called/Variants called based on subset genes list

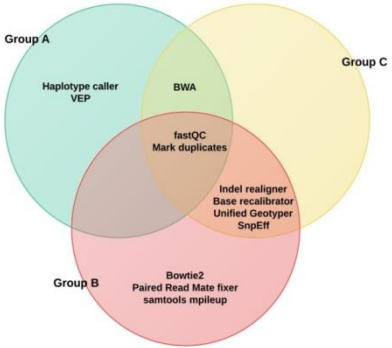
Sample	Total Variants/Variants in subset gene region			Common (%age concordance) - Ti/Tv ratio of SNVs only
	Group-A	Group-B	Group-C	
BELS1	44306/705	64766/778	80748/1035	524 (68%) – 2.03
BELS2	51800/657	53298/609	81144/1005	483 (75%) – 1.91
BELS3	57556/755	54915/662	83263/1074	536 (76%) – 2.08
NLDS1	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

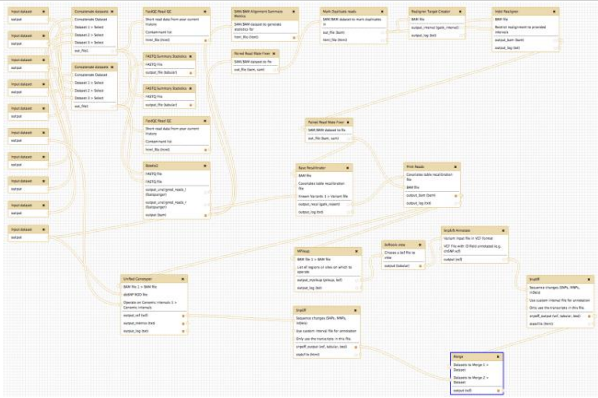
Results (cont.)

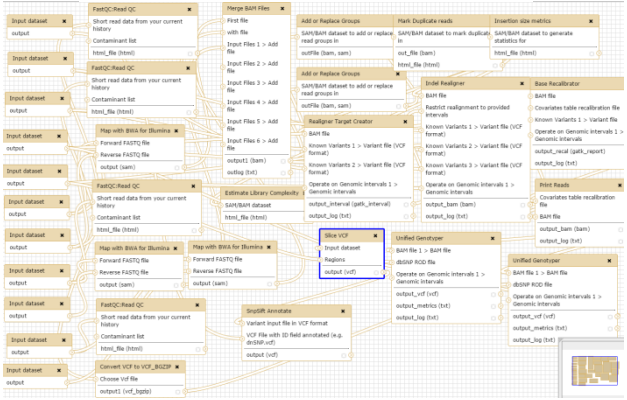
- No truth set available for the DSD patient data under analysis
- The current heterogeneity of computational genomics analysis
- Systematic approaches for workflow definition, evaluation and re-use are essential when moving into clinical diagnostics and evaluation

Overview of Workflows from the Three Groups

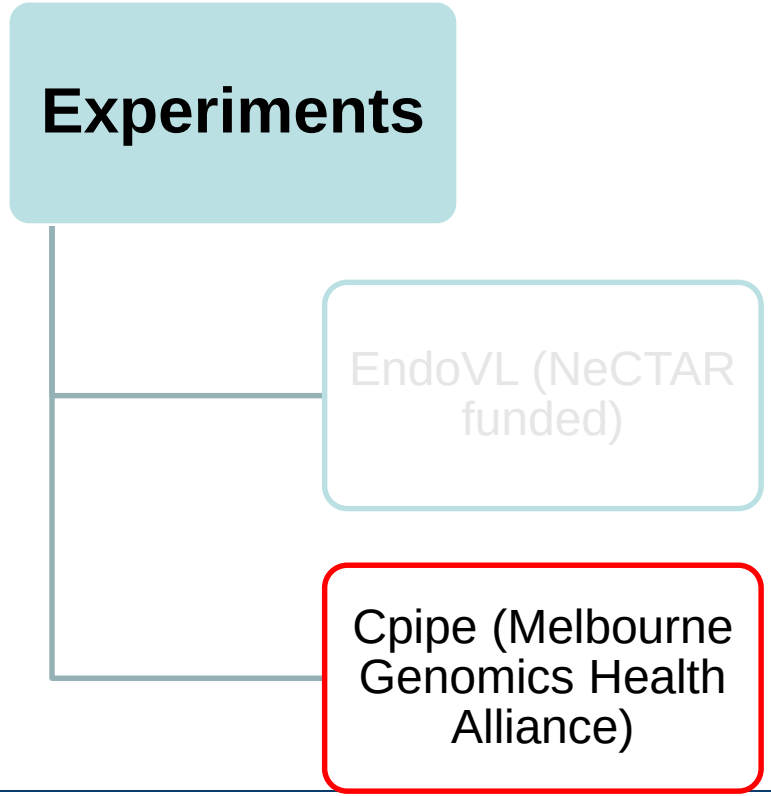








Case Study (Overview)



Case Study (Cpipe)

- Heterogeneity in the previous analysis
  - an enhanced workflow, which is now used by clinicians at the MGHA
- MGHA aims to integrate clinical research and genomic medicine
- Cpipe – a targeted bioinformatics pipeline
  - Reproducible and precise results at individual or population-wide scale

Cpipe on Cloud

- Setting up of Cpipe on a HPC Cluster is a complex process
- It is also essential that a genomic analysis can be independently repeated by others , especially when moving into clinical settings
- Cpipe was provisioned on the NeCTAR Research Cloud
- Complexity of installation and configuration of complex workflows will always be required
  - Cloud provides the capability to easily repeat the exact environment (Software as a Service (SaaS) paradigm)

Comparison of Analysis using Cpipe and endoVL Workflows

- The Genome in a Bottle dataset NA12878 was used to analyse and validate pipelines on Cloud

- NA12878 has been extensively studied and analysed to establish a validated truthset

- Workflows should ideally identify these variants that are *known* to occur

Results

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		

Correctly identified

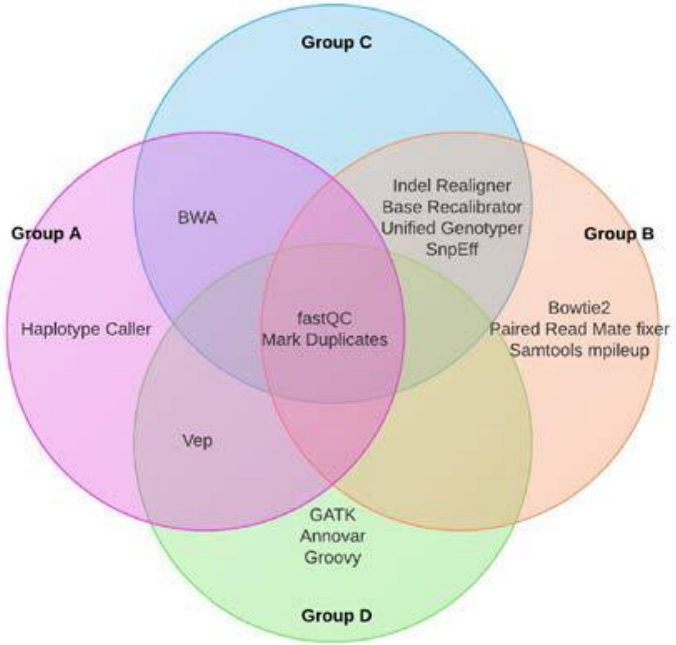
Correctly rejected

Table 3. The sensitivity, specificity and false di… …nts …

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

Results (cont.)

- Preference of workflow with the high sensitivity or specificity???
  - Depends on the clinicians and final use of workflow
- However, the systematic evaluation of workflows is important if these are being finally deployed to analyse patients data



Comparison of tools used for alignment, variant calling and quality control by the four groups

- Differences in the preference for tools
  - Group D – GATK
  - Group A and C – BWA
  - Group B – Bowtie 2
- Evidence in the diversity of possible workflows

Conclusions

- Literature involving use of NGS shows that there is absence of over-all agreement on how data should be analysed and presented
- An enormous number of tools and data processing workflow systems being developed
  - A little detailed assessment of the application of these to establish best practice and specifically, recommendations and practices
- Variable results -> Verification by clinicians or wet labs
  - Since the application of results in clinical/hospital settings have consequences for the patients

Conclusions (cont.)

- 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
- What is the best analytical workflow???
- The challenge of future acceptance of workflows in clinical domain

Thank You