Dr Liviu Constantinescu

Information Architect
Garvan Institute of Medical Research

l.constantinescu@garvan.org.au

Liviu Constantinescu completed his PhD in computer science at the University of Sydney as part of the Biomedical and Multimedia Information Technology Research Group, specialising in software development and multimedia technologies. His research focuses on improving the practice of healthcare through state-of-the-art networking and software development methods.

Implementing a clinical genomics infrastructure to sequence 18,000 human genomes per year

Liviu Constantinescu[a], Mark Cowley[a,b], Kevin Ying[a], Peter Budd[a], Derrick Lin[a], Warren Kaplan[a,b], Marcel Dinger[a,b]

[a]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, 384 Victoria St, Darlinghurst, NSW 2010, Australia
[b]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Darlinghurst, NSW 2010, Australia

SUMMARY

Clinical genomics is a rapidly evolving field focused on the use of genome sequencing information to guide patient diagnosis and treatment. Whole genome sequencing has been dubbed "the test to replace all genetic tests", since one sequencing run can identify all genetic variants present in a patient's genome. Implementing clinical-grade, whole genome sequencing across large patient cohorts represents a substantial big data challenge. We will present our "Sabretooth" plan for scaling operations in our centre from an estimated 800 to 18,000 genomes per year.

INTRODUCTION

Sequencing of patient genomes is anticipated to have a large impact upon healthcare and the delivery of personalised medicine in three key areas: stratifying patients for appropriate cancer treatment; diagnosing inherited genetic disease; and tailoring prescriptions by anticipating adverse drug reactions.

Recently, the Kinghorn Centre for Clinical Genomics (KCCG) purchased the Illumina HiSeq X™ Ten sequencing system, which has the capacity to sequence 18,000 whole human genomes at an average of 30x coverage, per year. This will generate 150 genomes every 3 days, or 1.4 PB per year.

Although we anticipate that the storage issue can be addressed via currently available computing architectures, the new challenge lies in the delivery of this architecture in a manner that is both sufficiently versatile to keep pace with the rapidly changing bioinformatics landscape and rigorous enough to fulfil the stringent regulatory requirements for clinical data. This presentation will focus on the implementation of modern software development processes and infrastructure adopted by the thought leaders in IT[5], to meet NATA quality standards and allow us the flexibility to continuously improve our processes and analytics.

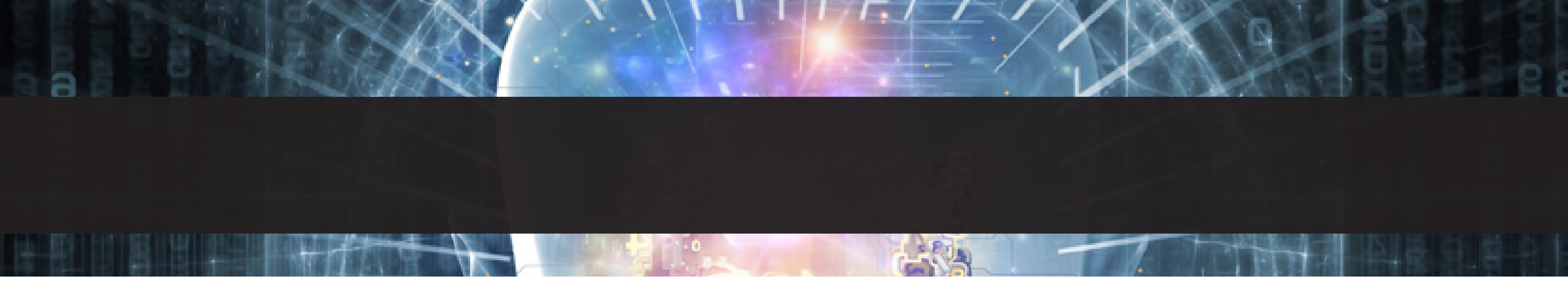
DESCRIPTION

Our bioinformatic workflow includes phenotype capture, read alignment, mutation calling, variant annotation and filtering by inheritance pattern, rarity, predicted functional impact and known disease association. Each stage utilises one or more software components, most of which are developed externally. These are supported by information systems that manage clinical data, laboratory processes and logistics. Every study traverses this "Sabretooth" pipeline, from accession to result.

Systems and modules in this pipeline undergo continuous, research-driven change, resulting in increased accuracy and diagnostic sensitivity. As the state of the art advances, obsoleted components must adapt or be replaced. This continuous change has a flow-on effect on subsequent components, and on the middleware interconnecting them. It poses four major challenges: managing software change; adapting and modularising workflows; generating auditable records; and allowing re-runs of legacy pipelines. The first two of these apply equally to clinical and research genomics, whereas the latter two are specific to a clinical context.

To manage software and requirements changes, the KCCG has put an agile software development process in place to continuously improve the modules, applications and information systems that make up our pipeline. By implementing daily stand-ups, feature backlogs, test driven development, automated testing suites, continuous integration and continuous deployment we gain confidence not only in the quality of the software we produce, but in our ability to manage the rapid release/deployment cycle of our systems, recover from hardware failures and roll out new features to the clinical and research arms of our group. Our implementation of the agile process strongly addresses the requirements and recommendations cited as critical to the development of high-quality bioinformatics software in the scientific literature [1,2,5,6].

For high-level management of repeatable, modular workflows, the KCCG have entered into collaboration with



the SeqWare working group at the Ontario Institute for Cancer Research (OICR). We've developed an in-house adaptation of their SeqWare framework, a set of infrastructure tools designed to guarantee the correctness of sequence analysis pipelines and deploy new versions on-the-fly. This framework supports a full hierarchy of functional, scientific and regression tests; retains history and metrics for every run; and incorporates a powerful query engine for interrogating our growing corpus of genome datasets[8].

Finally, a suite of agile process management and documentation tools centred around Atlassian's JIRA[3] augments our pipeline via automatic collection of business intelligence data regarding every stage of the process, guaranteeing end-to-end auditability and allowing clinical, analytical and management teams to tap into continuously updated information that traditional paper-based reporting cannot capture[4]. This information integrates release management, continuous integration and issue tracking, so the scope of every software and analytics change can be constantly monitored in terms of its impact on business and clinical outcomes.

CONCLUSION

KCCG is leading the charge toward the implementation of large-scale clinical genomics in Australia. We present Sabretooth as a case study in balancing the demands of clinical-grade informatics against the need to manage continuous change, so as to deliver the benefits of the most recent genomic research to all Australian patients in a cost-effective and reliable way.

REFERENCES

1. K. Rother, et al., A toolbox for developing bioinformatics software. Briefings in Bioinformatics 2011, 13(2), 244–257.
2. K. Beck, Test Driven Development: By Example. Addison-Wesley Professional, Boston, 2002.
3. Jira: Bug tracking, issue tracking, and project management. Available: http://www.atlassian.com/software/jira. Accessed 15/1/2013.
4. D. Larson, Agile Methodologies for Business Intelligence. Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications. IGI Global, 2012. 101-119. Web. 15 Jan. 2014
5. S. Baxter, S. Day, et. al., Scientific Software Development Is Not an Oxymoron. PLOS Computational Biology 2006, 2(9), e87.
6. D. Kane, M. Hohman, et al., Agile methods in biomedical software development: a multi-site experience report, BMC Bioinformatics 2006, 7:273
7. T. Nyrönen, J. Laitinen, et al. (2012). Delivering ICT infrastructure for biomedical research. Presented in the WICSA/ECSA '12: Proceedings of the WICSA/ECSA 2012 Companion Volume, ACM.
8. B. O'Connor, B. Merriman, et al., SeqWare Query Engine: storing and searching sequence data in the cloud, BMC Bioinformatics 2010, 11 Suppl 12, S2.