Dr Denis Bauer

Research Scientist
CSIRO

denis.bauer@csiro.au

Dr Bauer is interested in high-performance-compute-systems for integrating large data-volumes to inform strategic interventions for human health. She has a PhD in Bioinformatics and Post-Docs in machine-learning and genetics, published in Nature Genetics and Genome Research, was invited speaker at Bio-IT World Asia, and attracted more than AU$360,000 in funding.

Personalised cloud-computed genomics at health-system-relevant scale

Denis C. Bauer [a,b], Piotr Szul[c], Fabian A. Buske[d]

[a] Preventative Health Flagship, CSIRO, North Ryde, NSW, 2113, Australia
[b] Computational Informatics, CSIRO, North Ryde, NSW, Australia, 2113, Australia
[c] Computational Informatics, CSIRO, Marsfield, NSW, Australia, 2122, Australia
[d] Cancer Epigenetics Program, Cancer Research Division, Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, 2010, NSW, Australia

SUMMARY

Genomic information is increasingly incorporated into medical practice for diagnosis and personalised treatment. However, processing genomic information at a scale relevant for the health-system remains challenging due to computational requirements as well as high demands on data reproducibility and data provenance. Here, we present Next Generation Sequencing Analysis for Enterprises (NGSANE), a Linux-based, High Performance Computing (HPC) framework for production informatics, tailored to the demands and fast pace of personalised medicine, which is available as on-demand virtual cluster in Amazon's Elastic cloud.

INTRODUCTION

Unprecedented computational capabilities and high-throughput data collection methods promise a new era of personalised, evidence-based healthcare, utilising individual genetic or genomic testing to tailor health management as demonstrated by recent successes in rare genetic disorders[1,2] or stratified cancer treatments[3]. An analysis can take up to 4633 CPU hours per sample to process whole exome sequencing data and produce fully annotated genomic variants (see Figure 1A, CPU-single-threaded). The time, especially in the mapping stage, can be substantially reduced (7 fold) by utilising multithreading on High Performance Computing (HPC) clusters, where parallelisation between and within sample analysis can be easily implemented (see Figure 1A, CPU-multi-threaded).

To achieve minimal time delay between analysis tasks (i.e. mapping, recalibration, variant call, annotation) workflows are commonly automated by means of software 'pipelines'. While high demands are posed on data provenance and reproducibility of these pipelines, individual analysis components depreciate rapidly due to evolving technology and analysis methods, often rendering entire versions of production informatics pipelines obsolete.

Furthermore, the necessary parallelisation requires a large investment associated with compute hardware and IT personnel, which is a barrier to entry for small laboratories and difficult to maintain at peak times for larger institutes. This hampers the creation of time-reliable production informatics environments for clinical genomics. Commercial cloud computing frameworks, like Amazon Web Services (AWS) provide an economical alternative to in-house compute clusters as they allow outsourcing of computation to third-party providers, while retaining the software and compute flexibility.

To cater for this resource-hungry, fast pace yet sensitive environment of personalised medicine, we developed NGSANE, a Linux-based, HPC-enabled framework that minimises overhead for set up and processing of new projects yet maintains full flexibility of custom scripting and data provenance when processing raw sequencing data either on a local cluster or Amazon's Elastic Compute Cloud (EC2).

DESCRIPTION

Unlike currently available tools like Galaxy[4], BPipe[5], SeqWare[6] or Atlas2[7], NGSANE constructs pipelines based on Linux bash commands, which enables the use of hot swappable, modular components as opposed to the more rigid program-call wrapping by higher level languages or web-based services.

NGSANE separates project specific files from reference data, scripts, and software suites that are common to multiple projects. Access to confidential data is transparently handled via the underlying Linux permission system. A project specific configuration file defining the compute environment as well as the analysis tasks to perform facilitates the transaction between projects and framework. A full audit trail is generated recording performed tasks, utilised reference data, timestamps, software versions as well as HPC log files, including any errors.

Individual task blocks (e.g. read mapping) are packaged into bash script modules, which can be executed locally or on data subsets to test module code, submission parameters and compute environment in stages thereby mitigating the lack of debug-support from higher level languages/submission frameworks. During production, NGSANE automatically submits separate module calls for each individual data set to the HPC queue. This allows different existing modules, parameter settings, or software versions to be executed by changes to the project specific configuration file rather than the software code (hot swapping).

NGSANE gracefully recovers from unsuccessfully executed jobs be it due to failed commands, missing or incorrect input or under-resourced HPC jobs by enabling a clean restart from the most recent successfully executed checkpoint. Workflows can be fully automated by utilising NGSANE's control over HPC queuing systems and by leveraging the customisable interfaces between modules when submitting multiple dependent stages at once.

NGSANE supports the generation of a high-level summary (Project Card) to enable informed decisions about the experimental success. This interactive HTML report provides an access point for new lab members or collaborators, as well as a gold standard that can be used for testing purposes in a continuous integration server framework.

NGSANE is available as an Amazon Machine Image (AMI), which can be deployed to Amazon's EC2 by using, for example, MIT's StarCluster framework (http://star.mit.edu/cluster/) to launch a virtual cluster on demand (see Figure 1B). Other than regular on-demand instances, whose availability is guaranteed at a fixed price, StarCluster also offers command line-based sourcing of Spot Instances, where prices are based on current supply and demand. While Spot Instances can be acquired at a substantially lower price, their availability is not guaranteed. Hence NGSANE's checkpoint recovery is critical in such an unstable, competitive environment. Finally, NGSANE's HPC job partitioning and submission structure is independent from the program calls, therefore allowing new technologies (e.g. Hadoop) to be incorporated.

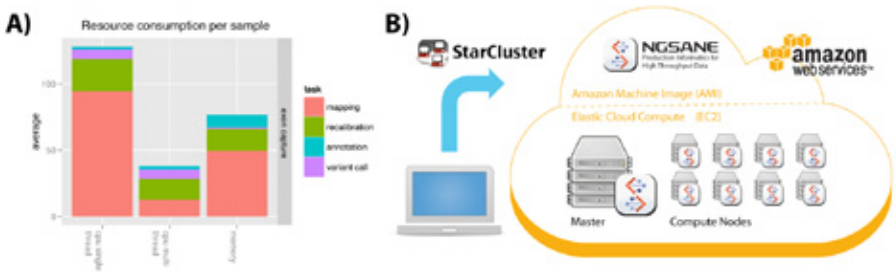

Figure 1. A) Resource consumption of the four steps involved in exon capture genomic data analysis. The average per sample is plotted in hours and gigabytes for CPU usage (single and multithreaded) and RAM memory usage, respectively. B) Schematic for a nine-node on-demand cluster with the NGSANE AMI deployed on every node on the EC2 service as launched by StarCluster.

CONCLUSION

NGSANE is a flexible HPC framework for NGS data analysis that is specifically tailored to the demands and issues of personalised genomics. NGSANE is implemented in bash and publicly available under BSD (3-Clause) licence via GitHub at https://github.com/BauerLab/ngsane. Currently implemented workflows include those for adapter trimming, read mapping, peak calling, motif discovery, transcript assembly, variant calling and chromatin conformation analysis.

NGSANE is available for local cluster installation or as an AMI to be deployed as an on-demand cluster on Amazon's EC2. This facilitates production-scale processing of large sample numbers and enables research at population scale to produce insights into individual disease risk and stratify treatment for common diseases with impact on the health system.

REFERENCES

1. Bainbridge, M.N., et al., Whole-genome sequencing for optimized patient management. Sci Transl Med, 2011. 3(87): p. 87re3-87re3.
2. Talkowski, M.E., et al., Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med, 2012. 367(23): p. 2226-32.
3. Pellatt, A.J., et al., Genetic and lifestyle influence on telomere length and subsequent risk of colon cancer in a case control study. Int J Mol Epidemiol Genet, 2012. 3(3): p. 184-194.
4. 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.
5. Sadedin, S.P., B. Pope, and A. Oshlack, Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics, 2012. 28(11): p. 1525-6.
6. O'Connor, B.D., B. Merriman, and S.F. Nelson, SeqWare Query Engine: storing and searching sequence data in the cloud. BMC Bioinformatics, 2010. 11 Suppl 12: p. S2.
7. Evani, U.S., et al., Atlas2 Cloud: a framework for personal genome analysis in the cloud. BMC Genomics, 2012. 13 Suppl 6: p. S19.