Digital Health: Changing the Way Healthcare is Conceptualised and Delivered
E. Cummings et al. (Eds.)

doi:10.3233/SHTI190776

SRA Down Under: Cache and Analysis Platform for Infectious Disease

Thom CUDDIHY[a], Brian FORDE[b], Nicholas RHODES[c], David PATERSON[d], Dominique GORSE[c,1], Scott BEATSON[b], Patrick HARRIS[d]
[a] *QFAB Bioinformatics - Research Computing Centre, University of Queensland, Australia*
[b] *School of Chemistry and Molecular Biosciences, University of Queensland, Australia*
[c] *QFAB Bioinformatics - Institute for Molecular Bioscience, University of Queensland, Australia*
[d] *UQ Centre for Clinical Research, University of Queensland, Australia*

Abstract. SRA, NCBI's Sequence Read Archive, is a valuable resource holding a near definitive collection of the world's collective sequenced reads for academic purposes. Increasingly, these reads are being used for both basic research and clinical investigations. When time is a critical factor in analysis, such as during bacterial outbreaks, the geographical separation between Australia and the offshore NCBI SRA servers can result in significant delays that may have adverse clinical outcomes. To address this, Queensland Genomics commissioned a pilot program for the establishment of a local Australian SRA Cache. Utilizing the hosting capabilities of the NeCTAR Research Cloud, QRIScloud's HTC infrastructure and the MeDiCI data fabric as a storage solution, and the software stack of Cromwell for workflow management, PostgreSQL database for sample and job metadata, and a coordinator Python Flask application, a local cache of seventeen bacterial species was established. Furthermore, the workflow capabilities of Cromwell were leveraged to provide analysis solutions for cached sample data, including quality control and taxonomic profiling, and individual and multiple sample analysis. Moving forward to a broader rollout of increased bacterial species, it was found that the initial storage estimation did not keep up with the exponential increase sequencing reads uploaded to NCBI SRA, which while highlighting the increasing availability and importance in modern research, will need to be addressed.

Keywords. SRA, Queensland Genomics, Cromwell, QRIScloud, MeDiCI

Introduction

The National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) represents one of the largest repositories for sequencing read data, with over 5.2PB of open-access data as of March, 2019[1].

There are many uses for such data, including study replication, exploratory studies and phylogenetic analysis. It can be particularly useful for both academics and clinicians during bacterial outbreaks or cases of multiple drug resistant infections, helping to identify potential sources of origin or clinical treatments.

[1] Presenting Author: QFAB Bioinformatics - Institute for Molecular Bioscience, Building 80, St Lucia Campus, Brisbane QLD 4072, University of Queensland, Australia; E-mail: d.gorse@qfab.org

Unfortunately, this repository is located off-shore and the access and download times for these large datasets are significant (i.e. several days for one single bacterial species). Furthermore, as most entries in the SRA do not have a corresponding assembly submitted, one must be generated, and this requires an additional amount of time, compute resources and expertise. For clinical situations regarding outbreaks or multi-drug resistant infections, this time delay could have tragic outcomes.

Queensland Genomics identified that a local source for SRA reads, as well as a curated set with pre-computed assemblies, single nucleotide polymorphism (SNP) profiling and other types of analysis, could aid in the identification and diagnosis of bacterial infections in real time and commissioned a pilot program for a national SRA Cache. The availability of the SRA Cache will enable the rapid comparison of a clinical sample to known samples at the DNA level and infer how this clinical sample relates to the known samples, both geographically and with clinical outcomes. This will allow health practitioners to determine whether a particular antibiotic resistance gene has spread to a new species, whether a particular strain of pathogen has been documented previously, or to detect both local or global outbreaks.

1. SRA Cache

1.1. Development, Deployment and Infrastructure

For the pilot program, it was determined that an Australian cache of SRA would be established from a subset of bacterial species. It would provide a local mirror of SRA samples and metadata for Queensland Genomics researchers to query and access, reducing the amount of time that they would need to retrieve the data from the NCBI servers.

Sixteen bacterial species were initially selected for inclusion in the pilot program, later increased to seventeen (refer Table 1), with the potential to add more species as the scope increases.

Table 1. Bacterial species selected for inclusion in the SRA Cache pilot and total counts as of March 2019.

Species	Count
Escherichia coli	112,778
Klebsiella pneumoniae	19,408
Enterobacter cloacae	1,875
Enterobacter hormaechei	803
Acinetobacter baumannii	7,906
Streptococcus pneumoniae	57,071
Streptococcus pyogenes	29,106
Streptococcus agalactiae	11,387
Streptococcus suis	2,663
Pseudomonas aeruginosa	10,741
Pseudomonas fluorescens	2,366
Pseudomonas putida	584
Pseudomonas syringae	582
Neisseria gonorrhoeae	10,820
Mycobacterium tuberculosis	67,933
Mycobacterium tuberculosis complex bacterium	4,733
Staphylococcus aureus	62,994
TOTAL	403,750

To achieve this, it was determined that two key pieces of software would be required: Cromwell[2], a workflow engine used to manage the compute steps, and a Python Flask application to coordinate the ingestion workflows and provide an internal Representational State Transfer (REST) Application Program Interface (API) for the Cromwell engine. Additionally, several internal endpoints were developed to store NCBI SRA metadata and log compute job steps in a local PostgreSQL database. Finally, several public endpoints were exposed for users to query the stored metadata, as well as to download sample data.

The National eResearch Collaboration Tools and Resources (NeCTAR) Research Cloud[3] was selected to host the primary virtual machine (VM), as it provided both scalability and integration with storage and compute facilities. It was then configured to submit jobs to the QRIScloud High Throughput Computer (HTC) Awoonga[4], and a 100TB Metropolitan Data Caching Infrastructure (MeDiCI)[5] data collection was provisioned and configured to be shared between the HTC compute nodes and the NeCTAR VM instance (see Figure 1).

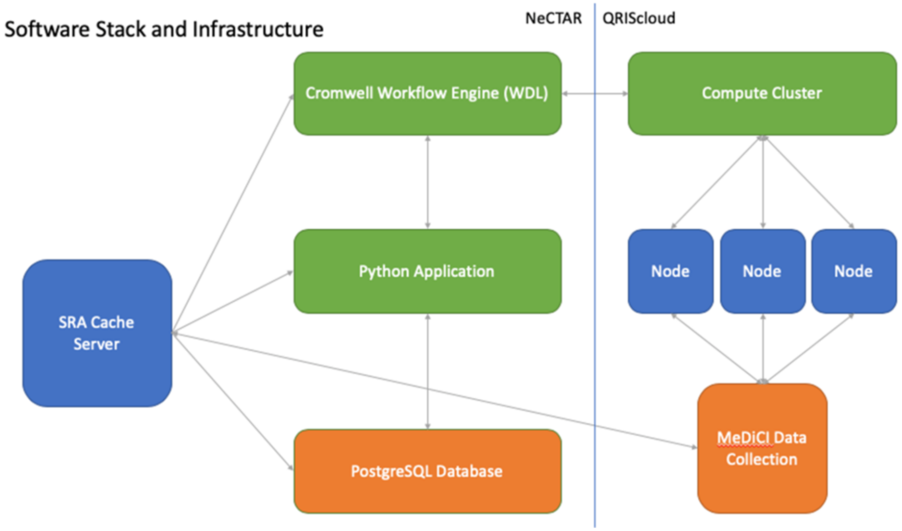

Figure 1. The software stack and infrastructure implemented for the Queensland Genomics SRA Cache. The primary server is running as a virtual machine instance on the NeCTAR cloud, containing the core Python application, database and Cromwell workflow engine. The workflow engine submits compute jobs to the QRIScloud HTC infrastructure, with both the compute nodes and primary server storing data on a shared MeDiCI data collection.

1.2. Ingestion Workflow

With the core software stack in place and the infrastructure setup, a Workflow Definition Language (WDL) script was written for Cromwell to ingest the SRA samples of the species of interest. Every week, a master job is triggered in Cromwell by the Python application. This job queries the Queensland Genomics SRA Cache via API to get a list of each bacterial species it should ingest. From this, it triggers a job for each species that queries all metadata of the NCBI SRA. Then, for each record that is returned, another job is spawned, that ingests that record metadata into the SRA Cache database. If it updates an existing record, or a new record is inserted, it then triggers another job to download and cache that sample using the NCBI SRA Toolkit[6]. A custom shell script was written to cancel any previously queued jobs for triggered samples still queued while the cache backlog was being established. All of these steps are recorded via the REST API job logging. A diagrammatic representation of this workflow is provided in Figure 2.

It was found that the initial backlog of SRA samples took 14 weeks to fully process, after which it was capable of maintaining new data on a weekly basis.

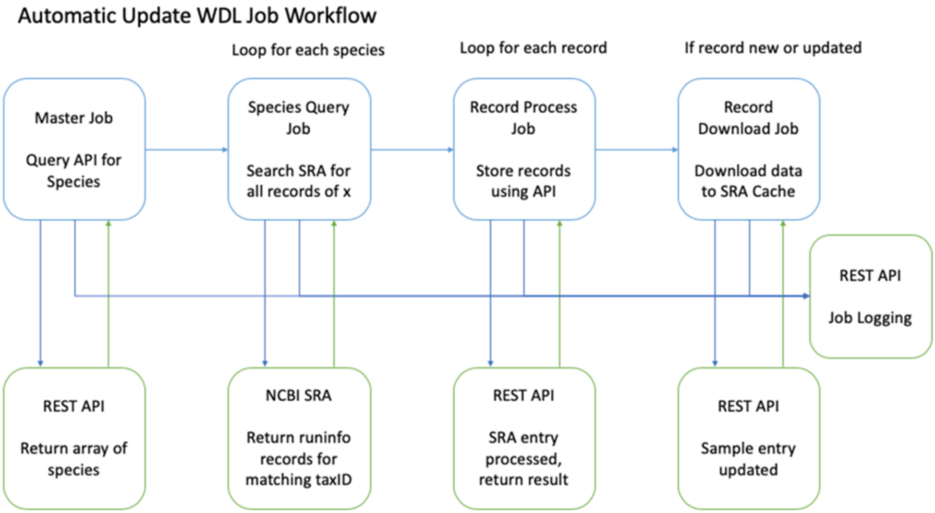

Figure 2. A diagram representation of the primary WDL discovery workflow and ingest sub-workflows.

1.3. REST API Access

As well as the internal API endpoints used to coordinate and log Cromwell jobs, and store SRA metadata, several public endpoints were developed for users to access the Queensland Genomics SRA Cache. A list of the endpoints currently publicly available as of March 2019 are listed in Table 2. All query endpoints are returned in JSON format, and support pagination, sort and order via meta-parameters. Generalized queries are supported via the ‘q’ parameter.

To maximise the utilisation of storage, the cache uses SRA reference-based compression where reads are aligned to known reference sequences and only their mapped location are recorded. While storage effective, reference-based compression does require decompression back to a sequence file (FASTQ) for most users. A single endpoint was developed that triggers both this staging process, and serves out the download URLs once completed. This was practical enough for the singular cases of API download requested, however this could be affected when large number of requests are submitted at once.

In addition to API access, several stakeholders with MeDiCI capability available to them were able to access the data directly via NFS within the QRIScloud NeCTAR availability zone, drastically improving their access times.

Table 2. API endpoints for Queensland Genomics SRA Cache. Base URL is https://sra-cache.qfab.org/api/v1

Endpoint	Method	Description
/species	GET	Get all species setup in cache
/list	GET	List all ingested SRA samples
/list/<ID>	GET	List all ingested SRA samples for species <ID>
/query?<parameters>	GET	Query SRA cache database with <parameters>
/fields	GET	Get all valid fieldnames valid for /query <parameters>
/get	GET	List all staged samples available for download
/get/<ACCENSION>	GET	Get download URLs for sample, or trigger staging
/get?id=<ACCENSION>	GET	Legacy method to trigger staging

2. Analysis Platform

2.1. Development

With the availability of sample data in Queensland Genomics SRA Cache next to compute resources, a WDL workflow was then developed to add value and utility to the data. In order to improve the quality and reliability of downstream analysis, a Quality Control (QC) workflow was first written to remove low quality sequence reads using Trimmomatic[7], and identify mis-assigned or contaminated samples using Kraken taxonomic profiling[8]. A more extensive analysis workflow was then developed that performed Multi-Locus Sequence Typing (MLST) and resistance profiling with srst2[9], de novo assembly with SPAdes[10] and SNP profiling (using curated references based on Kraken taxonomy) with Nesoni[11]. Finally, post multi-sample assembly analysis were performed with Abricate[12] and SNP distance network graphs established using Nesoni nways and custom Python code. Additional workflows are currently being developed with several collaborators to leverage the Queensland Genomics SRA Cache with their research goals.

3. Future Direction

3.1. Infrastructure

During the course of the pilot program, infrastructure was scoped to provide the most efficient means of covering the species selected. The multi-site deployment nature of MeDiCI caches was essential to the developed architecture, allowing access from both the NeCTAR cloud and QRIScloud HTC infrastructure. 100TB was initially scoped as being sufficient to house the selected bacterial species for the pilot program. As of March 2019, the total utilized storage of the Queensland Genomics SRA Cache was 112TB. The exponential increase in the NCBI SRA database growth[13] will need to be a factor in future storage solutions. In addition, with the excess of data, some records are now pushed to a tape back-end of the data framework, with the drawback of increased delay in access times for those records. For a full rollout of a national SRA Cache, the storage solution utilized would need to address this with an increase of the MeDiCI cache or the use of Zero Watt disk Storage for faster retrieval. Another improvement for a national rollout would be to store both the SRA reference-compressed data and pre-generated FASTQ sequence files, to eliminate the time needed to stage and serve.

Overall, the usage of NeCTAR for VM instance hosting and QRIScloud for compute were found to be suitable for both the pilot program and a potential increase in scope.

3.2. Web Platform and Custom Analysis

As the current deployment of the Queensland Genomics SRA Cache lacks a Graphical User Interface (GUI), extending the Python Flask application to serve a web GUI that covers the current query and download functionality would aid in usability. Additionally, extension of the GUI to display the results of individual and multi-sample analysis in a templated report format is planned. The web GUI could also be extended to include real-time read sequence querying. Recent advances in ultrafast SRA file searching have been implemented in the search tool BItsliced Genomic Signature Index (BIGSI)[14], which is able to scale to millions of SRA samples while maintaining millisecond kmer queries. By providing a web interface for BIGSI querying of the SRA Cache samples, researchers would be able to identify and explore new potential targets for investigation.

Finally, there is the potential to explore new and additional types of analysis. With the capacity of having SRA reads of entire species cached and available, emergent analysis may arise when additional groups and labs start using the service and asking different research questions. Extending the web GUI to allow the submission of custom WDL workflows, as well as an analysis-on-demand functionality, would enable researchers to take advantage of the Queensland Genomics SRA Cache in new and exciting ways.

Acknowledgements

This work was funded by Queensland Genomics, Queensland Health, Queensland Government. This research was also supported by use of the NeCTAR Research Cloud and by QRIScloud and University of Queensland's Research Computing Centre. The NeCTAR Research Cloud is a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

References

[1] NCBI, SRA database CSV, 2019. Retrieved from https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi

[2] K. Voss, G. Van der Auwera, and J. Gentry, Full-stack genomics pipelining with GATK4+ WDL+ Cromwell, *F1000Research* **6** (2017).

[3] National eResearch Collaboration Tools and Resources, Nectar Cloud - Nectar, 2019. Retrieved from https://nectar.org.au/research-cloud/

[4] UQ RCC, Awoonga - Research Computing Centre - The University of Queensland, 2019. Retrieved from https://rcc.uq.edu.au/awoonga

[5] D. Abramson, Caches All the Way Down: Infrastructure for Data Intensive Science, in: *Proceedings of the 26th International Symposium on High-Performance Parallel and Distributed Computing*, ACM, New York, NY, USA, 2017, pp. 27-27.

[6] R. Leinonen, H. Sugawara, M. Shumway, and C. International Nucleotide Sequence Database, The sequence read archive, *Nucleic Acids Res* **39** (2011), D19-21.

[7] A.M. Bolger, M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* **30** (2014), 2114-2120.

[8] D.E. Wood and S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome Biol* **15** (2014), R46.

[9] M. Inouye, H. Dashnow, L.-A. Raven, M.B. Schultz, B.J. Pope, T. Tomita, J. Zobel, and K.E. Holt, SRST2: Rapid genomic surveillance for public health and hospital microbiology labs, *Genome Medicine* **6** (2014), 90.

[10] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J Comput Biol* **19** (2012), 455-477.

[11] Victorian Bioinformatics Consortium, Nesoni cookbook, 2018. Retrieved from http://www.vicbioinformatics.com/nesoni-cookbook/

[12] T. Seemann, tseemann/abricate: Mass screening of contigs for antimicrobial and virulence genes, 2019. Retrieved from https://github.com/tseemann/abricate

[13] NCBI, SRA database growth, 2019. Retrieved from https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

[14] P. Bradley, H.C. den Bakker, E.P.C. Rocha, G. McVean, and Z. Iqbal, Ultrafast search of all deposited bacterial and viral genomic data, Nature Biotechnology 37 (2019), 152-159.