CGP as a Service:
From data submission to results using your web-browser

Keiran Raine, Adam P Butler, Peter Clapham, David Jones, Andrew Menzies, Lucy Stebbings, Jon Teague, Peter Campbell

ABSTRACT

The Cancer Genome Project (CGP) has been heavily involved in the work of the ICGC PanCancer Analysis Of Whole Genomes (PCAWG) project to characterise 2,800 cancers. CGP provided one of the three core somatic calling pipelines. As part of this effort we have successfully produced a codebase that is entirely portable, open-source, and available as a PCAWG workflow on www.dockstore.org. This workflow generates Copy-Number, Substitution, Insertion, Deletion and Structural Variant results optimised for tumour-normal paired somatic NGS data. CGP are now looking to provide an updated version of this workflow within a cloud enabled framework.

One of the key issues that faces investigators when working with large sequence data is the difficulty in transferring large datasets without the need to install dedicated software. In order to address this issue we plan to implement an in-browser, drag and drop process for data submission and retrieval. Following successful validation of data, mapping and analysis through the standard Whole Genome Sequence (WGS) somatic calling pipeline will be triggered. Here we present the current state of this work along with the current road-map for the next 2 years development.

COLLABORATIONS AND DATA

One of the main driving forces for CGP in continuing to develop our pipelines for public use in this way is for collaborative purposes. We regularly receive data from external groups to be processed through our internal pipelines. This is a large drain on our developer resources and can be problematic for a number of reasons including:

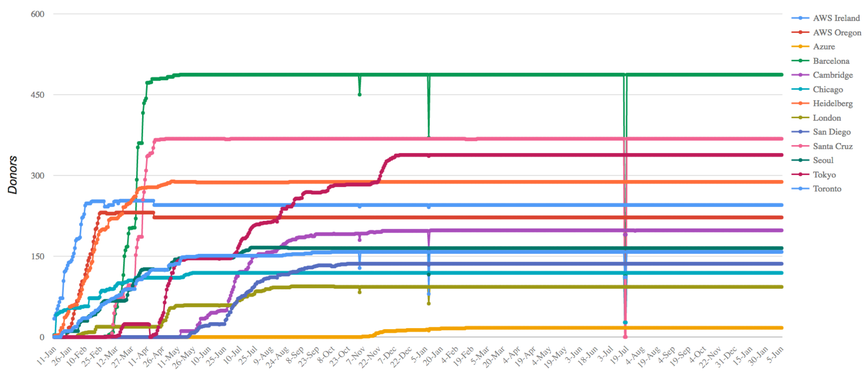
- Lack of informatics knowledge/support for small groups
  - Use of unix command line tools (sftp)
  - Data sent on hard-disk
  - Firewall changes
- Lack of understanding of data
  - Back and forth to correct errors
- Requirement for accurate metadata (run/lane/plex)
  - Specific to inclusion in our core pipelines

By addressing two of these areas we hope to reduce the overhead for this type of work as well as build an initial base platform that could be expanded into a public analysis platform.

FROM PCAWG TO DOCKSTORE

During 2014 the CGP developed the "Sanger" pipeline as part of the ICGC[1] PCAWG project. This involved converting all of our tools to work outside of any institute specific frameworks and making these available to the public (https://github.com/cancerit)

One of the technical successes of PCAWG was the large number of sites where the Sanger pipeline was successfully run. Much of this success was down to the tools developed by OICR (http://oicr.on.ca) to manage workflows on a heterogeneous collection of hardware and virtualisation frameworks.



During this process Docker was adopted to simplify distribution of tools. OICR have built on this to produce Dockstore[2] (https://dockstore.org/). All of the pipelines have workflows registered on Dockstore for future reference, however, these have been frozen with new developments, enhancements, corrections and performance improvements not being included.

CGP have continued to develop our tools and algorithms, fixing edge cases and improving efficiency. As part of this we have created three Dockstore tools we plan to support into the future:

Dockstore tool	Purpose
dockstore-cgpmap	Provides a BWA mem mapping flow accepting fastq, BAM or CRAM as input, producing BAM/CRAM.
dockstore-cgpwxs	Exome analysis tools generating SNV and Indel calls.
dockstore-cgpwgs	Adds SV, CNV, genotyping and gender checks to dockstore-cgpwxs.

SNV analysis is provided by CaVEMan[3], Indel by cgpPindel[4], CNV by ascatNgs[5] and SV by BRASS[6]. The underlying codebases for all of these can be found in the cancerit github organisation[6].

Both Dockstore and these tools are already being used in another largescale project, the ICGC Pan Prostate Cancer Group (PPCG) project.

One of the key differences between PCAWG and PPCG is that there is no single group managing the running of tools with individual centres with compute being required to setup and execute the analysis. The use of docker and Dockstore are critical to this process.

SOLVING DATA SUBMISSION

ceph docker Dockstore

Moving data is always suboptimal, but for smaller research groups with little local compute it is inevitable. Our primary focus is to simplify this process down to "Drag 'n Drop" within a web-browser.

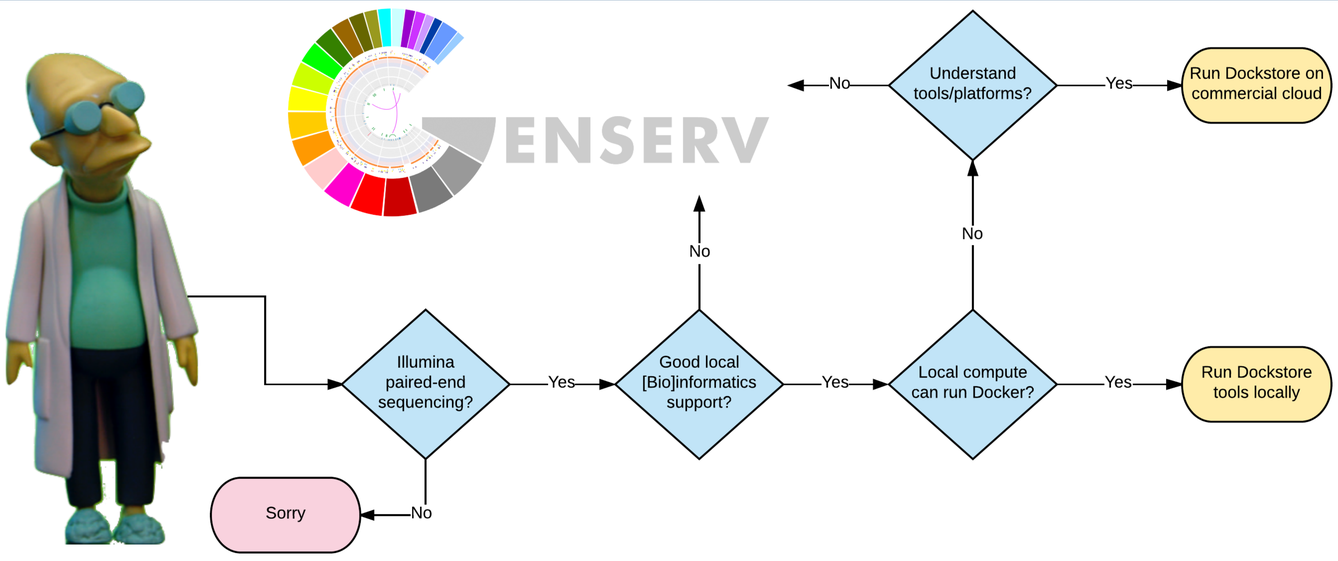
Additionally we've simplified the upfront sample information down to file groupings (data for the same sample) and tumour/control designations. For user ease the manifest will be standard Excel format with embedded validation as well as instant feedback on submission. The upload system will be completely web driven and accept fastq[.gz], BAM or CRAM as input. Data previously mapped and combined in BAM/CRAM is also acceptable.

The data will be uploaded to the Sanger Flexible Compute[7] Ceph storage using S3 transfer protocols. Once received, data is passed through a QC process running in an auto scaling environment managed by wr[8] (workflow runner) within OpenStack.

QC failures will automatically be fed back with recommendations, however, we attempt to automatically handle many common issues.

The first roll out will focus solely on passing data to our internal systems for processing.

LOOKING AHEAD



Once the data import element of the project has been confirmed as an improvement on current methods by users, we plan to expand the system to handle the actual mapping and analysis over several phases.

Doing this will allow a slow scale up of features over time, but the current starting point is BWA mapping of Human GRCh37 for WXS and WGS data.

Our ultimate goal is to deliver a service with a very low barrier to entry. Ideally, providing compute along with bioinformatics support and responsive inclusion of new tools.

Phase	Phase I	Phase II	Beyond
Availability	Close collaborators	Close collaborators	Open (new features to collaborators first)

Features	• Manifest validation
	• Sequencing data upload
	• Raw data QC
	• Data progress views
	• Import to legacy CGP pipelines
		• Implement Dockstore tools
		• Human GRCh37
		  • Mapping
		  • WXS (SNV, Indel)
		  • WGS (WXS + CNV, SV, Genotyping)
		• Data sharing (owner shares to other user)
		• Download of results.
			• Post mapping automated QC
			• Human GRCh38 + other species
			• Emerging sequencing types
			• Interactive workflow generation (approved tools only)
			• Open requests for tool support.
			• Mapped data passed to ENA/EGA on users behalf

REFERENCES

1. http://icgc.org/
2. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000res. 2017*
3. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics. 2016*
4. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics. 2015*
5. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics. 2016*
6. https://github.com/cancerit/BRASS
7. https://hpc-news.sanger.ac.uk/
8. https://github.com/VertebrateResequencing/wr

wellcome trust sanger institute

CancerIT (https://github.com/cancerit) is part of:

Cancer, Ageing and Somatic Mutation (http://www.sanger.ac.uk/science/programmes/cancer-genetics-and-genomics)

This work was supported by the Wellcome Trust grant [098051].