SevenBridges

ACCELERATING GENOMIC ANALYSIS ON THE CLOUD

Enabling the PanCancer Analysis of Whole Genomes (PCAWG) consortia to analyze thousands of genomes

Enabling the PanCancer Analysis of Whole Genomes (PCAWG) consortia to analyze thousands of genomes

In October 2015 Seven Bridges joined the technical working group of the PanCancer Analysis of Whole Genomes (PCAWG) project to help the project meet its analysis deadlines. We rapidly and reproducibly analyzed over 1,350 whole genomes using the cloud-based Seven Bridges Platform. The output files are included in the final PCAWG project data set where they will contribute to the next generation of cancer research.

INTRODUCTION

Seven Bridges is the biomedical data analysis company accelerating breakthroughs in genomics research for cancer, drug development and precision medicine. We build the scalable, cloud-based Seven Bridges Platform, which empowers rapid, collaborative analysis of thousands of genomes in concert with other forms of biomedical data. We also build the Cancer Genomics Cloud, one of the US National Cancer Institute's cloud pilot systems. Seven Bridges is used by thousands of researchers in government, biotech, pharmaceutical and academic laboratories.

In October 2015 Seven Bridges joined the technical working group of the PanCancer Analysis of Whole Genomes (PCAWG) project to contribute to their data production program. This project is run by a large academic consortium who needed to move their analysis to the cloud in order to meet program deadlines.

Seven Bridges joined PCAWG as the first commercial participant in the technical working group to provide official files for downstream analysis, with the task of running a variant calling workflow rapidly and at large scale. Here we describe our involvement in the PCAWG project, and the rapid delivery of variant calls for over 1,350 whole genomes.

PANCANCER ANALYSIS OF WHOLE GENOMES

The PCAWG project is an international collaboration investigating patterns of mutation in over 2,800 cancer patients. The PCAWG dataset includes donors from 47 different cancer research projects with 20 different primary tumor sites. Tumor–normal samples from each donor have undergone whole-genome sequencing and all tumor and matched normal genomes are analyzed using a uniform set of alignment and variant calling algorithms. These genotypes will be integrated with additional clinical and molecular data (such as RNA-Sequencing) from the same cases.

PCAWG is largest and most complete investigation of the molecular mechanisms that contribute to cancer, and the first to combine thousands of whole genome sequencing analyses. Made possible by the collaboration of hundreds of researchers worldwide, it is designed to explore the nature and consequences of somatic and germline variations in both coding and noncoding regions, with specific emphasis on cis-regulatory sites, noncoding RNAs, and large-scale structural alterations (Box 1). However, the computational requirements for analysis and storage are orders of magnitude greater than prior analyses. Such analyses can be done on large academic clusters, but are ideally suited to the cloud.

PCAWG KEY AREAS OF STUDY

1. **Discovery of driver mutations outside of the protein-coding regions of the genome;**
2. **Integrating mutational signatures across tumor types and mutation categories;**
3. **Characterizing subclonal structures and patterns of genome evolution across cancers;**
4. **Investigating relationships between germline and somatic mutations;**
5. **Investigating biological pathways targeted by driver mutations.**

THE SEVEN BRIDGES PLATFORM

To support the PCAWG technical working group, we used the Seven Bridges Platform to quickly perform variant calling on more than 1000 whole genome sequences. The Seven Bridges Platform is a cloud-based environment for conducting bioinformatic analyses. It allocates storage and compute resources on demand to meet the needs of a given analysis. While the platform is infrastructure-agnostic, for the PCAWG study we took advantage of the ICGC dataset hosted on AWS which allowed us to quickly, securely, and cost effectively access ICGC data.

The Seven Bridges Platform is home to hundreds of popular community-developed bioinformatics tools and workflows that researchers can readily customize and refine to meet their specific application. Additionally, the platform hosts a robust Software Development Kit (SDK), which allows users to add their own applications. The PCAWG analysis used a custom variant calling workflow, which was brought to the Platform using the SDK.

Additionally, the Platform has an Application Programming Interface (API) that lets users build automated workflows to perform quality control checks, run analysis tools, and simplify the upload and metadata capture process. The PCAWG analysis was done using our newly released API version 2 (http://docs.sevenbridges.com/docs/the-api).

The Seven Bridges Platform is HIPAA-compliant, and Seven Bridges is the only commercial Trusted Partner of the NIH, which means we can store, manage, and grant access to controlled data from The Cancer Genome Atlas dataset.

BUILDING THE WORKFLOW

We received aligned BAM files from the PCAWG Consortium, with the task of running standard BAM cleaning and applying a custom variant calling workflow provided by the Broad Institute. The resulting output files were to be labeled according to a specified scheme and uploaded to the University of Chicago for inclusion in the final PCAWG data set. Speed of completion was a priority, and the donor files were allocated among computation providers according to their capacity to rapidly process them.

The Seven Bridges Platform describes data analysis tools and workflows using the Common Workflow Language (CWL)—a set of open-source specifications that we have helped to develop, which also record version numbers and parameter settings. CWL tools and workflows are fully portable, and not locked into any execution environment. For a project such as PCAWG in which output standardization matters this level of reproducibility is essential.

While some of the tools in the Broad variant calling workflow are publically available, some tools are only available under Materials Transfer Agreement and are provided in a docker container. Our bioinformatics team was able to build docker containers or use existing ones to make all of the workflow tools easily deployable. Our team then described the tool inputs, outputs and parameters using the CWL specification and then linked each of the tools together to create a reproducible, portable and scalable workflow.

PCAWG DATA AVAILABILITY

Data Type	# Donors	# Files	Format	Size
Simple Germline Variations	2,819	8,132	VCF	490.88 GB
Structural Germline Variations	2,819	5,203	VCF	5.34 GB
Aligned Reads	2,834	8,815	BAM	726.69 TB
Simple Somatic Mutations	2,834	18,604	VCF	181.89 GB
Copy Number Somatic Mutations	2,834	5,887	VCF	127.91MB
Structural Somatic Mutations	2,834	12,715	VCF	1.16 GB
Total	**2,834**	**59,356**	**NA**	**727.35 TB**

Data from ICGC data portal https://dcc.icgc.org/pcawg [accessed 1 April 2016]. Seven Bridges analyzed over 1,350 genomes from 655 donors, and provided over 5,000 VCF files to the final PCAWG data set.

ANALYZING CANCER WHOLE GENOMES

To validate the workflow and our CWL description, we tested it on tumors and normal tissue from four test donors, totaling around 20 samples. The testing process also served as a pilot program ahead of the larger file-processing effort. Analyzing each sample was found to require 2–6 days to complete, depending on a number of factors including file size, which varied across samples. AWS instances have fixed-size disks attached, so the instance has to be matched to the required disk space. The workflows curated on our Platform are optimized to run on particular instances depending on computational and storage needs.

We allocated the PCAWG samples across AWS machine types according to file size. Because instances are priced differently depending on the amount of disk space and memory, being able to select instances according to the computational and storage needs results in a cost saving for the user (in this case we estimate >20% saving). Such optimized resource allocation is an important feature of our Platform.

We initiated the analysis as soon as we obtained access to the relevant files and permissions. A small team of bioinformaticians ran and monitored batches of 300–600 tasks. We determined batches based on file size, which allowed them to run on different AWS instances depending on which was most cost efficient. The automatic selection of instance types is a major benefit of running analyses on the Seven Bridges Platform.

By working together with Seven Bridges, the PCAWG Consortia was able to meet its aggressive computation timeline requirements. Seven Bridges delivered more than 1,350 whole genome variant calling results to the Consortia for use in research studies around the world.

DELIVERING THE VERIFIED FILES

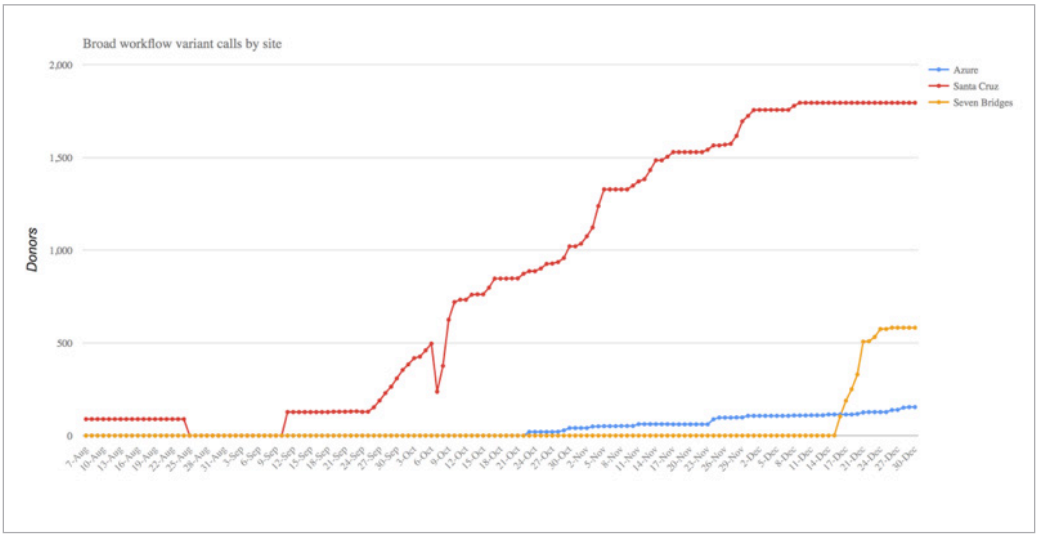
For each donor, we verified and uploaded over 40 separate files comprising the vcf outputs from the MuSE and the Broad workflows, packaged intermediate files, and associated metadata.

The PCAWG Consortium provided detailed instructions for the post-processing of output files, packaging into directory structure and rsyncing to the University of Chicago servers. To make file delivery as efficient as possible, our team of bioinformaticians worked with the consortium to define dedicated output scripts to fit the specific metadata schema and requirements of production grade processed data that will be used in downstream analyses by researchers worldwide.

When the analysis tasks completed, the results and output files were saved on Amazon S3. For each donor, we downloaded the files onto a new instance, and ran our scripts to verify the output and rename and create the correct directory structure. Three different output types required uploading, each in a prescribed format:

1. The vcfs and associated files from the Broad workflow;
2. Specified intermediary files from the Broad workflow, packaged into a tarfile;
3. The vcfs and associated files from the MuSE workflow.

FIGURE 2



COMPLETING THE ANALYSIS

The rich feature set of the Seven Bridges Platform (i.e. workflow versioning, automation, instance configuring, and task monitoring) enabled us to rapidly complete analysis of over 1,350 whole genomes, and to provide the verified output files to the PCAWG Consortium. The whole genomes were rapidly analyzed to strict specifications and, thanks to CWL, the analysis methods are transparent and reproducible.

Moreover, Seven Bridges was proud to use our expertize to participate in the full array of cancer genomics data generation and delivery. As a part of the PCAWG Consortia, we helped create the data which teams of researchers will use to better understand cancer at the molecular level.

FIGURE 3 | PCAWG FILE MANIFESTS

TEAM@SEVENBRIDGES.COM
SEVENBRIDGES.COM

SevenBridges