PCAWG - Pancancer Analysis of Whole Genomes

Request for Collaboration: Commercial Compute Resources for Pan-Cancer Analysis of Whole Genomes.
Date: 13 July 2015
Expressions of interest due: 24 July 2015

The Steering Committee of ICGC's Pan-Cancer Analysis of Whole Genomes (PCAWG; https://pcawg.icgc.org/pcawg) consortium is seeking industrial partners to help meet the compute demands of the project. We seek computing expertise, compute time and temporary storage from the commercial partners, who will, in return, receive publicity and access to over 700 bioinformaticians involved in the consortium with state-of-the-art expertise on cancer genomes.

The PCAWG project has generated a uniform set of whole genome alignments for over 2800 tumor-normal pairs and is currently engaged in cataloguing somatic and germline variations among the tumor genomes. In addition to genomic data, a substantial portion of the donors have additional supporting -omics data including sequenced RNA and arrayed DNA methylation.

To date, eight academic compute centers have contributed to the project, but the overall number of dedicated compute nodes is limited. An estimate of the total amount of computation needed puts the project months behind schedule. For example, running a single group's somatic variant caller took over two months. The group is planning to run at least two more variant callers, and likely up to half a dozen callers, to identify somatic variants. The results of several germline callers are needed as well. Various fundamental core analysis of the transcriptome that use the RNA-Seq data (e.g. alternative splice products, gene levels, and fusion genes) also require a large demand for compute.

The new cloud policies announced last month by the NIH make possible the involvement of commercial providers to significantly accelerate the work of the PCAWG project. For these reasons, the SC is seeking commercial providers to participate in at least, but not limited by, two aspects:

1. Knowledgeable engineers with expertise in creating portable code are needed. The project has employed the use of Docker to create algorithm containers, or lightweight

virtual machines, that are able to be shipped and run on a variety of compute platforms. Dozens of algorithms need to be wrapped into the Docker specification to ready them for cloud or cluster compute. The amount of time needed for this wrapping procedure varies with the sophistication of the algorithm pipeline, but experience in the project (e.g. wrapping the Broad's MuTect and associated pipeline tools) suggests the time can be appreciable such as weeks to months to complete. Thus, the SC is seeking engineers to work with PCAWG member groups to wrap tools into Docker.

2. The PCAWG steering committee is seeking commercial providers to donate compute time and temporary storage on Docker-ready cloud systems. Commercial providers would make cycles available to members of the PCAWG project free of charge. No restrictions would be placed on the output of the algorithms to maximize their utility for downstream analysis. The types of analysis could include help with "core pipelines" (such as variant calling and RNA-Seq quantification) or with downstream analyses of interest to member groups.

Interested commercial providers should expect to donate a minimum amount of compute and temporary storage to ensure the project is aided, and not hindered, by the participation. To be competitive the provider is requested to donate 2.3 million core hours over a maximum of 10 weeks, which is sufficient for processing 1/3rd of the donors in the PCAWG set. The providers will also donate the costs of data transfer (~200 TB incoming and ~20 TB outgoing).

Due to the current rules governing data accessibility and distribution, providers will be expected to delete all PCAWG related data once the PCAWG group has finished with its analysis. However, there will be near-future opportunities to engage in a process to authorize the redistribution of a significant portion of the raw and analyzed data.

Contributions made to the project will be acknowledged both verbally at conferences and in written form in all major manuscripts communicated by the PCAWG group. The expected high-profile nature of the PCAWG project (e.g. papers have been negotiated to appear in *Nature*) should garner visibility and credibility for contributing providers.

Over the next couple of weeks, the SC will collect a list of interested commercial entities from which a set will be selected and named on a revised data use agreement application with the NIH. Any commercial provider interested in helping should contact Jennifer Jennings (jennifer.jennings@oicr.on.ca) by Friday, July 24. A meeting with the SC will then be arranged and a memorandum of understanding drafted and circulated.

Sincerely,

The ICGC PCAWG Steering Committee

Peter Campbell, Lincoln Stein, Jan Korbel, Gad Getz, Josh Stuart