Title: Working with Big Cancer Data in the Collaboratory Cloud

Date: November 4
Location: Vancouver, BC
Half-day: 8:30 am – 12:30 pm
Class size: 50 people
Cost: $100+tax

Target Audience: Graduates, postgraduates, staff bioinformaticians, and PIs who need to learn how to access and work with cloud compute infrastructure in support of their research on big data sets (e.g. cancer data from the International Cancer Genome Consortium (ICGC) or PanCancer Analysis of Whole Genomes (PCAWG)). Familiarity with the Unix command line interfaces is recommended but not required.

Course Description: The Cancer Genome Collaboratory is a compute cloud that was set up to facilitate complex analyses on big cancer genome data projects, including the ICGC and PCAWG. The Collaboratory provides access to configurable virtual machines (VM) with which to compute on this data (thereby removing the need to purchase and maintain your own compute cluster). To navigate through working in this new compute space, the CBW has developed a half-day course providing a hands-on introduction to launching and configuring your own virtual machine (VM), accessing Cloud-based data sets, and work with your data. Cloud-computing best practices will also be discussed.

Participants will gain practical experience and skills to be able to:

- Launch their own virtual machine (VM)
- Configure a VM with prepackaged tools
- Pull in data sets from Cloud repositories
- Follow best practices in data and workflow management
- Run a data analysis pipeline in the Cloud

Module 1: Introduction to the Cancer Genome Collaboratory (George Mihaiescu)
Lecture (45 min):

- Introduction to Cloud Computing and Virtual Machines

- The Cancer Genome Collaboratory Cloud
- Introduction to Docker and Dockstore

Lab practical (1 hr):

- Setup and launch a VM within the Collaboratory environment
- How to execute Docker containers stored on Dockstore
- Setup and configuration for worked example

Coffee Break – 30 min

Module 2: Big Data Analysis in the Cloud (Sohrab Shah lab)
Lecture (45 min)

- Large-scale biological activities that generate datasets used in the Cloud
- Searching for ICGC data stored on Collaboratory
- How to access data in the Cloud and accessing non-protected data

Lab practical (1 hr): Solomon Shorser and Sohrab Shah lab

- Initiate the sequence alignment task on cell line data