Cologne Center for Genomics

Workshop

On April 10th 2018 the Lal Research Group at the Cologne Center for Genomics (CCG) will be hosting a coding workshop on

Hail — an open-source, scalable framework for exploring and analyzing very large genomic data.

With the introduction and evolution of next-generation sequencing platforms, it is now feasible to analyze exomes and genomes of thousands and soon millions of people. This remarkable development is accompanied by great computational challenges to analyze the huge data sets. Alongside the new challenges there are also new opportunities. A team from the **Neale lab** at the Broad Institute of Harvard and M.I.T., US has developed the **Hail Project**. The aim of the Hail project is to harness the flood of sequenced genomes in order to unravel the genetic architecture of disease. Their open-source framework is already being used to analyze the largest genetic data sets available, to power dozens of major academic studies, and to meet the exploding needs of hospitals, diagnostic labs, and industry.

What is the course about?

In this course, participants will gain an overview of the functionality of Hail and hands-on experience using Hail. Hail is a library for data analysis in Python and runs on Apache Spark. Because it can scale from one computer to thousands, Hail has had enormous impact in the analysis of the largest sequencing and genotyping datasets, but Hail is useful for analyzing data of any scale.

Who is the course instructor?

Tim Poterba is a software engineer in the Hail Team at the Broad Institute of Harvard and M.I.T. Before this role, he was a computational biologist in the Neale lab at the Broad Institute.

Course details: The course will be held April 10th and will take place at the CMMC (http://www.cmmc-uni-koeln.de/home/) in the large seminar room. It will start with a seminar giving an overview on “Hail — an open-source, scalable framework for exploring and analyzing very large genomic data” open to everyone. Right after the seminar the workshop will begin at the same place, however, this will be open to registered participants only.

More information about Hail: More background including tutorials can be found on the official Hail.is homepage.

Cologne Center for Genomics

hail

Agenda

Talk (12:00-12:45), open to everyone: Hail — an open-source, scalable framework for exploring and analyzing very large genomic data

Workshop (13:00-17:00), open only for registered participants:

Sections:
1. Background: applications of Hail such as GWAS
2. Types, expressions, and missingness
3. Table: representation of one-dimensional data
4. Filtering and annotation
5. Aggregation
6. Joins
7. MatrixTable: representation of genetic matrices
8. Generic example and exercises: film dataset analysis

Target audience: Graduate students, postdoctoral scholars, clinical scientists, and principal investigators currently working with genomic data, or about to embark on projects that require analysis of such data.

Prerequisites: Basic familiarity with UNIX/Linux environments is required. Participants should be comfortable with Python (more than basic familiarity). **Important:** For individuals with no training/preparation in Python nor strong coding skills in a related language this hands-on workshop ***will not be useful***. However, interested individuals can still join the seminar before the course.

Costs: Attendance at the course is free.

How to apply: Due to space constraints, we can only accept a limited number of participants. Please send your CV and a half page personal statement/essay about why you want to participate and to which degree you fulfill the course prerequisites to dlal@broadinstitute.org.