Contribution ID: **46** Type: **Tutorial**

NGS Data Analysis Training Workshop

Wednesday, 11 November 2015 13:30 (2 hours)

Summary

"Big data" is one of today's hottest concepts, but it can be misleading. The name itself suggests mountains of data, however big data consists of three V's: volume of data, velocity of data processing, and variability of data sources. These are the key features of information that require big-data tools to make use of high networking and computing power.
Researchers working with genomics in medicine, agriculture and other life sciences are producing big bio-data, mainly by the application of Next-Generation Sequencing (NGS) to give answer to important biological issues. NGS technologies are revolutionizing genome research, and in particular, their application to transcriptomics (RNA-seq) is increasingly being used for gene expression profiling as a replacement for microarrays. In order to deal with big bio-data, , current approaches in Life Science research favor the use of established workflows which have been proven to facilitate the first steps in data analysis.
We propose to setup a Training Workshop focusing towards the particular needs of the researchers active in the field of NGS data analysis, but so far have limited to no experience with the use of EGI resources for this task.

Description

The workshop will consist of two parts; initially there are going to be presentations from key applications and workflows that are currently established in the EGI ecosystem that cater to the particular needs of NGS analysis. This first part will give an in depth idea of the state-of-the-art in this field to the participants, and prepare them for the second part of the session, i.e the hands-on exercises. The exercises will be carefully selected both in terms of generality (i.e. applicable to a wide range of NGS data and analyses), as well as time constraints (i.e. small enough to conclude within the context of the session). The scope of these exercises will address issues such as input and reference data management, use of established analysis tools both in a Grid as well as a Cloud infrastructure, and tools for retrieving and further analyzing the produced output.

All exercises will be performed on the Chipster platform (http://chipster.csc.fi/) using EGI FedCloud resources. The process of connecting to the FedCloud and launching Chipster will be very briefly addressed in the context of this tutorial. However, there is a dedicated tutorial on "Running Chipster data analysis platform in EGI Federated Cloud" before this session (https://indico.egi.eu/indico/contributionModification.py?contribId=25&sessionId=26&confId=2544) so interested participants are encouraged to attend both tutorials.

Impact

The goal of this workshop is to attract life science researchers from different fields (such as agro-biotechnology, health and nutrition among others) that are (a) actively using NGS data analysis workflows in their research, and (b) have little experience employing EGI resources. Although researchers active in NGS are currently limited in number, as compared with the number of life scientists, the rise of NGS data across all Life Science domains leads to an increasing demand of both trained personnel as well as novel tools and approaches.

Timetable

Part 1: NGS Data Analysis in EGI (60')
10' Intro to EGI resources
15' Data Replication
15' Cloud Applications
15' Grid Applications
5' Discussion / Wrap-up

Part 2: Hand-on training (60')
- Exercise #1: Connect to a Cloud VM
- Exercise #2: Select ref data and replicate
- Exercise #3: Upload test input NGS data

- Exercise #4: Execute workflow
- Exercise #5: Post-workflow analysis

Summary

The hands-on exercises of the NGS Analysis Training Workshop will be based on the BioLinux environment, with the necessary tools and databases pre-installed. Indicative tools that will be employed are TopHat, Cufflinks and Bowtie2. In terms of reference databases, sample workflows will be executed on model genomes such as Homo Sapiens and Arabidopsis Thaliana.

Primary authors: Dr ARGIRIOU, Anagnostis (Institute of Applied Biosciences / CERTH); Dr HADZIDIMITRIOU, Anastasia (Institute of Applied Biosciences / CERTH); Dr PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki); MATTILA, Kimmo (CSC); Prof. STAMATOPOULOS, Kostas (Institute of Applied Biosciences / CERTH)

Co-authors: Dr VARDI, Anna (Institute of Applied Biosciences / CERTH); KOUMANTAROS, Kostas (GRNET)

Presenters: Dr HADZIDIMITRIOU, Anastasia (Institute of Applied Biosciences / CERTH); Dr VARDI, Anna (Institute of Applied Biosciences / CERTH); Dr PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki); KOUMANTAROS, Kostas (GRNET)

Session Classification: Tutorial: NGS data analysis