Contribution ID: **23** 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, but that's just the start. Overall, data can be 'big'for three reasons –often referred to as the three V's: volume of data, velocity of processing the data, and variability of data sources. If any of these key features are present, then big-data tools are necessary, often combined with high network bandwidth and massive compute systems.
Researchers working with genomics in life sciences (biomedicine, agri-food science etc) 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. In order to deal with big bio-data, current approaches in Life Science research favor the use of established workflows facilitating the first steps in data analysis.

This Training Workshop will focus on the particular needs of the researchers active in the field of NGS data analysis, but so far have limited or no experience with the use of big data tools and large compute systems.

Description

The workshop will use compute resources from EGI, a publicly funded infrastructure that offers compute and storage resources and services for researchers in academia and industry. The workshop will consist of two parts; initially there are going to be presentations from key applications and workflows currently established in the EGI ecosystem that cater to the particular needs of NGS analysis. This first part will give the participants an in-depth idea of the state-of-the-art in this field, and prepare them for the second part, 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 such as quality control, filtering and trimming of reads, assembly, annotation and differential expression), 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 in 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 cloud resources from the EGI Federated Cloud infrastructure. The process of accessing cloud resources and launching Chipster will be briefly addressed in the context of this tutorial. However, there is a dedicated tutorial on "Running Chipster data analysis platform in EGI Federated Cloud" (https://indico.egi.eu/indico/contributionDisplay.py?sessionId=26&contribId=25&confId=2 before this session so interested participants are encouraged to attend both tutorials.

Impact

Researchers active in NGS are currently few in number, as compared with the number of life scientists. However, the rise of NGS data across all Life Science domains leads to an increasing demand of both trained personnel and novel tools and approaches. With this in mind, the goal of this workshop is to attract life science researchers from different fields in life sciences who are (a) actively using NGS data analysis workflows in their research, and (b) have limited experience in employing large scale computer systems or specifically EGI resources.

Timetable

Part 1: NGS Data Analysis in EGI (40')
10' Quick Introduction to EGI resources, NGS Analysis workflows
10' Data Replication
10' Cloud Applications
10' Discussion / Wrap-up

Part 2: Hand-on training (80')
• 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

Additional Information - Requirements

The exercises of the NGS Analysis Training Workshop will be solved using the Chipster platform. The computational resources required will be provided by the EGI Federated Cloud Infrastructure.

The participants will be required to have a laptop (any OS with Java installed) which will be used to access the EGI training resources through SSH (for the launch of the VM) and the Java interface (for the exercises).

Links, references, publications, etc.

[1] Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A., "Differential expression in RNA-seq: a matter of depth", Genome Res. 2011 Dec;21(12):2213-23. doi: 10.1101/gr.124321.111. Epub 2011 Sep 8.

[2] Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. and Thurston, M. 2006. Open Software for biologists: from famine to feast. Nature Biotechnology 24, 801 - 803.

Additional information

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