IHEC Data Portal 2016 update: datasets quality control, permanent sessions, public API

David Bujold1, Catherine Côté1, Jonathan Laperle2, Carol Gauthier2, Michel Barrette2, David Morais2, Tony Kwan1, Alain Veilleux2, Pierre-Etienne Jacques2, Guillaume Bourque1

1: McGill University, Montreal, Quebec, Canada
2: Université de Sherbrooke, Sherbrooke, Quebec, Canada

Abstract

The IHEC Data Portal (http://epigenomesportal.ca/ihec) is the integrative online resource to navigate through datasets produced by the International Human Epigenome Consortium. As of May 2016, the Portal hosts over 7500 human datasets, for which more than 5000 are in the IHEC core set of assays. With an average of 150 unique sessions weekly, it has become the central access point to visualize and obtain IHEC datasets. In order to increase information quality and accessibility, multiple new features have been added over the last year.

First, datasets quality assessment features have been implemented. These include a visual correlation tool that was introduced last year, now released on the main portal server. A quality control pipeline was also developed, verifying datasets quality by evaluating multiple metrics. Using only publicly accessible tracks, it identifies potential problems such as incomplete coverage, high background noise, and poor correlation to other tracks with similar metadata. This pipeline will soon be added to the Portal data integration workflow, and will allow dataset-producing consortia to be warned of inconsistencies that could be validated, prior to release. An IHEC-wide analysis using this pipeline has demonstrated that even with large inter-consortia variability in methods for library preparation and downstream analysis, and differences in cell types, diseases and other factors, a useful qualitative assessment can be offered to data producers and users.

Other improvements include the addition of bi-yearly persistent releases and permanent sessions. Publicly-accessible tracks are now being served directly from the Portal enabling reliable navigation sessions, when compared to the previous distributed model, as there is no more dependency on multiple remote servers statuses. Users can also save their Portal sessions permanently, allowing them to obtain an ID that links to their datasets selection and filtering options. These IDs can then be directly citable in papers making use of the IHEC Data Portal datasets. Lastly, a publicly accessible Web API allows users to query available datasets metadata, in both JSON and human-readable formats.

The IHEC Data Portal is a service hosted by GenAP (https://genap.ca), developed and maintained by the McGill Epigenomics Data Coordination Centre (http://epigenomesportal.ca). It is funded under the CEEHRC, by the CIHR and by Genome Quebec, with additional support from Genome Canada. The correlation matrix computation approach was developed by the Université de Sherbrooke, and funded by NSERC. The computing and networking infrastructure, and part of the software development, are provided by Compute Canada and CANARIE.