Using the IHEC data through the Genomic Efficient Correlator (GeEC) tool

Jonathan Laperle[1,2], Marc-Antoire Robert[2], David Bujold[3], David Morais[3,4], Michel Barrette[4], Charlotte Bastin[2], Marie Harel[2], Christophe Morin[2], Guillaume Bourque[3,5], **Pierre-Étienne Jacques[1,2,4]**

1) Département d'informatique, Faculté des sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada
2) Département de biologie, Faculté des sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada
3) McGill University Genome Québec Innovation Center, Montréal, Québec, Canada
4) Centre de calcul scientifique, Université de Sherbrooke, Sherbrooke, Québec, Canada
5) Department of Human Genetics, McGill University, Montréal, Québec, Canada

Abstract
The Genomic Efficient Correlator (GeEC) tool, aimed at efficiently performing pairwise correlation of thousands of epigenomic datasets, has been used over the last year to pre-calculate correlation matrices that are incorporated to the IHEC Data Portal (http://epigenomesportal.ca/ihec/) (Bujold et al., submitted). GeEC was also proven useful for some members of IHEC (Breeze et al., submitted) to demonstrate that even if the experimental procedures are not necessarily consistent through all projects, the generated datasets are still overall highly comparable since they tend to cluster based on the assay type and sample cell type, rather than on the producing consortium. Moreover, the correlation data were used to identify potentially mislabeled or problematic datasets, as part of a quality control pipeline implemented in the IHEC Data Portal.

In addition to visualizing the pre-computed correlation scores from the Data Portal, users can now also compare their own epigenomic datasets to IHEC ones using a public version of GeEC recently launched. This could be useful, for instance, to help in the characterization of datasets, or as a quality control. The various features of GeEC, integrated into the Galaxy framework of the Genetics and genomics Analysis Platform (GenAP, genap.ca) project, include the support of many genomic file formats (bigWig, WIG, bedGraph, BAM), the possibility to compute correlations at different resolutions (from 1 Kb to 10 Mb), using different metrics (e.g. Pearson, Spearman), on different subsets of regions (e.g. genes, TSS, user-defined) or the complete genome, and the post-processing of the generated correlation matrix using for instance different clustering algorithms to display the results as an annotated heatmap and/or dendrogram. We also provide a user-friendly interface facilitating the selection of the desired datasets, and we plan to offer datasets from model organisms generated by international consortia such as modENCODE as well as data downloaded from GEO/SRA and uniformly processed. We will present the design and implementation of GeEC as well as a performance comparison with other tools and some of the key results obtained so far.