OpenGDC: standardizing, extending, and integrating genomics data of cancer

E. Cappelli, F. Cumbo, A. Bernasconi, M. Masseroli, E. Weitschek

Nowadays, several organizations are focusing on collecting heterogeneous biomedical big data. Among them, we focus on the Genomic Data Commons (GDC), an organization directly funded by the US government that collects a huge amount of experimental and clinical data about more than 30 types of cancer from several research centers in the United States. These data are maintained by GDC that supports two main research projects: The Cancer Genome Atlas (TCGA), containing experiments about thousands of patients affected by different kind of tumors, and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), including genomic cancer data of children.

Because of the heterogeneous nature of data (different formats and annotations), it is not easy to access and analyze them with state-of-the-art bioinformatics frameworks or programs. Here we propose OpenGDC, an open-source Java software to automatically extract and standardize public accessible GDC genomic and clinical data allowing researchers to easily perform ad-hoc integrated genomic analyses.

Our tool is able to convert these genomic data to the most common Browser Extensible Data format (BED) extending them with additional information retrieved from other public repositories (i.e., HUGO, GENECODE, and miRBase), comprising all the original GDC experimental data (i.e., Somatic Mutations, DNA methylation, Copy Number Variations, Gene-, Isoform-, and miRNA-Expression Quantification).

Additionally, OpenGDC provides a standardization procedure for clinical and biospecimen data (metadata) that includes in a single tabular file with also additional data retrieved from the GDC APIs. It is able to automatically manage data redundancy (e.g., the 'gender' of the patient related to a particular tissue is often replicated in both clinical and biospecimen sources).

The OpenGDC converted data are fully supported by bioinformatics frameworks like the GenoMetric Query Language (GMQL) system that exploits a SQL-like declarative language to make integrative queries on heterogeneous genomic data; a valid example about how our data standardization approach makes integrative analyses easy to be performed by ad-hoc bioinformatics frameworks.

In addition, we provide an open access FTP repository containing all the public accessible genomic and clinical data from GDC already converted into the BED format, resulting in more than 1.5 TB of data. An automatic procedure to maintain the repository up to date with GDC has been implemented. The repository is accessible through the following link: ftp://bioinformatics.iasi.cnr.it/opengdc/bed/.

Finally, OpenGDC is freely available at http://bioinf.iasi.cnr.it/opengdc/.