Nicolas Servant 21 October 2014

Hi-C pipeline

Development step 1

The current pipeline is available under SVN. A test dataset is available in the branches folder. The new beta-version (v2.1) of the pipeline is now available.

https://svn.curie.fr/svn/U900/pf-integration/ngs/tools/HiC-pip

Briefly, the new version was improved following the specifications summarized below :

- **Parallelization.** The Hi-C processing part from the raw reads to the list of valid interaction is now parallelized. Input reads have to be splitted and are processed independently. All files are then merged per sample to generate the contact maps.
- **Data format.** All unnecessary file conversions were removed. New and standard format for aligned reads is now used. Standard format for contact maps was proposed to improve load and data storage.
- **Generation of contact maps.** The script used to build the contact map was optimized in C++.
- **Workflow.** The current analysis workflow was improved to remove false positive mapped reads, and to remove duplicated sequences.

Altogether, these optimizations first allow us to run a complete Hi-C processing, and second to do it in a very efficient time.

The pipeline will be updated according to feedbacks from users. The following improvements are already planned for the coming weeks :

- ICE normalization at the end of the worflow
- Quality Control graphs during the Hi-C processing

Development step 2

The secondary analysis of Hi-C data are regularly updated through the HiTC Bioconductor package. However, the visualization of contact maps is still a missing functionality. To my knowledge, no stand alone and simple application is available today.

What would be this tool ?

- A stand-alone application which can be easily install on any laptop (with RAM !)
- A visualization tool able to :
    - Load a set of contact maps (a single map or a genome-wide map)
    - Switch from heatmap view to triangle view

        - Change contrast color for representation
        - Zoom in/out
        - Change the contact maps resolution according to zoom. For instance, for genome-wide visualization 1Mb bins should be enough. For TADs visualization, up to the max resolution of the input file.
- Add a genome track (BED/WIG/GFF)
- Link to external viewer (UCSC)

What would **not** be this tool ?

This tool is not for Hi-C analysis ! But only for Hi-C visualization.