INF-BIO5121/INF-BIO9121 - High Throughput Sequencing technologies and bioinformatics analysis

Autumn 2017

The home exam date is the 3rd of November. Before this day you have to solve the task described here, and prepare a 10- to 15-minute presentation to be held on the exam date. In addition to the presentation, you must be prepared to answer questions about the subject, to show intermediate data files and to explain the commands you have used.

Exam task title: Analyzing the relation between susceptibility loci of ADHD and chromatin states.

Chromatin profiling has emerged as a means for detection of regulatory activity, e.g. as discussed in the following article*: "Mapping and analysis of chromatin state dynamics in nine human cell types"* (Pubmed ID: 21441907). Along with this article, publicly available annotations of chromatin states along the human genome have been created for 9 different cell types (http://compbio.mit.edu/ENCODE_chromatin_states).

Genome wide association study (GWAS) is a statistical method for associating single-nucleotide polymorphisms (SNPs) with a specific trait, based on genetic variants from a sample of individuals with or without the trait. When used for diseases, the associated SNPs are considered to mark a region on the genome that influences the risk of disease. Attention-deficit/hyperactivity disorder (ADHD) is a brain disorder marked by an ongoing pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development. GWAS SNPs for ADHD, as well as other diseases and traits, are collected in *"A Catalog of Published Genome-Wide Association Studies"*, which is available at: http://www.ebi.ac.uk/gwas/.

- The ADHD SNPs might be expected to show a preference for certain chromatin states, in particular cell lines. The aim of the task is to do an analysis of the ADHD associated SNPs in relation to chromatin states, looking into whether the SNPs display a preference for certain chromatin states. One can choose whether to look at a particular cell type or across cell types, or a combination of the two. For more information about the cell types, you can visit the ENCODE cell type overview: https://genome.ucsc.edu/ENCODE/cellTypes.html

- The full analysis should be documented in a Galaxy Page, where all parts of the analysis should be well described and possible to reproduce[1]. The full analysis includes finding the appropriate data from the articles/databases, parsing the data into forms suited for analysis, performing the analysis and interpreting the results. (NB: The student should also provide relevant

[1] Note: The implementation of Galaxy Pages contains a bug complicating the creation of pages (resulting in content not being displayed in the output). In our experience, some simple measures avoids the issue: First, when beginning a page, add several blank lines and end the document with a period: "." After having embedding histories or other elements, make sure to manually jump to the next pre-entered line (by clicking) instead of creating a new line from the end of the line with the embedded content.

descriptive statistics of the datasets. One should discuss the implications of the different methods and parameters (e.g. null models) that have been used. The student should prepare a presentation of the results on the Galaxy Page, with the use of additional PowerPoint (or similar) slides if needed.

- Ideally, data should be found directly from the original source, and all steps of processing (conversion, lift-over, filtering etc.) from original data should be documented. If one is not able to achieve this, there are already processed versions of the data available, which could be used in further analysis. The Chromatin state segmentation data is available in the track repository of the Genomic HyperBrowser, for human genome build hg19, under the following track names: *"Chromatin:Chromatin state segmentation"*. An ADHD SNPs dataset can be downloaded from https://www.dropbox.com/s/0ge4ihdbg6s5o95/ADHD_hg19.bed?dl=0 (Important note: Use these data only after exhausting all options on trying to get the data from the original source. Document the steps and be able to show that you have made a reasonable effort before referring to the back up datasets.)

  Note that if data is processed from the original source as part of this task, it is not necessary that it end up exactly like the already processed versions mentioned here. What is important is just that the processing is well documented. Note also that a deep biological interpretation of the analysis is not a requirement - the focus is on data handling and basic statistical interpretations.

- Note that the positions in the GWAS catalogue correspond to the recent hg38 human genome assembly. In order to carry out an analysis on the hg19 genome (which the chromatin state segmentation tracks use), you will need to use a lift-over tool to convert to hg19. You can find a lift-over tool of your own choice. Or, use the lift-over tool at https://usegalaxy.org/. This supports only BED files, so you will need to convert the GWAS data into the BED format.

- Tip: track format conversion tools are available in the Genomic HyperBrowser (look for in the tools menu). One might use a more universal format as an intermediate step (e.g. when converting from GWAS original data to BED format).
- Tip: For processing the chromatin state segmentations datasets, the Galaxy "Select" tool in the “Filter and sort” section is relevant
- Tip: For the main analysis the “Analyze genomic tracks” should be used.
- Tip: For other parts of the analysis (e.g. descriptive statistics) certain GSuite tools can be useful.
- Tip: Analyses can take some time to execute, start working on the exam in time.
- Extensive collaboration with other students is NOT allowed.

- If you have any questions on the exam formulation, if something is not working correctly or you get stuck at some point, do not hesitate to send an email to borissim@ifi.uio.no.

Supplementary reading:

1. "The complex language of chromatin regulation during transcription" (PMID:17522673)
2. "Discovery and characterization of chromatin states for systematic annotation of the human genome" (PMID: 20657582)