# 1 Project Applied High-throughput Analysis

## 1.1 Goal

The main goal of this course is to be able to independently analyze “omics” data, which is therefore also the central topic of this project. For this group work, you should work in interdisciplinary pairs (i.e. with different background/track) whenever feasible. When you found a partner, you can use the “Discussion” section on Ufora to let us know with whom you are working. If you haven’t found a partner yet, you can also use this section to find someone (which will probably come in handy this year with the COVID pandemic limiting social interaction). Exceptions are possible upon request, e.g. for PhD students.

## 1.2 Task

Find at least four publicly available omics datasets that are somehow related. Analyze the data using the tools introduced during the course or similar, integrate/compare the results and discuss (from a biological and technical point of view). For example, you can evaluate whether expression differences observed in cancer may be attributed to either epigenetic or genetic differences; whether copy number variations associated with crop yields are fully reflected at the RNA level or somehow counteracted by epigenetics; to which extent gene expression measurement methods lead to the same conclusions in a specific cancer study... Most likely, it will be necessary to combine only partially related datasets, which is not a problem for this project’s sake.

Important conditions: a) the four datasets should be obtained by at least three different omics technologies, including at least one microarray and one sequencing based method. A comparison of high-throughput qPCR, 2 types of microarrays and RNA-seq for gene expression analysis is for example allowed as it involves three different platforms; b) the datasets should be actual omics datasets, qPCR experiments for very few loci, targeted resequencing etc. is not allowed as the amount of variables involved is too low; c) datasets used by different students should be avoided, once you have identified your data post a very brief description on Ufora in the "Discussion" section; including GEO, ArrayExpress, ... IDs) so your fellow students can easily identify no longer available data.

## 1.3 *The report*

Make a report based on the analyses, consisting of two parts: a) the general information, i.e. research question, data sources (including scientific references) and basic characteristics, used methods (also from an experimental point of view), most important results and figures, interpretation and discussion, conclusions. This part should be 5 to maximally 10 pages long and should be **generally understandable** for e.g. molecular biologists, medical specialists, ... Note that the goal of this part is to describe the different rationales behind the data-analysis and used methods/options. The first part should end with the integration/comparison/discussion of all results, to answer the central research question (if possible). The second part should contain the b) used codes and commands, packages, etc. Also "supplementary" figures and tables, clarifying a certain point made in part a) can be included here. This part should be well annotated, and is not featured by a length restriction (but should not contain actual data!). Essentially, part a) should contain what you've done and why, whereas part b) should demonstrate how you did everything and include the "supplements". Finally, don't forget to accurately refer to the relevant literature or tools.

## 1.4 *The oral examination*

Part of the oral examination will consist of a discussion of the report, focusing on a) does the student understand the essentials and rationale behind the used methods, b) are the results and conclusions from the report sufficiently supported by the analyses

## 1.5 *Further remarks*

- Using one method per dataset is sufficient, unless QC indicates that alternatives are required (QC results can be included in part b of the report, basic conclusions regarding QC in part a).
- Where relevant (e.g. for sequencing based methods) a basic overview of relevant data characteristics (e.g. coverages, mapping %ages) should be provided as a table in part b).
- For sequencing experiments: if count data is already available as such, these may be used, but additional raw data characteristics (e.g. from publication) should still be provided. Use of fpkm/rpkm and similar processed values is discouraged.
- Several methods (particularly GWAS related) are computationally intensive, this can be a valid reason to select an alternative method. Limiting the dataset (number of variables / cases) is allowed, but preferably only based on a clear rationale.
- Lack of significant results is a result as such, but discuss the possible reasons, from a biological point of view where relevant but particularly from a data-analytical point of view (e.g. bad quality data, low power, erroneous design)
- Where relevant for a dataset, a brief comparison of your results with those from the authors (which will typically at least slightly differ) can be included in part a).
- Your report may contain experimental data types or data-analytical methods (even web services) not discussed during the lectures, as long as they clearly belong to the state-of-the-art "omics" realm (note the "state-of-the-art"!). Note that the use of rpkm/fpkm (based methods) is discouraged, unless supported by a clear rationale.
- Feel free to contact the lecturer(s) regarding this project and its specific modalities, but don't expect troubleshooting from our side - this is an important part of the project itself. Your search engine is your best friend!
- Deadline Friday December 18th, midnight. Start early with this project, as problems often take lots of time to solve! Late or incomplete submissions should be avoided and are penalized, yet are preferred above no submission. In case of a late submission please inform us by mail.