# Final Project Outline

# ***Name:***

Please complete this project outline and submit via canvas. The instructors and TAs will review your plan to make sure that you have identified a reasonable and appropriate study and that the scope of work is appropriate for a final project. If you have any questions, please do not hesitate to reach out.

## Citation

* Add citation information for chosen study here
* [Include](about:blank) link to manuscript here
* Add other relevant links such as supplemental materials, key references, etc.

## Brief Summary of Published Study

*High-level summary (2-3 paragraphs) of the existing study. Be sure to add relevant contextual background details and information about how the study was performed*

## Biological Questions

* What *specific* biological questions were asked by the authors of this study using the dataset(s) that you will re-analyze?

# Data Availability

## Data Source

*Where are the raw and processed data available for download?*

## Accession Numbers / Links

*Please list all accession numbers or other identifying IDs for the data you will use for your project.*

# Sample Information

*Create a table of all sample information. Be sure to list sample generation information and as much relevant metadata including conditions of interest and*

## Sample Types

*What type of samples/data are available (RNA-Seq, ATAC-Seq, ChIP-Seq, etc). Be sure to describe how the libraries where made, and from what types of input. The more information you have here about sample generation, the better you will be able to select/identify tools for your workflow.*

# Study Design

## Biological Questions to ask

*What* ***specific*** *biological questions will you ask of these data for your project? The first may be similar to those asked by the authors of the original study, but at least one must be unique/novel*

## Explicit comparisons/contrasts

*What comparisons/contrasts will be made? Will you be comparing one (or more) conditions to another? Which samples will be used as a control and why? You may have multiple sections here describing different experimental designs for each biological question you are interested in asking.*

# Proposed Workflow

*For each subsection, try and identify additional resources that might need to be created or acquired to enact your workflow (e.g. reference genome/transcriptome for read alignment, gene annotations, etc)*

## Data Acquisition

*Where and how will you acquire the 1) raw sequencing data, and/or 2) processed data provided by the authors?*

## Data Organization

*How will the data be organized and made available within your project? Plan ahead for an organized project structure.*

## Read Alignment/Mapping

*What approaches/tools will you be using to map the reads to a reference (e.g. ungapped alignment, splice-aware alignment, pseudoalignment?) Can you use a publicly available reference index to map? If so, where is this located and how can you retrieve it? If not, what steps/tools/data will you need to construct a reference index required for mapping of the raw sequencing data? How will you assess the quality of your read mapping?*

## Feature Quantification/Tabulation

*Once you have mapped reads to a reference, how will you quantify features of interest? For DNA tasks, how will you define peaks/regions of interest? For RNA tasks, which features are you interested in (e.g. transcripts, aggregated genes, splice variants, etc)? What tools will you use to translate your aligned reads to quantified features? What other data sources will you need to fully annotate features of interest (i.e. Gene IDs alone may not be helpful, where can you find additional information to better annotate features with things like Gene Names, associated pathways or ontologies, etc). Describe how you will QC these quantifications to assess such things as expected distributions, feature selection (e.g. how will you remove poor or erroneous features, or select a subset of features for downstream processing?)*

## Data Analysis

*Add additional subheaders here as needed for additional tasks or analysis steps that are specific to your project workflow.*

*e.g. Data Preprocessing & cleaning, Feature QC, Feature annotation, Dimensionality reduction, Differential analysis, Gene set enrichment analysis, Pathway analysis, etc*

# Required Software/Tools/Packages

*Please try and be as comprehensive as possible here. This will be a useful resource for you to not only think about how the data will 'flow' through your project (ie. how tools will be chained together), and help you think about which tools can be used to answer your biological questions, but also provide quick references for their use and troubleshooting.*

*Use the provided examples below to help get you started.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Software/Tool/ Package*** | ***Task*** | ***Software Location*** | ***Link to***  ***Documentation*** | ***Link to Vignette/Tutorial*** | ***Notes*** |
| Hisat2 | Spliced read alignment |  | [Hisat2 Manual](http://daehwankimlab.github.io/hisat2/) | \* [Mapping with Hisat](https://wikis.utexas.edu/display/bioiteam/Mapping+with+HISAT2)  \* [Hisat2 tutorial](https://rpubs.com/selveyad/hisat2_tutorial) |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Anticipated Results (Deliverables)

*Describe what specific results/figures/etc you would like to generate to address your biological questions. How might these results be presented? Do you have any a priori hypotheses that you expect to test? Will this require specific statistics to test/reject?*

# Additional Metadata resources

*Here you should curate* ***and link*** *to any additional resources that you might need external to your target datasets. Examples might include:*

* Reference Genome/Transcriptome (what version)?
* Key gene sequences/annotation?
* Curated relevant gene sets for enrichment analysis?