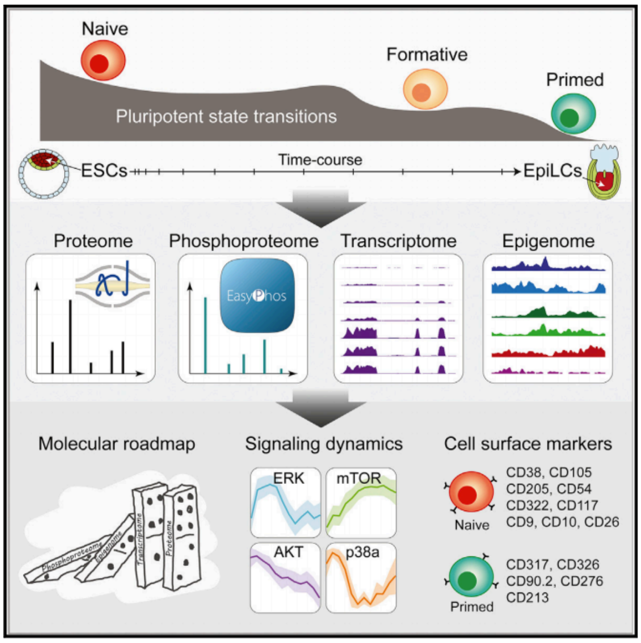
oOLET5602 Final Project Outline

Project: Transcription factor target gene prediction through multi-omics datasets

Aim: Prediction is a central application in many real-world data analyses. In this project, we will aim to apply classification techniques for predicting novel transcription factor target genes.

Background: Transcriptional regulation is a vital process in all living organisms. It is orchestrated by transcription factors along with other factors that in concert drive mRNA expression. Distinct transcriptional networks drive development, lineage specification, and cell fate decisions during early embryonic development (Theunissen and Jaenisch, 2017). Recent advances in omics technologies have made it possible to profile genome-wide transcriptional and epigenetic events for investigating transcriptional networks. A key goal is to identify the target genes of transcription factors that drive key transcriptional events over a course of time.



High-Temporal-Resolution Profiling of the Proteome, Phosphoproteome, Transcriptome, and Epigenome during ESC to EpiLC Transition (Yang et al., 2020)

The time-course multi-omic profiling of embryonic stem cell differentiation (Yang et al., 2019) provides a unique opportunity to reveal previous unknown aspects of stem cells during pluripotency progression. Multiple studies have shown transcription factors such as Sox2 (Masui et al., 2007) and Nanog (Masui et al., 2007) are key regulators of stem cell maintenance and differentiation. Oct4, Sox2, and Nanog (collectively called 'OSN' factors) are well known regulators of stem cell maintenance (Yeo and Ng, 2012). To this end, we aim to predict novel substrates of Sox2 and Nanog involved in the progression of pluripotency by using temporal multi-omic datasets.

- **Data set details:** (saved as Final_Project_ESC.RData)
- Transcriptome
- Proteome
- H3K4me3, H3K27ac, H3K27me3, H3K4me1, H3K9me2, PolII (Epigenome)

The main multi-omic datasets. Rows denote genes and columns denote the timepoint of differentiation.

Dataset	Details
Transcriptome	Time-course differentiation mRNA profiles at 0hr, 1hr, 6hr, 12hr, 24hr, 36hr, 48hr, and 72 hr of ESC differentiation. Time point 0hr corresponds to stem cells prior to differentiation stimulation.
Proteome	Time-course ESC differentiation proteome profiles at 1hr, 6hr, 12hr, 24hr, 36hr, 48hr, and 72hr compared to time point 0hr. Timepoint 0hr corresponds to stem cells prior to differentiation stimulation.
Epigenome	Time-course ESC differentiation epigenome profiles of 6 histone marks at 0hr, 1hr, 6hr, 12hr, 24hr, 36hr, 48hr, and 72 hr. The histone marks include H3K4me3, H3K27me3, Pol2, H3K4me1, H3K27ac, and H3K9me2. Time point 0hr corresponds to stem cells prior to differentiation stimulation.

- OSN_target_genes_subset

This variable contains identifiers of known Sox2 and Nanog target genes (included in Final_Project_ESC.RData).

Instructions:

Work with your group members and design a predictive model for identifying novel Sox2/Nanog substrates using the above datasets (and any other data sources that you found to be useful).

Summary of key aspects:

- Required outputs**:**
  (1) Probability of each gene as a substrate of Sox2/Nanog; and
  (2) A predicted list of substrates for Sox2/Nanog
- Evaluations:
  (1) Compare to previous predictions from Kim et al. (2020) (OSN_target_genes; included in Final_Project_ESC.RData).
  (2) Benchmark the prediction accuracy.
- Additional outputs:
  (1) All other visual analytic results.
  (2) All other tables of results that are useful.

Detailed instructions:

I. Data exploration. The features are the time-course expression profiles after induction of in vitro differentiation. We have time-course data from three types of omics data. This provides a unique opportunity to predict target genes of transcription factors at multiple omics levels. Explore the global structure of each type of omics data. We have provided you with a list of known Sox2/Nanog target genes. Visualise the expression profile of these target genes.

II. Target prediction and validation. Only a subset of substrates of Sox/Nanog is known. Develop a classification algorithm to predict additional substrates of Sox/Nanog. Consider how to deal with potential imbalance class distribution. Evaluate and benchmark your predictions by comparing your predictions to the prediction in our previous study (Kim et al., 2020).

Write a detailed Rmd (R markdown) report to document the project. Include problem description, your implementation in R codes, detailed comment of your R codes, and discussion on each decision you have made on performing classification. Prepare a presentation based on your project report.

Presentation and Report Rubric:

Marks for the presentation and report will be determined by how well the above key points are addressed. Each group will have a maximum of 10 mins for the presentation. Each group will receive a single mark for the presentation and the report and all members in a group will have the same mark unless group members are noted to contribute unequally. Both the presentation and reports will be peer-reviewed and marked. Note that when the number of total students enrolled is insufficient to form groups, the assessments (presentation and report) will be conducted and assessed individually.

The presentation should be recorded following the guidelines in `Basic Instructions on How to Add Narration to a PowerPoint Presentation.pdf` and submitted online. Your report should be knitted as a html report from RMarkdown (which shows your code and code output).

Key points	**Exceptional**	**Proficient**	**Fair**	**Developing**	**Inadequate**
(1) Data exploration	5	4	3	2	1
(2) Propose and execute a classification procedure for prediction	5	4	3	2	1
(3) Validation and benchmark of prediction results	5	4	3	2	1
(4) Presentation/report quality	5	4	3	2	1

References

1. Masui, S., Nakatake, Y., Toyooka, Y. et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625–635 (2007). https://doi.org/10.1038/ncb1589
2. Yeo, J., Ng, H. The transcriptional regulation of pluripotency. Cell Res 23, 20–32 (2013). https://doi.org/10.1038/cr.2012.172
3. Thorold W. Theunissen, Rudolf Jaenisch. Mechanisms of gene regulation in human embryos and pluripotent stem cells. Development 144, 4496–4509 (2017). https://doi.org/10.1242/dev.157404
4. Chappell, J. and Dalton, S. Roles for MYC in the Establishment and Maintenance of Pluripotency. Cold Spring Harb Perspect Med. 3(12) (2013). https://doi.org/10.1101/cshperspect.a014381
5. Yang, P., Humphrey, S., Cinghu, S. et al. Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. Cell Sys 8, 427–445 (2019). https://doi.org/10.1016/j.cels.2019.03.012
6. Kim, H., Osteil, P., Humphrey, S. et al. Transcriptional network dynamics during the progression of pluripotency revealed by integrative statistical learning. Nuc Acid Res 48(4), 1828–42. (2020). https://doi.org/10.1093/nar/gkz1179