EMAIL 1:

I am wondering if we could collaborate to look at our current data. We are reproductive biology lab and does not have much background on coding. My PhD student, Ryan Finnerty (cc’ed here), has the oviduct samples from mice that were collected at day 0.5, 1.5, and 2.5 after mating. We have performed bulk-RNA-seq, scRNA-seq, and luminal fluid proteomic analyses on these samples. Currently, we analyzed these datasets separately and found that there is an increased in pro-inflammatory cytokine signals at day 0.5 (the presence of the sperm in the oviduct) and then the signal quickly dissipated at days 1.5 and 2.5 (now the sperm are gone from the oviduct and embryos are present). When analyzed separately, the data seem to corroborate between transcription and translational levels. However, we are wondering if we could use machine learning technology in your lab to integrate all datasets that we have and we could obtain the similar outcome. This would strengthen our findings.

EMAIL 2:

Here's the follow-up email about what we discussed (my apologies for sending this later than I thought). We discussed that we could use ML to predict the relationship between predict bulk-RNA seq, scRNA-seq, or proteomic data or vice versa.

Action items for my lab:

* Please find attached file for the example of data file that you'll receive, likely .csv files that contain a combined normalized expression values for all datasets in one table. Would you like the genes/proteins that are not present in some dataset to be set as "null"?
* We will provide combined datasets for testing data and training data.

After receiving data from our lab, Yanli and Frimpong can help us:

1. find correlation value between datasets using Pearson or Spearman analysis
2. design AutoEncoder to predict the relationship between each dataset
3. test the data.

And of course, we will include all personnel that are involved in this project as co-authors for the manuscript.

EMAIL 3:

First of all, sorry for the delay getting you the data. This is taking longer than we anticipated.

Data files are in the shared folder: [​Folder icon ML GRN Cheng Collaboration](https://mailmissouri.sharepoint.com/:f:/s/WinuthayanonLab-Ogrp9/EgROC1-YndlDjGnbwgGCt8oBP-CGjw4NFpEFZMllA9E-pA?e=gKVhXA) Let me know if you have trouble accessing it. All of the data derived from mouse oviduct tissues.

**Data keys**🡺 also see experimental design.png in the folder:

* IA = tissues/cells from Infundibulum + ampulla regions of the oviduct
* IU = tissues/cells from Isthmus + Uterotubal junction regions of the oviduct
* Estrus = estrus stage of the ovarian cycle 🡪 mice ovulated naturally
* SO = superovulated 🡪 meaning that mice had been treated with hormones to exogenously force ovulation (this is usually done to synchronize mice and always done in the IVF clinic to ensure ovulation)
* 0.5 – Right after fertilization (considered 1st day of pregnancy)
* 1.5 – 2nd day of pregnancy
* 2.5 – 3rd day of pregnancy
* Pseudo = pseudopregnancy – this is when we mated the mice with vasectomized males (no sperm) to study whether the impact that we observed was due to the presence of sperm/pregnancy or changes in hormonal levels
* Preg = pregnancy

We try to not manipulate the data as much as possible so that we don’t have any bias going in before ML.

* Bulk RNA-seq data are represented as featureCounts from Galaxy (HISAT2 align) for all genes – not detected genes are listed as 0
* Protein data are represented as abundance values – not detected proteins are listed as NA
* scRNA-seq data are listed as either “score” = z-score of that group vs. the rest and also LogFoldchange (which is log 2FC of that group vs. the rest) – not detected genes are listed as 0

**The file contains multiple worksheets:**

1. Bulk RNA-seq literature – this is the data from [published literature](https://nam02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubmed.ncbi.nlm.nih.gov%2F33864778%2F&data=05%7C01%7Cyw7bh%40missouri.edu%7C194d2034f2a5495bb64308dba9bf179b%7Ce3fefdbef7e9401ba51a355e01b05a89%7C0%7C0%7C638290410422839353%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=eNrb81JXGbioFbxW%2B5fbpeWrj8RC82rigHfT%2BOOo1Ac%3D&reserved=0) from other group. They only have an n of 2 for each group. We downloaded the raw data from NCBI and did the similar run using Galaxy as same as bulk RNA-seq from our group.
   1. IA n1
   2. IA n2
   3. IU n1
   4. IU n2

1. Bulk RNA-seq Winuthayanon – this is data from our group. We have an n of 3 for each group
   1. Estrus IU
   2. Estrus IA
   3. IU 0.5 preg
   4. IU 0.5 pseudo
   5. IU 1.5 preg
   6. IU 1.5 pseudo
   7. IU 2.5 preg
   8. IU 2.5 psedu
   9. IU 3.5 preg
   10. IU 3.5 pseudo
   11. IA 0.5 preg
   12. IA 0.5 pseudo
   13. IA 1.5 preg
   14. IA 1.5 pseudo
   15. IA 2.5 preg
   16. IA 2.5 pseudo
   17. IA 3.5 preg
   18. IA 3.5 pseudo

1. Protein from luminal fluid in the oviduct at different timepoints. For this dataset, the fluid inside the oviduct tube was flushed and collected for mass spec analysis. This one we only have ASSESSION number, but not the protein name. Data from each time point were a pool of 3 mice.
   1. Estrus
   2. 0.5
   3. 1.5
   4. 2.5
   5. SO estrus
   6. SO 0.5
   7. SO 1.5
   8. SO 2.5

1. scRNA-seq – we found 10 cell clusters of cells in the oviducts. Data were concatenated/harmonized for all timepoints (0.5, 1.5, 2.5 preg) and all regions (IA and IU). The UMAPs of cell clusters (Leiden), Times (0.5 preg, 1.5 preg, 2.5 preg, and Ctrl (0.5 pseudo)), and regions are also in the shared folder. **All scRNA-seq data in the excel file are “SO” (superovulated).**Data in excel file from each cluster contain all timepoints and all regions. Data from each timepoint contain all cell clusters from all regions. Each group is a pool of cells from n of 3-4 mice.
   1. 0\_secretory1 vs rest
   2. 1\_secretory2 vs rest
   3. 2\_cilated vs rest
   4. 3\_Fibroblast1 vs rest
   5. 4\_Immune1 vs rest
   6. 5\_Ephx2 vs rest
   7. 6\_Endothelial
   8. 7\_Immune2 vs rest
   9. 8\_Secretory3 vs rest
   10. 9\_Fibroblast2 vs rest
   11. 0.5 pseudo vs rest (we don’t have 1.5 pseudo or 2.5 pseudo for this dataset)
   12. 0.5 preg vs rest
   13. 1.5 preg vs rest
   14. 2.5 preg vs rest
   15. IA all timepoint vs rest
   16. IU all timepoint vs rest

scRNA-seq datasets not included here:

* + - We also have another dataset of scRNA-seq “estrus entire tube (did not separate into IA or IU)” natural ovulation that we didn’t include here yet.
    - We also have another dataset of scRNA-seq “2.5 preg” that is not superovulated that we didn’t include in here yet.