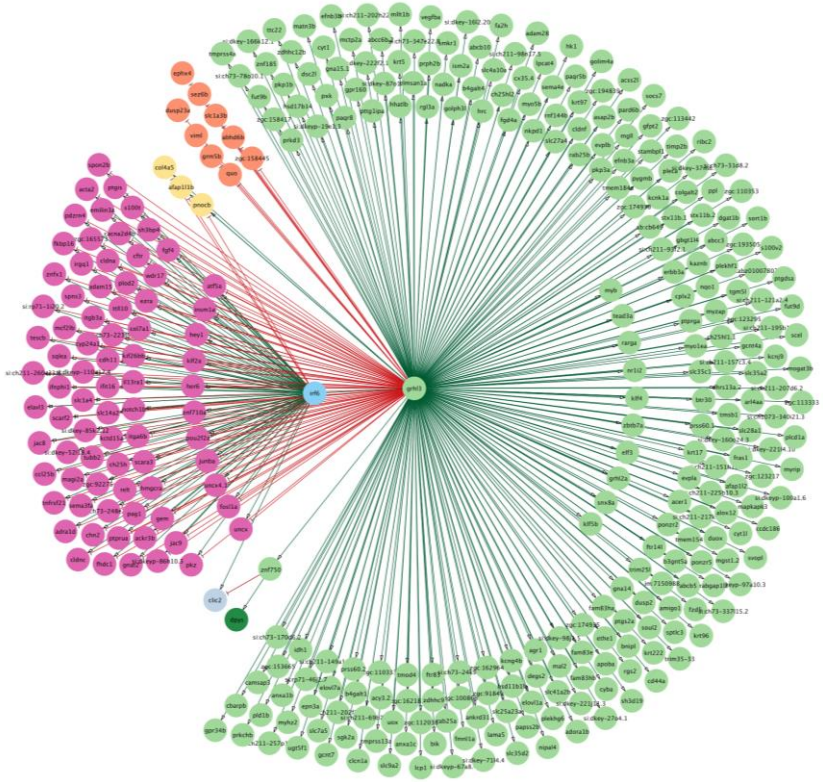
Inferring full scale machine learning (ML) model of transcriptional regulatory network (TRN) underlying specification of the zebrafish enveloping layer (periderm)

Lira, Ph.D. | Sr. Manager in Stat.

October 20th, 2022

PHARMALEX

Outlines



- Motivation of the study
- Data Description
  - scRNA-seq
  - scATAC-seq
- Transcriptional Regulatory Network (TRN) Construction
- Validation of the TRN
  - Identification of Gold Standards (GS)
  - Computation of Partial area under Precision-Recall curve (PA-PRC)
- Summary and Conclusions
- Key References

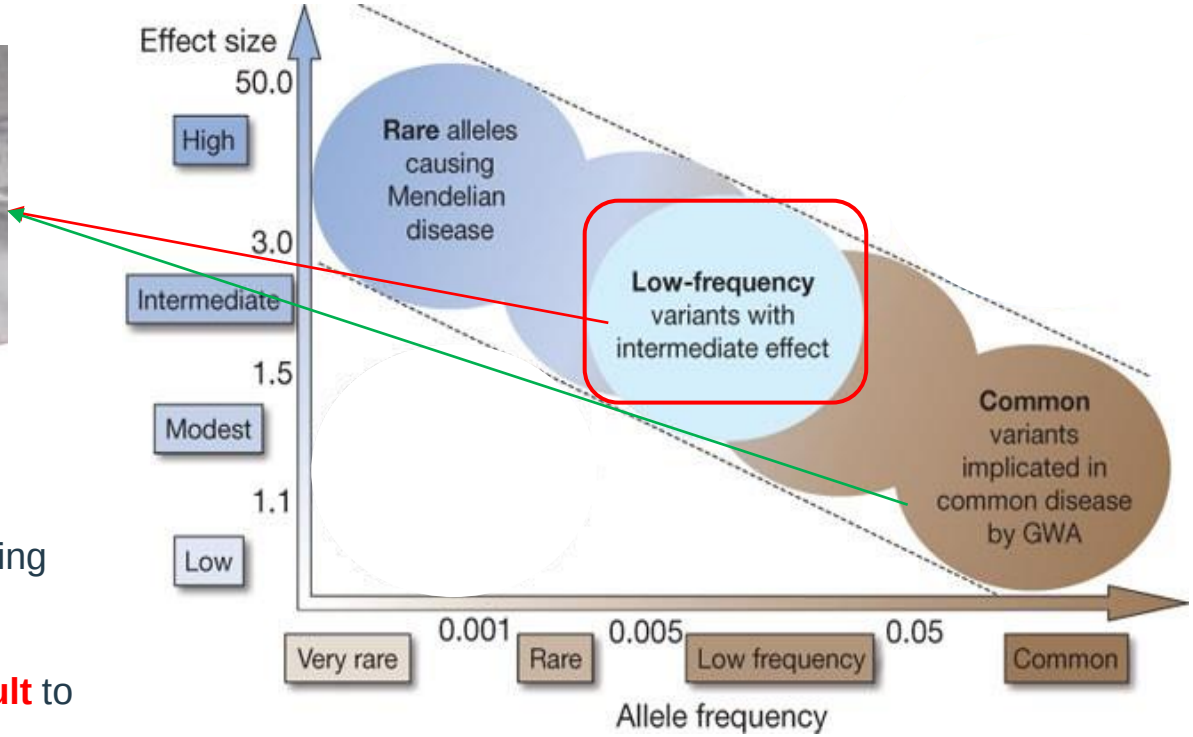
Introduction I

- Orofacial Cleft (OFC) may cause problems with feeding, ear disease, speech and socialization.
- It is curable with a relatively simple surgery but in some parts of the world still unaffordable: high costs for the family and eventually for society.
- Knowledge gap: most of the heritable risk for **non-syndromic OFC** has not been assigned to a gene.

Syndromic

Non-Syndromic

Introduction II

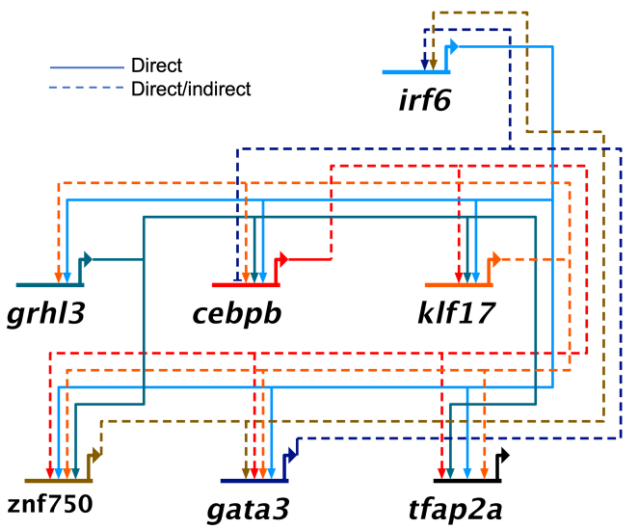


- While **GWAS** is appropriate to detect the heritable risk stemming from common variants, **low-frequency variants with intermediate effect are difficult** to be identified by GWAS.

Manolio et al. Visscher, 2009

Introduction III

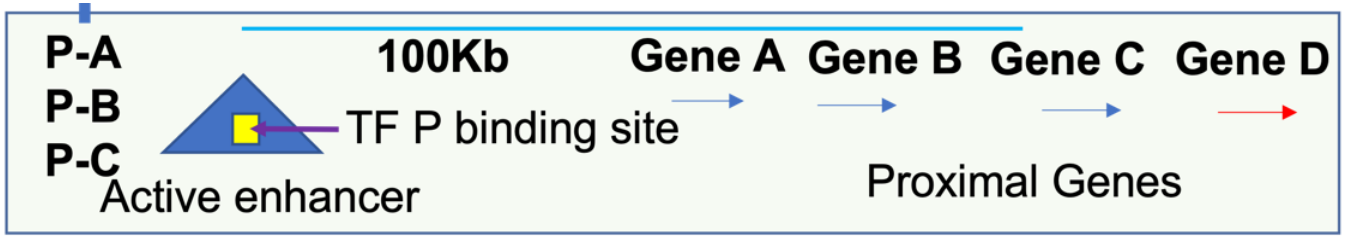
- To overcome the limited capability from GWAS, transcriptional regulatory networks (TRN) have been proposed to describe the control of gene expression variations by transcription factors (TFs).



Data Description

- scRNA-seq gene expression
  - 3,787 target genes X 394 EVL cells

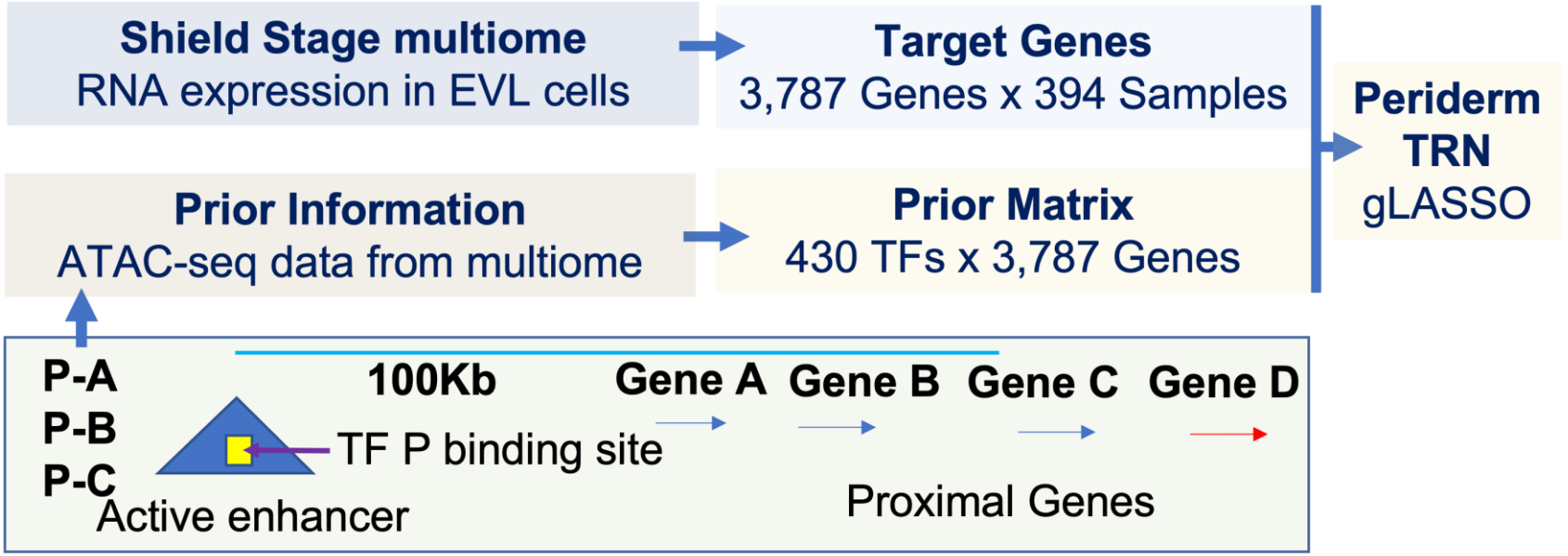
- scATAC-seq open-chromatin data
  - Generated binary matrix involving 430 TF's X 3,787 genes in the following cis-regulatory.



  - Hi-C (a high-throughput genomic and epigenomic technique, a derivative of a series of chromosome conformation capture) data
    - Was considered to detect genome-wide chromatin interactions in trans-regulatory, but data specific to EVL cell-type couldn't be found.

TRN (Transcriptional Regulatory Network)

- Overview (Miraldi et al.)



Graphical LASSO (gLASSO)

- A sparse penalized maximum likelihood estimator (MLE) for the precision matrix (inverse of covariance matrix, $\Theta = \Sigma^{-1}$) of a multivariate elliptical distribution is obtained to maximize the log-likelihood,

$$\log det\ \Theta - tr(S\Theta) - ||\Theta * P||_1$$

Where S denotes the empirical covariance matrix and P is the regularization amount for each variable (Friedman et al. 2007). The entries of the P matrices were one of two values: the non-negative value ρ for TF-gene interactions without evidence from prior; otherwise bias*ρ for TF-gene interactions with support in the prior.

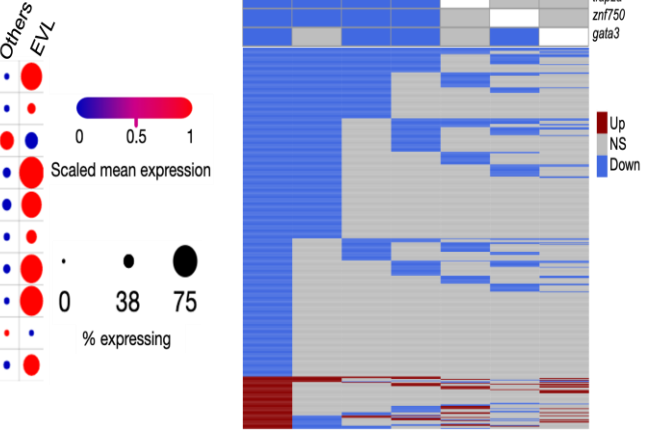
The penalty parameter ρ was determined by cross-validation.

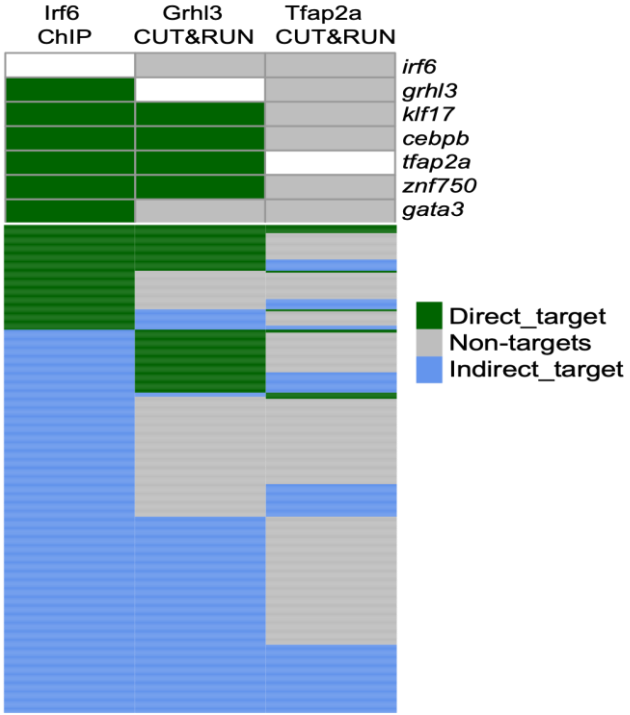
We tested five different bias depending on prior reinforcement strength.

Generation of Gold Standard (GS)

- Identification of key TF's
- Discovery of genes with enriched expression in EVL cells
- Identification of genes regulated by key TF's from bulk-RNAseq
- Identification of genes directly regulated by Irf6, Grhl3, or Tfap2a by comparing with ATAC-seq peaks

PWM	% of targets	p-value	TF family	Best match
	27.3	1e-348	Grhl	Grhl3
	25.2	1e-233	Tead	Tead3a
	14.9	1e-163	Fos	Fosab
	14.0	1e-145	Klf	Klf17
	14.6	1e-85	Tfap2	Tfap2a
	20.0	1e-71	Ets	Ets2
	10.0	1e-68	Gata	Gata3
	12.3	1e-66	Cebp	Cebpb
	25.6	1e-34	Nr2e3	--------
	5.0	1e-14	Irf	Irf6





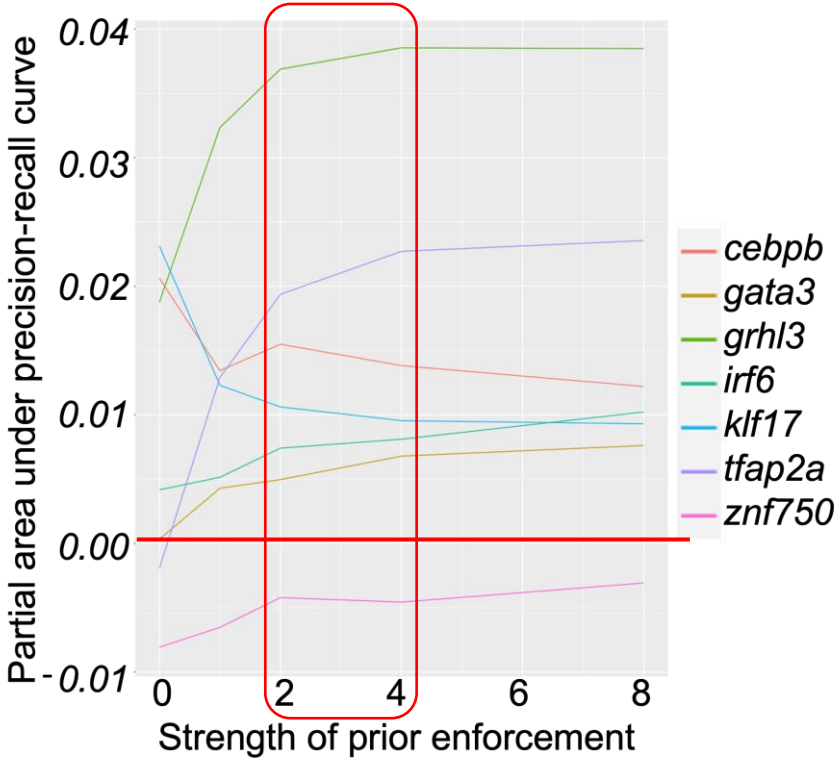
Tuning of computationally inferred TRN by using GS network

- By comparing the significant pairs selected by gLASSO with TF-gene relations from GS outpus, true positives (TP), false positives (FP), and false negatives (FN) were estimated.

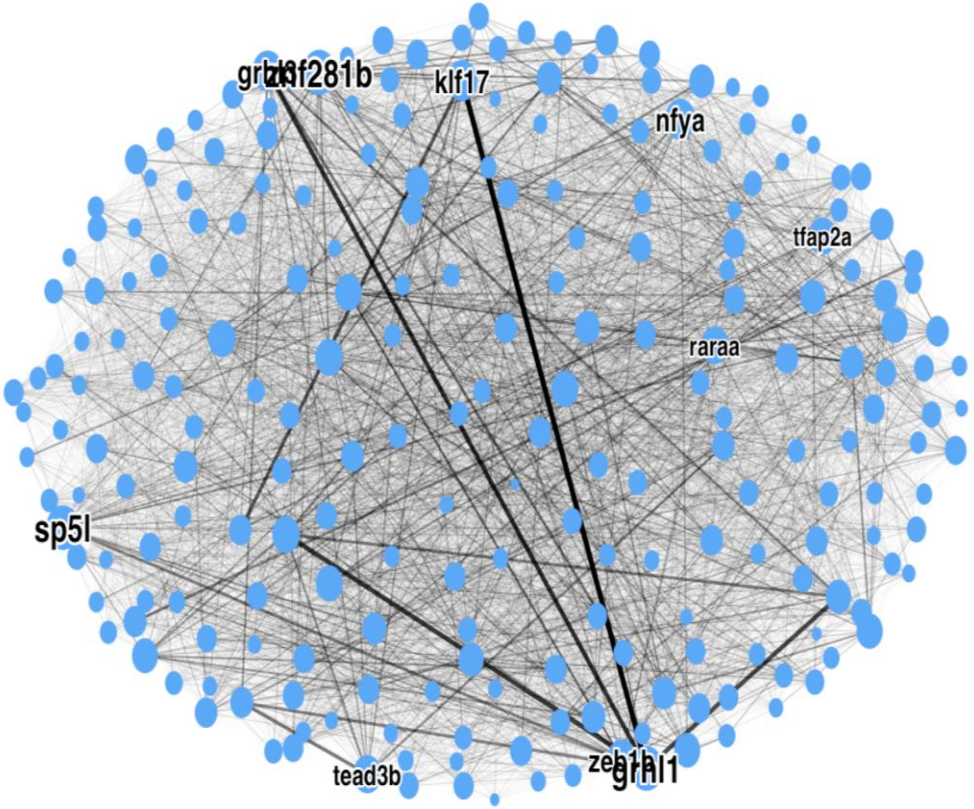
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)

- 0 partial area indicates random guessing (baseline) while the partial area >0 shows improvement over random guessing.

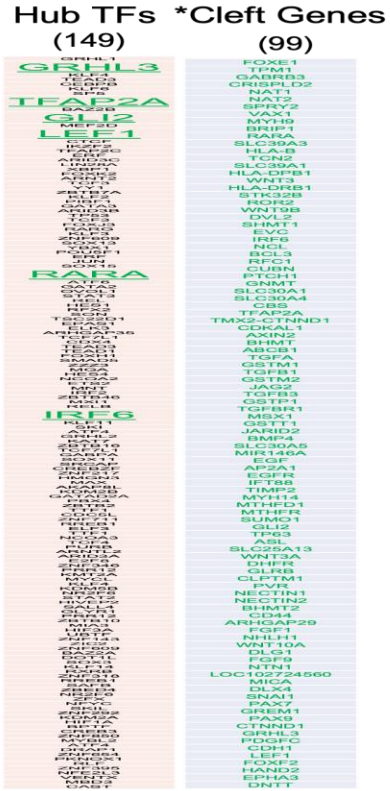
- Overall, moderate prior reinforcement derives better performance than others (no reinforcement or strong reinforcement).

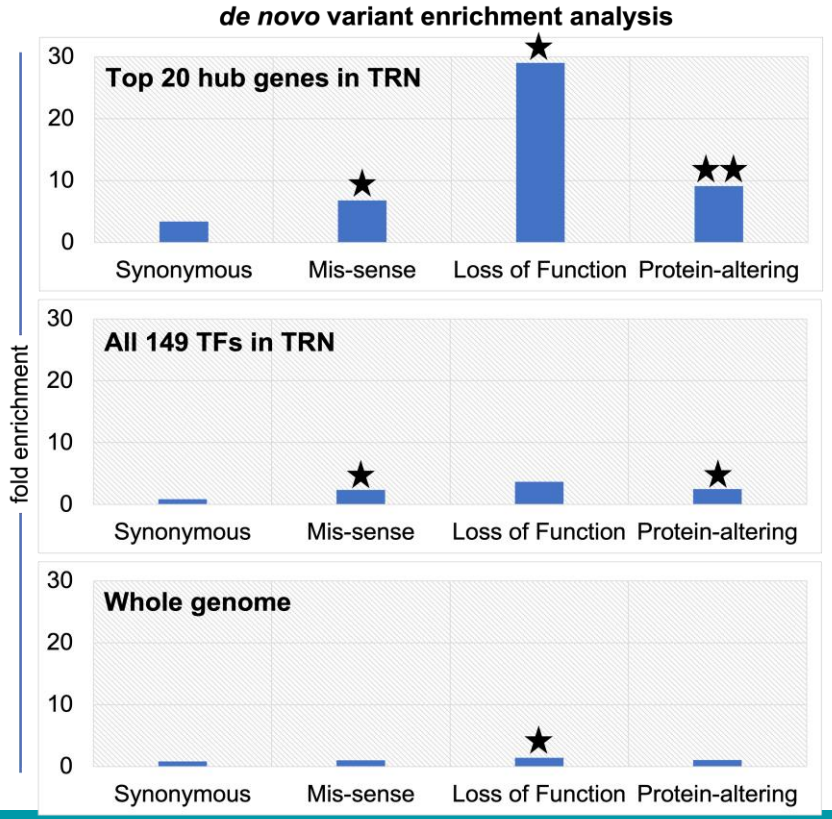


TRN with 2 Prior Reinforcement



Highly-connected TFs in the EVL TRN are enriched for known cleft genes and *de novo* variants in OFC patients





* https://toppgene.cchmc.org/ [toppgene.cchmc.org]

Conclusions

1. Highly connected (hub) genes in the EVL TRN, relative to randomly selected TFs in it, are enriched for 99 known OFC risk genes* (p-value ≈ 3.08e-05). * https://toppgene.cchmc.org/ [toppgene.cchmc.org]

2. Hub genes of EVL TRN, relative to all genes in the human genome, are also enriched for loss of function (LOF) *de novo* variants detected in OFC patients (p-value ≈ 2.28e-03).

3. **These results indicate that other highly connected genes in the EVL TRN are candidates to be OFC risk genes.**

Key References

- Miraldi et al. 2019 Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. Genome Research. 29:449-463.

- Friedman et al. 2007 Sparse inverse covariance estimation with the graphical lasso. https://tibshirani.su.domains/ftp/graph.pdf

Acknowledgements

Cornell lab:
Robert Cornell
Priyanka Kumari
Frank Radella
Jenny Jiang
Abby Kay
Sunil K Singh

Former lab members:
Huan Liu
Kaylia Duncan
Lira Pi
Greg Bonde
Colin Kenny

Collaborators:

University of Iowa:
Patrick Breheny
Annika Halverson

Emory University:
Elizabeth Leslie
Sarah Curtis

Children's Hospital of Philadelphia/ Harvard Medical School:
Eric Liao
Edward Li

R01 DE023575
R01 DE027983

National Institute of Dental and Craniofacial Research

NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases