Resolving Higher Order Interactions Between Quiescence-Associated Transcription Factors

Sean Cheah[1], Mithun Mitra[1,2], and Hilary A. Coller[1,2]

[1]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA

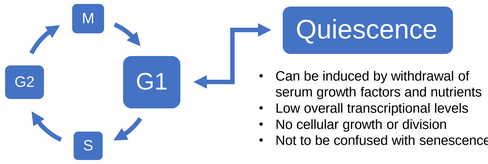
[2]Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, CA

Highlights

- Quiescence is the reversible arrest of cell proliferation and is important across biological processes ranging from stem cell maintenance to cancer cell dormancy.
- To explore the quiescence-associated regulatory networks, we devised a pipeline that identifies potential cooperative binding interactions between transcription factors (TFs) using ChIP-Seq data.
- This was done by training a binary classifier to distinguish “real” from “fake” TF binding patterns[1]
- The resulting model was interpreted using SHAP[2], which explains model outputs in terms of per-feature contributions.
- The pipeline was used to explore TF interactions in the context of differentially expressed (DE) genes from in-house RNA-Seq data for quiescent and proliferating primary human dermal fibroblasts.

Introduction

1 Quiescence: Reversible Cell Cycle Arrest[3,4,5]



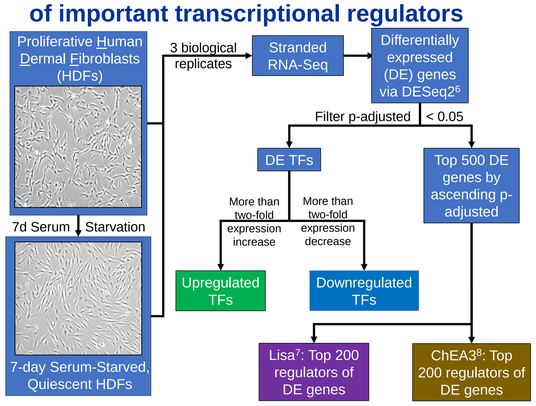
- Can be induced by withdrawal of serum growth factors and nutrients
- Low overall transcriptional levels
- No cellular growth or division
- Not to be confused with senescence

Senescence	Quiescence
• Irreversible • Cause: cellular overactivation • Phenotype: cellular hypertrophy and loss of proliferative potential	• Reversible • Cause: lack of growth stimulation • Phenotype: can be restimulated to induce proliferation

2 Goals of This Study

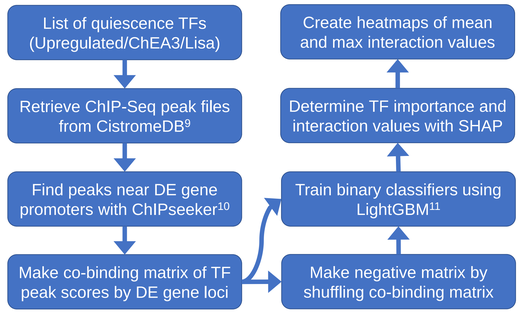
1. **To identify transcription factors (TFs) that experience significant differential expression during quiescence**
2. **To analyze publicly-available ChIP-seq data for differentially expressed (DE) TFs to determine their possible gene targets**
3. **To identify cohorts of TFs that cooperatively regulate target genes that are differentially expressed during quiescence**

3 Quiescence RNA-Seq data enabled identification of important transcriptional regulators



Methods

4 Workflow relies upon publicly available tools



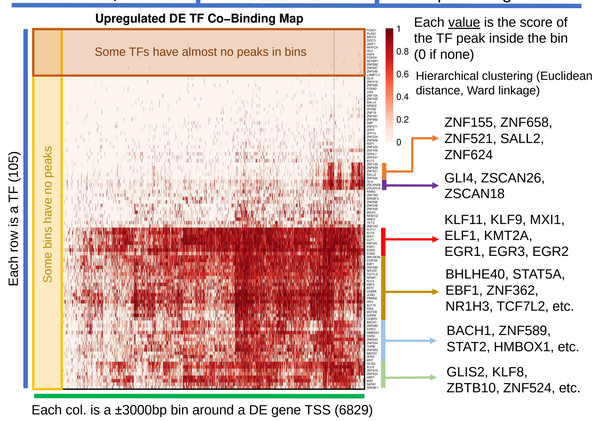
5 Limited ChIP-Seq data for quiescence-related TFs necessitated cell-line agnostic analysis

Genes	Up	Down	Lisa	ChEA3
All	3265	3564	–	–
TFs	158	61	200	200
Data*	105	42	199	135
HDF+	2	0	13	5

*High quality TF ChIP-Seq data from any cell type in CistromeDB
+Human dermal fibroblast ChIP-Seq data

6 Groups of TFs have similar binding patterns near promoters of differentially expressed genes

1. Rank peaks: By ascending MACS2 q-value
2. Score peaks: $\frac{(\#\text{ of peaks}) - \text{rank}}{\text{rank} - 1}$
3. Subset peaks: Isolate peaks within 3000bp of DE gene TSS



Each value is the score of the TF peak inside the bin (0 if none)

Hierarchical clustering (Euclidean distance, Ward linkage)

ZNF155, ZNF658, ZNF521, SALL2, ZNF624

GLI4, ZSCAN26, ZSCAN18

KLF11, KLF9, MXI1, ELF1, KMT2A, EGR1, EGR3, EGR2

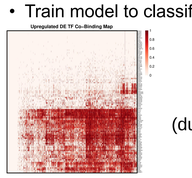
BHLHE40, STAT5A, EBF1, ZNF362, NR1H3, TCF7L2, etc.

BACH1, ZNF589, STAT2, HMBOX1, etc.

GLIS2, KLF8, ZBTB10, ZNF524, etc.

7 Reframing the search for TF interactions as a binary classification

- Train model to classify “real” from “fake” TF co-binding patterns

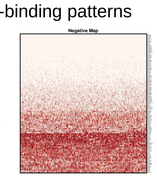


Positive Set (1)

True co-binding matrix containing real TF co-binding patterns

Each TF is a feature

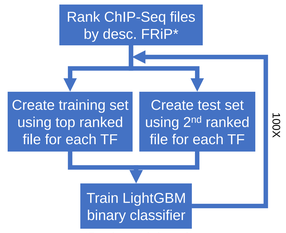
Each DE gene bin (duplicated for negative set) is a sample



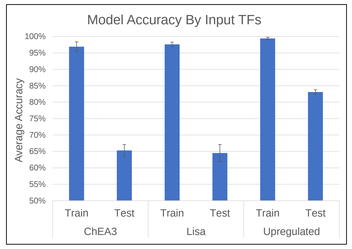
Negative Set (0)

Independently shuffle each row of co-binding matrix to disrupt real TF co-binding patterns

8 Using only upregulated TFs consistently yielded the highest training and test accuracies

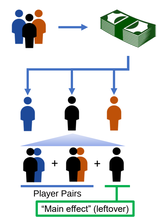


For each TF set, trained 100 models with different random negative sets to average out chance interactions



Average accuracy scaled to 50%, the baseline accuracy of classifying all samples as 0 (or all samples as 1)

9 SHAP interaction values quantify TF co-binding



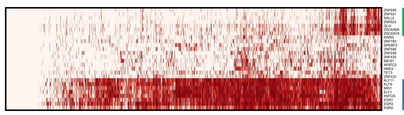
Treat feature values as players cooperating to win a payout: the model output.

SHAP values fairly allocate the payout to individual players by contribution

Interaction values fairly allocate each player’s SHAP value to player pairs (symmetric) by contribution

About SHAP
- Based on Shapley values (game theory)
- Accurate + consistent feature attribution
- Calculated locally (per sample)
- TreeSHAP: decision tree implementation

Interpretation
- SHAP value for a TF represents its “combinatorial potential”
  - Depends on # and significance of co-binding events with other TFs
- Interaction values represent the significance of a TF pair’s co-binding



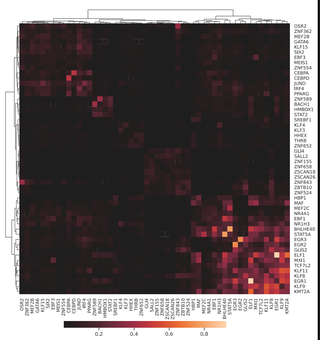
Strong local co-binding
High **max** interaction value across all peaks

Strong global co-binding
High **mean** and/or **max** interaction value across all peaks

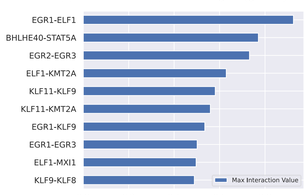
Results

10 Top SHAP interaction values corresponded to both known and novel upregulated TF interactions

Maximum Interaction Value

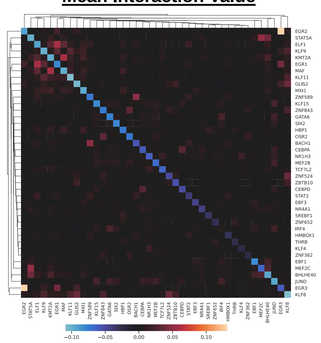


- Showing only TFs with at least 1 maximum interaction value >0.1 (same in 11)
- Many maximum (local) interaction effects disappear in heatmap of averages (global)

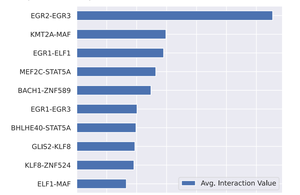


- Top 10 maximum interaction values shown
- Reported as lists in 11 (high to low)

Mean Interaction Value



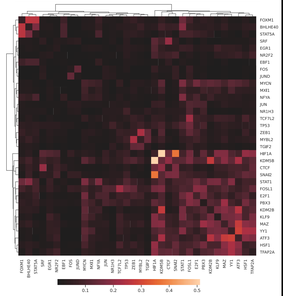
- Showing only TFs with at least one average interaction value >0.01 (same in 11)
- Nearly all main effects are negative: model’s predictive power derives from TF interactions



- Top 10 average interaction values shown
- Reported as lists in 11 (high to low)

11 Different input TF sets resulted in different top predicted interactions

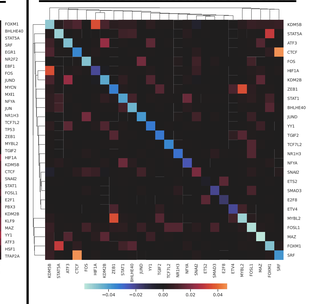
ChEA3 Max Interaction Value



Top 10 Maximum Interactions

HIF1A-KDM5B, HIF1A-SNAI2, ATF3-YY1, ATF3-MAZ, KDM2B-KDM5B, BHLHE40-FOXM1, FOXM1-STAT5A, ATF3-KDM2B, HSF1-YY1, MYBL2-ZEB1

ChEA3 Mean Interaction Value



Top 10 Average Interactions

CTCF-SRF, HIF1A-KDM5B, MYBL2-ZEB1, FOXM1-STAT5A, ATF3-KDM2B, *SNAI1**, FOS-JUND, NFYA-STAT1, ATF3-YY1, E2F8-ETS2

*Rare case of a main effect being highly positive

Lisa Max Interaction Value

Top 10 Maximum Interactions

KLF1-NFE2, HIF1A-KDM5B, CTCF-RAD21, KDM5B-SP1, RAG2-SUPT5H, SMC1A-SP1, KDM5B-NFE2, HIF1A-RAD21, RAD21-SMC1A, KLF1-SP1

Lisa Mean Interaction Value

Top 10 Average Interactions

KLF1-NFE2, E2F3-IRF1, IRF1-MED12, HIF1A-KDM5B, CTCF-RAD21, RAG2-SUPT5H, CHD2-SP1, E2F3-MED12, PHF8-RBBP5, EGR3-IKZF2

Blue = downregulated in quiescence (p-adj. < 0.05); Green = upregulated in quiescence (p-adj. < 0.05)

Conclusions

Main Findings

- Model was highly sensitive to list of input TFs and availability of ChIP-Seq data
- SHAP model interpretation identified previously characterized TF associations
  - EGR2 and EGR3: frequently co-mentioned in papers related to immunology
  - HIF1A and KDM5B: frequently co-mentioned in papers related to cancer
- SHAP model interpretation offered strong evidence for intriguing novel TF interactions
  - BHLHE40, a transcriptional repressor, and STAT5A, a transcriptional activator
  - CTCF, which binds to insulators, and SRF, which binds to the serum response element in promoters of target genes

Future Directions

- Train models on other, potentially more informative TF/gene sets
  - Top Lisa/ChEA3-predicted regulators of genes upregulated in quiescence
  - Top TFs by significance of differential expression (instead of log2 fold change)
- Scour other ChIP-Seq databases for TFs that are missing data
- Train models on cell-type specific ChIP-Seq data (not human dermal fibroblast)
- Characterize differentially expressed genes where TFs have high interaction values
- Integrate chromatin data from ATAC-Seq experiments on quiescent cells
- Develop new model to predict differential expression using TF binding data
- Combine results to piece together the transcriptional regulatory network governing the differential expression observed in quiescence

References

1. Gerstein M., Kundaje A., Hariharan M. et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91-100.
2. Lundberg S. M. and Lee S. (2017). A Unified Approach to Interpreting Model Predictions. *NIPS 2017*, 4765-4774.
3. Coller H.A., Sang L., and Roberts J. M. (2006). A New Description of Cellular Quiescence. *PLOS Biology*, **4**(3), e83.
4. Lemons J. M. S., Feng X., Bennett B. D. et al. (2010). Quiescent Fibroblasts Exhibit High Metabolic Activity. *PLOS Biology*, **8**(10), e1000514.
5. Blagosklonny M. V. (2011). Cell cycle arrest is not senescence. *Aging*, **3**(2), 94–101.
6. Love M.I., Huber W., and Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.
7. Qin Q., Fan J., Zheng R. et al. (2020) Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol*, **21**, 32.
8. Keenan A. B., Torre D., Lachmann A. et al. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research*, **47**(W1), W212–W224.
9. Zheng R., Wan C., Mei S. et al. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Research*, **47**(D1), D729–D735.
10. Yu G., Wang L., He Q. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**(14), 2382-2383.
11. Ke G., Meng Q., Finley T. et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS 2017*, 3149-3157.