GR-Trace · glucocorticoid response

Kenyon Shi, Kevin Shi

1 · PROBLEM

what and why

Glucocorticoid is a steroid-hormone pathway involved in stress and inflammation, and synthetic glucocorticoids such as dexamethasone are widely used as medicines. After a cell is exposed to a glucocorticoid, some genes respond quickly while others change only hours later. We want to test whether those early transcriptional changes contain enough information to forecast the later response, and whether a simple, interpretable model can reveal which early patterns matter most.

2 · APPROACH

model / method, and why this one

The core experiment is intentionally simple: use early changes in gene expression after dexamethasone exposure to predict each gene’s later response. We will first reproduce the basic time-course response from the published data, then compare an interpretable linear model with a nonlinear tree-based model. • For each gene, calculate dexamethasone-versus-control expression change at 1, 3, 5, 7, 9 and 11 hours, averaging the two replicates at each time point. • Primary prediction task: use the 1, 3 and 5 hour measurements plus simple trajectory features (early slope and variability) to predict the mean late response at 9 and 11 hours. • Models: Elastic Net regression as the main interpretable baseline and Random Forest regression as a nonlinear comparison. Gradient boosting will be added only if the first two models are stable. • Secondary analysis: group genes into sustained up-regulated, sustained down-regulated and weak/transient- response categories, then compare which early patterns separate these groups. • Interpretation: inspect Elastic Net coefficients, permutation importance, and representative gene trajectories instead of relying only on a single performance score. • External stress test: compare the strongest sustained-response genes with an independent dexamethasone RNA-seq experiment in primary human airway smooth muscle cells (GSE52778). This dataset will not be used to train the model.

3 · DATASET

source, size, preprocessing

Primary dataset: NCBI Gene Expression Omnibus GSE104714. Human A549 cells were exposed to dexamethasone or vehicle control for 1, 3, 5, 7, 9 and 11 hours, with two replicates per condition at each time point (24 RNA-seq samples total). GEO provides a processed expression matrix, so the first version of the project can stay self-guided and avoid a full raw-read alignment pipeline. Preprocessing: use the processed matrix; match dexamethasone and control samples by time point; filter genes with very low or unstable expression; compute log-scale treatment effects; standardize model inputs using training data only. If the processed values are not suitable for direct fold-change analysis, fall back to the raw- count workflow described by the original study. External dataset: GSE52778 contains RNA-seq from four primary human airway smooth muscle cell lines under untreated, dexamethasone, albuterol, and combined treatment conditions. Only the untreated-versus- dexamethasone comparison will be used as a cross-dataset biological check. Kenyon Shi + Kevin Shi | GR-Trace

4 · METRICS

how you’ll know it worked

observed late gene response. • Strong-response classification: precision-recall AUC, ROC-AUC, F1 score, and confusion matrix after freezing a biologically reasonable response threshold before final testing. • Stability: repeat train/test splits and bootstrap confidence intervals so one lucky split does not determine the conclusion. • Cross-dataset check: percent sign agreement and overlap among the top sustained glucocorticoid-responsive genes in GSE52778. • Baseline: compare every model with simple rules such as “use the 5-hour response as the 10-hour prediction.” A model is only useful if it beats this baseline consistently.

· Late-response regression: Spearman correlation, mean absolute error (MAE), and R-squared between predicted and

5 · COMPUTE

`training time, memory, dependencies`

The processed GEO files are small, so the core analysis is CPU-friendly. A typical run should fit comfortably in 8- 16 GB RAM and finish in minutes; repeated bootstraps, hyperparameter sweeps, and optional extensions can run on Exea compute. Dependencies: Python 3.11+, pandas, NumPy, SciPy, scikit-learn, Matplotlib, and Jupyter. Optional: statsmodels, XGBoost, and R/DESeq2 if raw-count reprocessing becomes necessary. Estimated storage: under 2 GB including processed data, notebooks, cached results, figures, and environments.

6 · DELIVERABLES

`model, paper, demo`

reproducible result table comparing the baseline, Elastic Net, Random Forest, and any optional model. • Figures showing representative response trajectories, predicted-versus-observed late response, model errors, and cross-dataset comparison results. • A 5-8 page research paper / preprint describing the biological question, method, results, limitations, and what did or did not generalize. • A short presentation and lightweight gene-response explorer that lets a user select a gene and view its time-course and model prediction.

· A clean GitHub repository with data-loading, preprocessing, modeling, evaluation, and figure-generation code. • A

7 · RISK

`what breaks first, and plan B`

what breaks first, and plan B What could break Why it matters Plan B Only two replicates per time point Fold-change estimates may be noisy. Average replicates, filter unstable genes, bootstrap results, and avoid claiming sample-level clinical prediction. Early response does not predict late The headline model may have modest Treat this as a result: identify which response well accuracy. response patterns are predictable vs. delayed, and shift emphasis to trajectory clustering. External dataset disagrees Cell type and experiment differences may Report it as a cross-cell-type stress test; keep dominate. the main conclusion limited to A549 cells. Bioinformatics scope grows too fast Raw RNA-seq processing could consume the Stay with GEO processed matrices and two project. simple models. Raw-read processing and pathway analysis are optional extensions. Kenyon Shi + Kevin Shi | GR-Trace SCOPE / RESEARCH INTEGRITY This is an exploratory computational biology project, not a clinical model. The main claim will be limited to whether early transcriptomic response can forecast later response within the available cell experiments. Both researchers will keep a contribution log, understand the code and analysis, verify every reported result, and disclose any use of generative AI for coding or writing assistance. 4-6 WEEK EXECUTION PLAN Week Main goal 1 Read the original studies, download the processed matrices, reproduce basic dexamethasone-vs-control response plots, and lock the preprocessing steps. 2 Build the simple 5-hour baseline and Elastic Net model. Check whether early response actually contains predictive adding complexity. 3 Train Random Forest, compare errors across genes, and test stability across repeated splits / bootstraps. 4 Analyze sustained vs. transient response patterns and identify representative genes that the models predict well or poorly. 5 Run the GSE52778 external stress test in airway smooth muscle cells; make final figures and document limitations. 6 Clean the GitHub repo, write the paper, prepare the gene- response explorer, and present results for feedback. WHAT

COUNTS AS A GOOD RESULT A useful outcome does not require the Random Forest to “win.” If a simple early-time rule or Elastic Net performs just as well, that is a stronger interpretability result. If the external airway dataset does not match the A549 response, we will report that directly and use it to show where the A549 response does not carry over across cell types. REFERENCES 1. McDowell IC, Manandhar D, Vockley CM, et al. Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS Computational Biology. 2018;14(1):e1005896. doi:10.1371/journal.pcbi.1005896. Data: GEO GSE104714. 2. NCBI Gene Expression Omnibus. GSE104714: time-series RNA-seq of human A549 cells exposed to dexamethasone or vehicle control. 3. Himes BE, Jiang X, Wagner P, et al. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. PLoS ONE. 2014;9(6):e99625. doi:10.1371/journal.pone.0099625. Data: GEO GSE52778. Kenyon Shi + Kevin Shi | GR-Trace