Nice paper title

# Introduction

DNA content is practically the same across all cells in an organism, but different cell types are capable of executing different functions. This is due to extensive regulation of the several processes involved in the production of a functional protein from its encoding gene.

* Biology of gene regulation
  + Transcription process
  + Transcription Factors
  + DNA methylation
  + Histone modifications / chromatin state

Epigenetic modifications, such as DNA methylation and histone modifications, also impact DNA structure and thus change how accessible it is to the transcriptional machinery.

* Methods for learning gene regulatory networks
  + Correlation as intro
  + Limitations of correlation
  + Linear regression vs mutual information

Analysis of gene-gene correlation is limited in multiple ways. Firstly, the expression of a gene is usually regulated by a variety of factors (e.g. different transcription factors)1, but gene-gene correlation is limited to independent analysis of gene pairs. Secondly, correlation is a symmetric measure of association that cannot be used for quantitative predictions. Thirdly, many indirect effects (i.e. correlations that exist between two genes due to the correlation of both to a third gene) are included in this way, which can lead to many false positive associations. The use of linear regression approaches can overcome these limitations by modelling quantitatively the impact of multiple genes in the expression of another (target) gene. By incorporating the L1 penalty in a regularized regression model, the number of predictive genes for a target can be minimized. Additionally, computation of partial correlations allows for the capture of indirect effects that can then be removed from the network.

In this work, we show that gene correlation is largely invariant across tissues. This motivates the learning of a global Gene Regulatory Network (GRN) based on regularized regression, expected to capture regulatory relationships between genes that are valid across different tissues and cell types. We show that this is indeed the case.

* Choice of data
* Loss of gene regulation with age

# Results

## Gene-gene correlation is largely invariant across tissues and cell types

* Why we expect gene-gene correlations to be the same
* How we tested this in the GTEx and Tabula Sapiens data
* Why those specific groups of genes
* Also with tissue / cell-type specific genes?

<PAIRWISE CORRELATION OF HAND-PICKED GROUPS OF GENES>

## Gene regulatory networks capture invariant gene-gene relationships

Gene-gene correlation analysis is limited to one gene pair at a time, includes many indirect effects and offers no predictive information. In order to overcome these issues, we used regularized regression (Lasso) combined with stability selection to identify stable predictors for each gene. These were then combined into a quantitative linear model that explains expression patterns of a given gene based on the expression patterns of their predictors (see Methods).

<PERFORMANCE OF NETWORK ON DIFFERENT GTEx TISSUES>

<PERFORMANCE OF NETWORK ON PSEUDOBULK FROM TABULA SAPIENS>

<PERFORMANCE OF NETWORK ON TISSUE/CELL-TYPE-SPECIFIC GENES>

Polygon

Description automatically generated

Figure : Distribution of centrality measures in the full human network. Residual edges were removed (see Methods).

* In-degree distribution is bimodal with indirect effects and becomes a power-law when they are removed. Why? We expect biological networks to be scale-free (i.e. the degree distribution can be modelled with a power law distribution), so this would suggest the network without indirect effects is more meaningful.
* Genes with high out-degree also have high in-degree (hubs).
* Transcription factors are not enriched in high out-degrees (not even after removing indirect effects), suggesting we actually don’t capture direct relationships between regulator and target, but rather a connection between the targets.

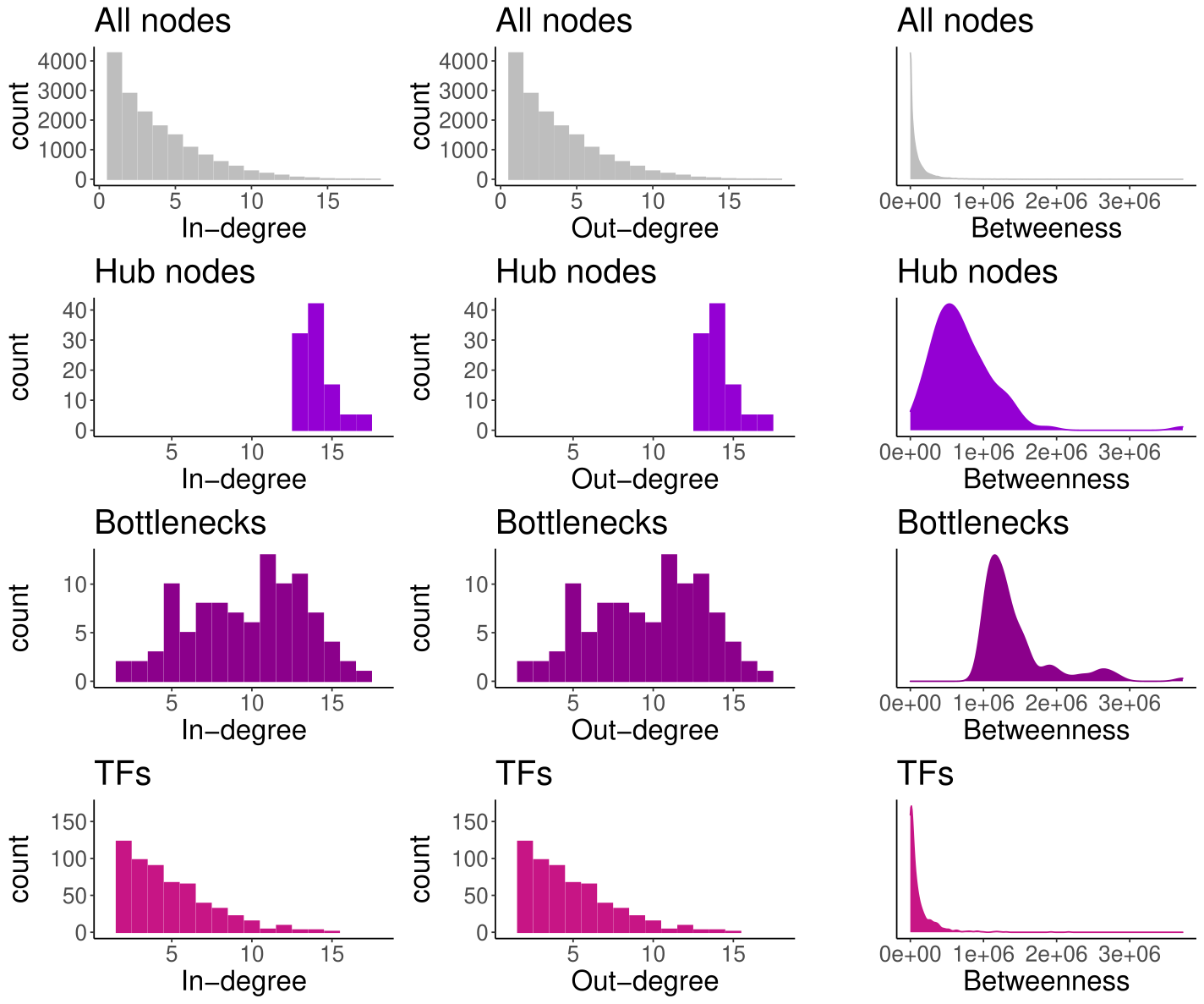


Figure : Distribution of centrality measures in the human network after removal of indirect effects. Residual edges were removed (see Methods).

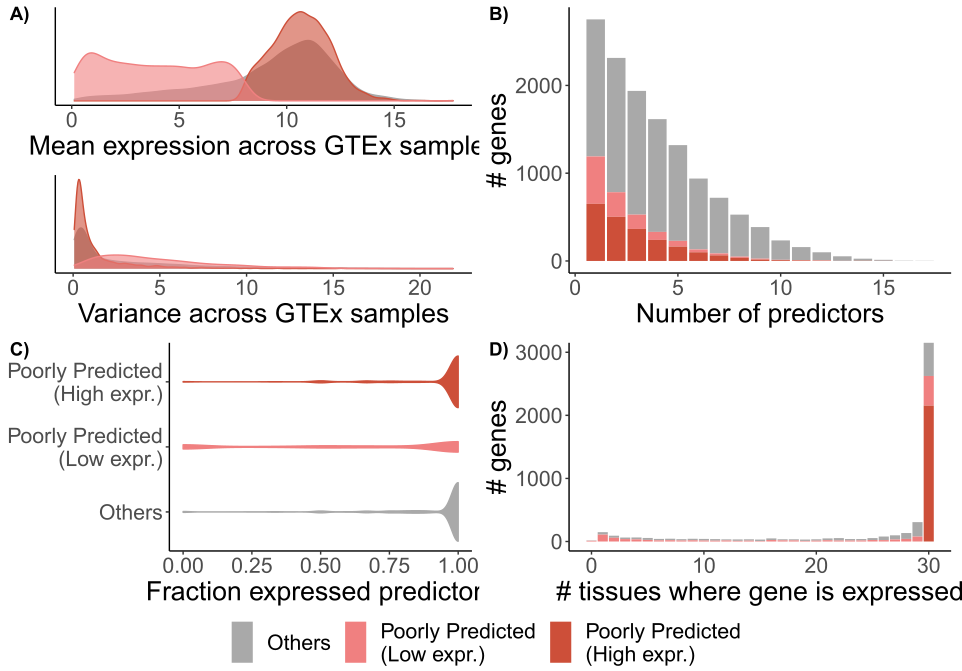


Figure 3: Characterization of genes poorly predicted by the regulatory network after removal of indirect effects. Poorly predicted genes were separated into highly (dark orange) and lowly (pink) expressed. A) Expression mean and variance of poorly predicted genes across GTEx samples. B) Distribution of number of predictors (in-degree) of poorly predicted genes. Bars are stacked. C) Fraction of predictors for each target gene that are expressed in the GTEx data, at mean > 2. D) Number of GTEx tissues where each target gene is expressed (mean > 2).

* Genes that are not well captured by the network are split into two groups: lowly expressed and highly expressed.
* Lowly expressed, poorly predicted genes, are themselves poorly expressed and have predictors that are also poorly expressed.
* Highly expressed, poorly predicted genes have extremely low variance and are ubiquitously expressed across tissues (housekeeping).

## Gene modules

We performed spectral clustering <REF> for gene module detection. Briefly, the network’s adjacency matrix was decomposed into its eigenvectors. Only the eigenvectors with largest absolute eigenvalues were considered for h

* Spectral clustering for gene module detection
* Modules mildly represent different biological functions (GO terms)
* Big module encompasses many things
  + Activity across cell types
  + Which GO slim terms does it include?
  + When further split, what is it?

<ENRICHMENT FOR TF TARGET GROUPS WITHIN EACH MODULE?>

<LINK BETWEEN ABSOLUTE EXPRESSION AND PREDICTABILITY>

## Deviations with age

<GENES WITH DEVIATIONS TO PREDICTION>

<OVERLAP BETWEEN TISSUES>

<COMMON TARGETS OF ANY TFS?>

<COMPARISON WITH EXPRESSION CHANGES>

<GENOMIC POSITION OF GENES WITH DEVIATION TO PREDICTION>

<GROUPING OF GENES OF SAME TAD>

<OVERLAP WITH MARKERS OF CLOSED CHROMATIN IN YOUNG AND OPEN IN OLD?>

<WHAT GENES ARE GAINING REGULATION AND HOW CAN THIS BE EXPLAINED?>

# Discussion

It is necessary to point out that not all mechanisms of gene expression regulation are reflected in transcript levels. For instance, post-translation modifications of transcription factors that activate their regulatory function are independent of the expression level of those transcription factors and thus cannot be captured in such models. Given the limited availability of several molecular levels of high-throughput data, the current work is limited to modelling relationships between transcript levels only, at the risk of underestimating the complexity of regulatory interactions.

* Non-linear effects on gene regulation (e.g. TFs competing for the same binding site)
* Temporal delay from regulation to product - doesn’t matter because we are in steady state, right?

# Methods

## Learning gene regulatory networks

## Removal of indirect effects

## Additional processing of the network

Firstly, residual edges were removed. This was done by computing, for each edge pair (i,j) (j,i), the ratio of between the lowest and the sum of the absolute edge weights. When this ratio was below 0.1, the weakest edge (lowest absolute value) was removed from the network. The largest connected component of the resulting network was then determined using the function *components()* from the *igraph* R package version 1.2.6, with default parameters. All subsequent analyses were performed on the largest connected component.

## Topology analysis

After removing residual weights and reducing the network to its largest connected component, in-degree, out-degree and node betweenness were computed

## Module detection

Module detection was performed using spectral clustering. We chose to remove directionality from our network in order to avoid complex eigenvalues and summarize the weights of edge pairs between two nodes (edge i->j and edge j->i) as the sum of both weights. This captures the full strength of the connection between two nodes. The resulting adjacency matrix was row-wise normalized, by dividing each row by the sum of all its entries (it’s degree). This avoids that highly connected nodes dominate the represented topology <RONJA’S AND KONSTANTINA’S PAPER>. The row-normalized adjacency matrix was then decomposed into its eigenvectors and eigenvalues.

## GTEx data preprocessing

## *Tabula Sapiens* data preprocessing

* Downloaded from , in the form of
* Separated by tissue and sequencing method

Only cell types with at least 1000 cells (10X protocol) or 100 cells (Smart-seq2 protocol) were included in the analysis. Cells were filtered for quality based on their library size and percentage of mitochondrial reads. In order to be included in the analysis, cells had to pass the following filters, defined for each cell type in each tissue separately: 1) library size larger than 3 median absolute deviations (MAD) below the median, in order to exclude dead cells and empty droplets; 2) library size smaller than 3 MAD above the median, in order to exclude doublets; 3) percentage of reads from the mitochondrial genome smaller than 4 MAD above the median, in order to exclude dead cells.

Genes undetected across all cells of the same cell type were removed.

# References

1. Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M. & Aravind, L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.* **360**, 213–227 (2006).

# Supplementary Material

<FILTERS FOR TABULA SAPIENS DATA, EXAMPLE IN ONE CELL TYPE>