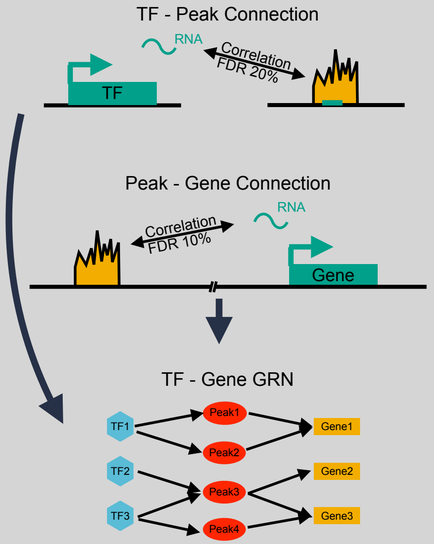
Reconstruction and evaluation of cell type specific gene regulatory networks using chromatin accessibility and RNAseq data

Aryan Kamal, Christian Arnold, Sophia Mueller-Dott, Neha Daga, Annique Claringbould, Maksim Kholmatov and Judith Zaugg*
European Molecular Biology Laboratory, Heidelberg, Germany

Background

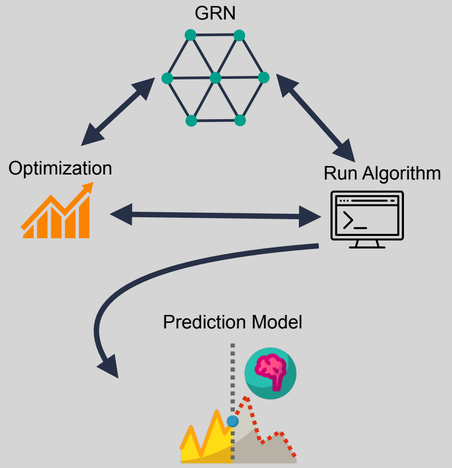
Since most of the genetic variations fall in non-coding regulatory regions of the genome, and the effect of environmental influences are infeasible to measure comprehensively, it is highly important to use predictive regulatory models which help interpret the effect of genetic and epigenetic variation on cellular response to external signals. To address this, we propose a framework for building a unified, unbiased regulatory network using cell-type and condition-specific edges subjected to a universal benchmarking strategy.

GRN Reconstruction



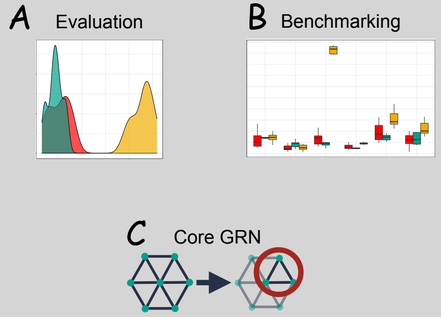
Phase1: Overview of GRN reconstruction using ATACseq and RNAseq data

Prediction Model



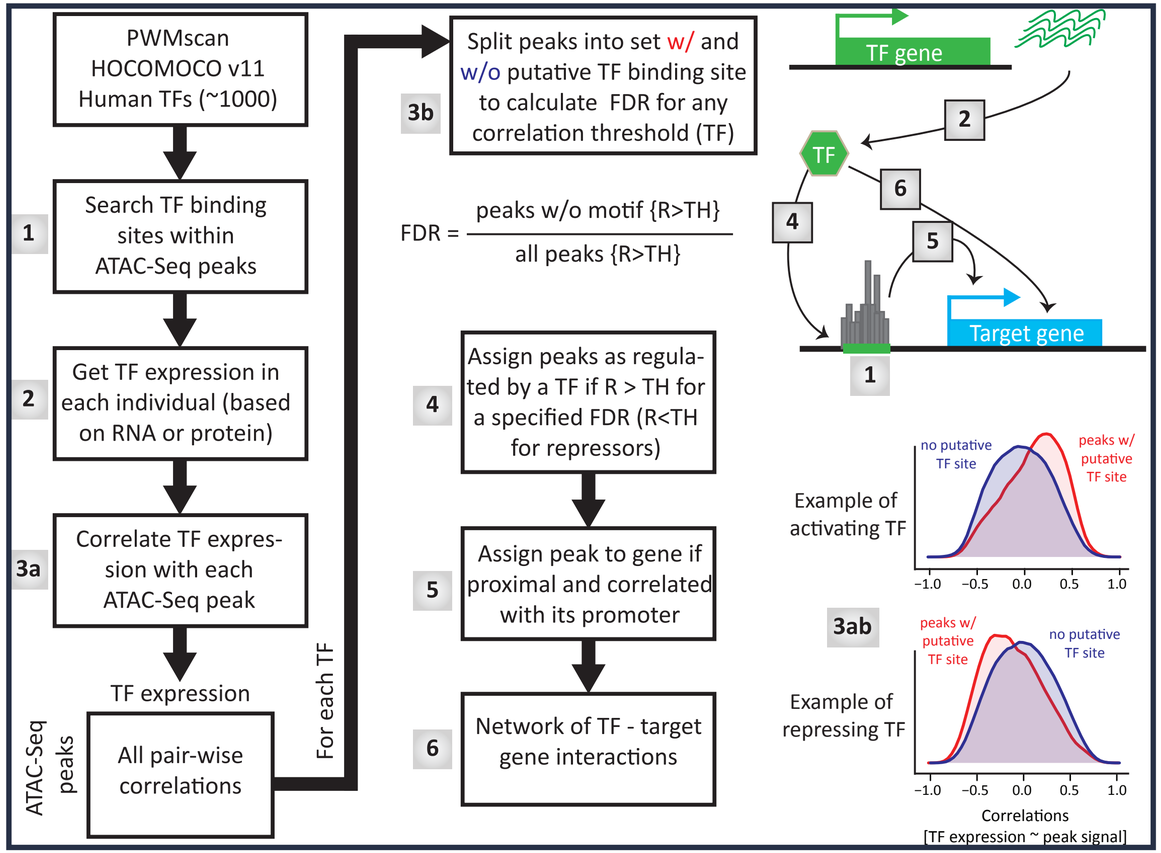
Phase2: Overview of GRN Prediction Model. In this phase, we constructed a model based on GRN to predict differential expression values between two conditions.

Application



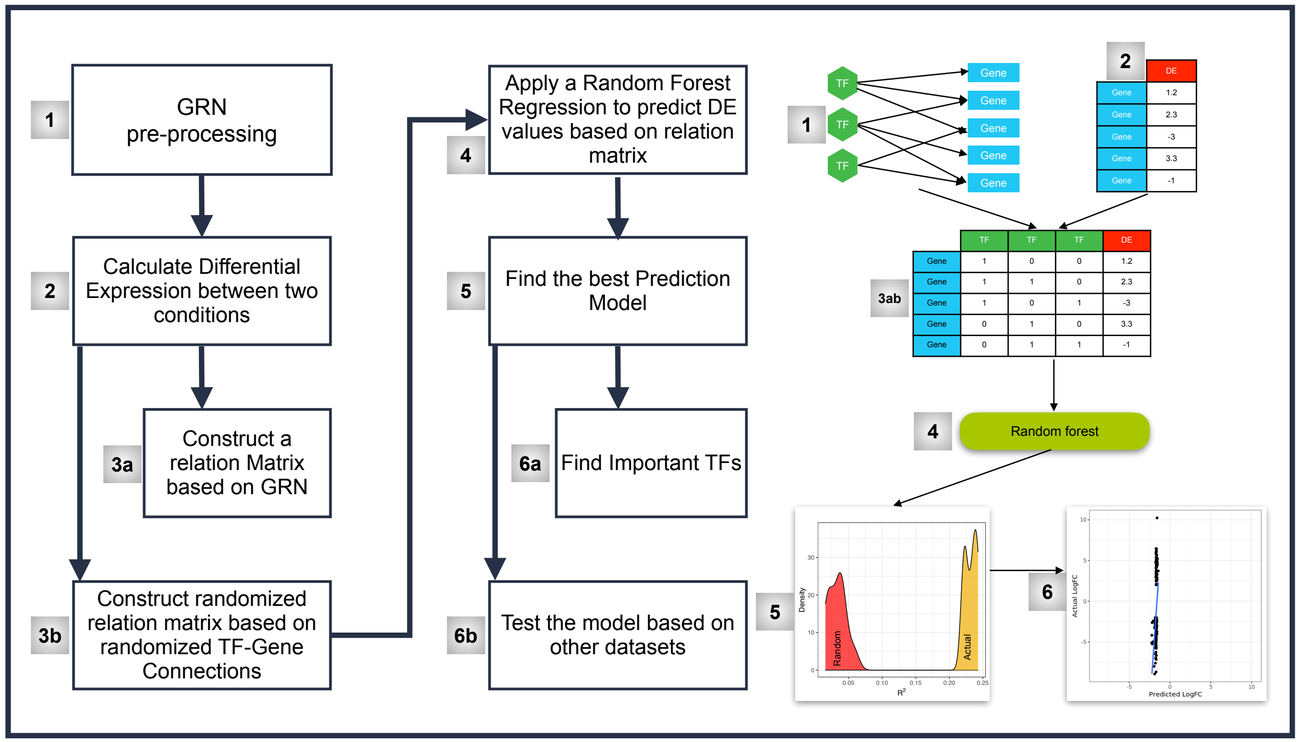
Phase3: (A) evaluation of GRN using the prediction model and compare it with edge-specificity control. (B) Benchmarking and comparing different GRNs based on the prediction model. (C) Pruning GRN using important TFs and TF-Gene connections which results in a smaller but more significant GRN.

EMBL



We construct an unbiased unified regulatory network, based on a collection of cell-type specific models, which will enable us to study general and cell-type specific regulatory mechanisms.
The network has three types of nodes: TFs, regulatory elements, and genes. These are connected by two types of edges: trans-edges (TF to regulatory element) and cis-edges (regulatory element to target gene).

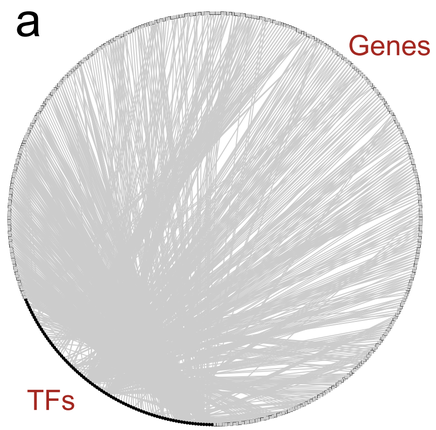
The key point of our approach is that we integrate cis- and trans-edges inferred from multiple complementary methods and thus create an unbiased unified regulatory network.



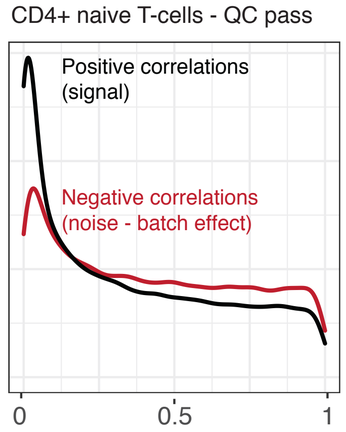
Here we proposed a novel benchmark based on assessing the added values of regulatory networks for predicting cell-type specific expression response to perturbations.
Briefly, we train a (RF) model to learn the coefficients for each TF, in predicting DEG for all genes, based on the TFs-gene links in each specific network.
We are able to train this model for each network (including randomized versions) separately, and compare their performances.
This will provide a network-specific score of predictive value for a specific differential expression response.

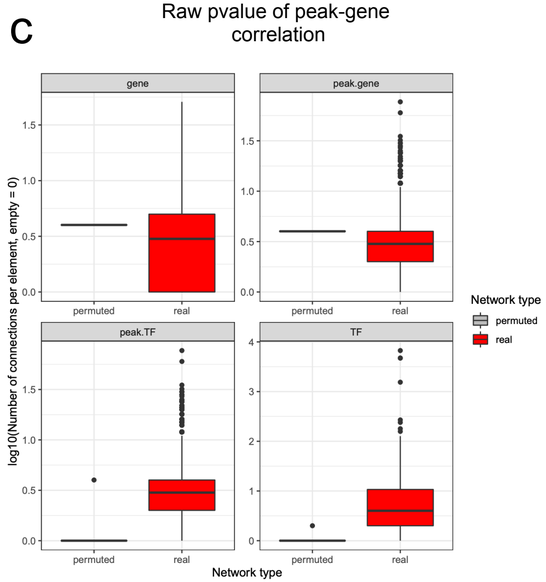
EMBL

Result

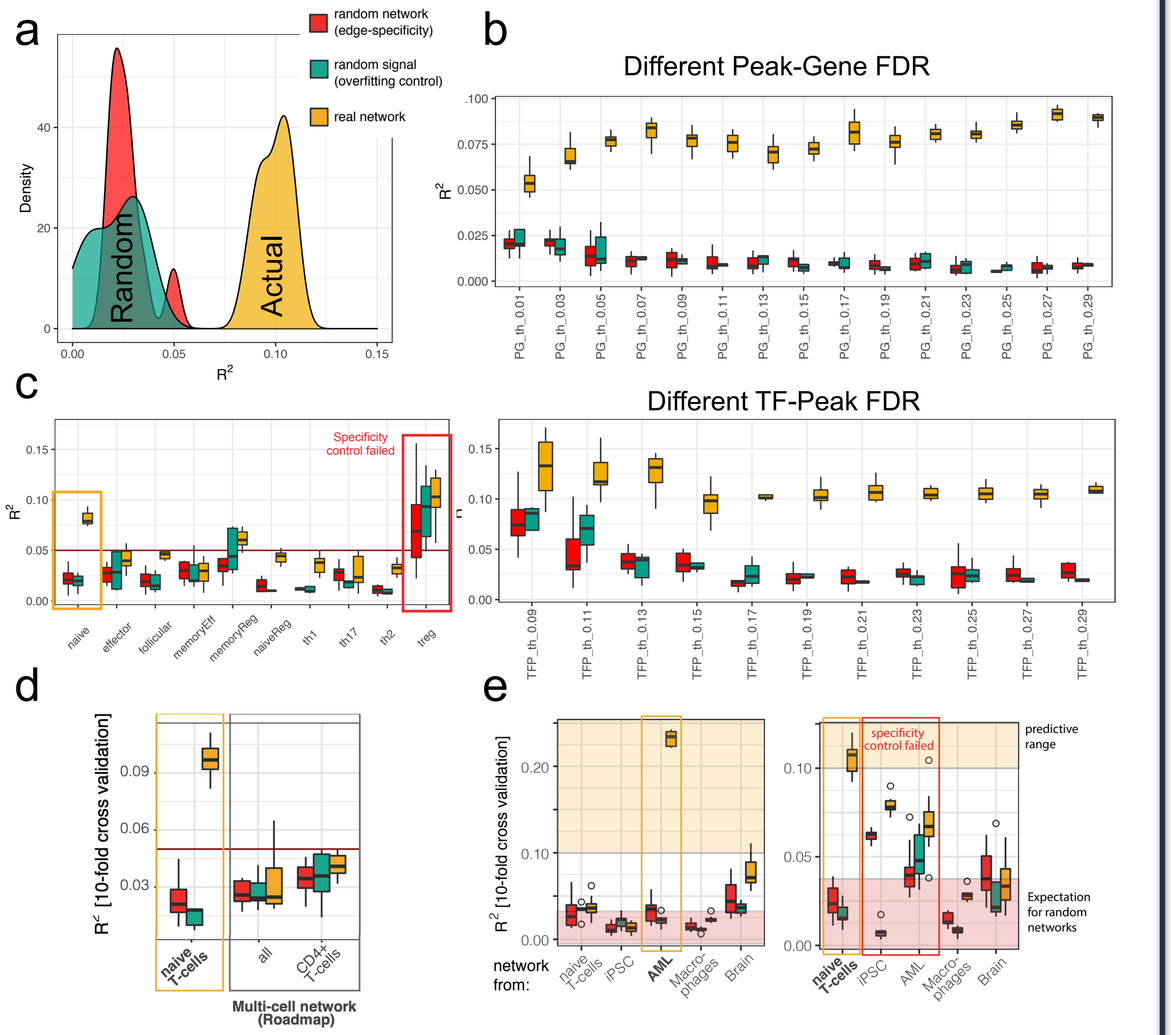


TF-Peak FDR : 0.2
Peak-Gene FDR: 0.1
No of TFs - 174
No of peaks - 3589
No of genes - 3362
Total edges - 12862



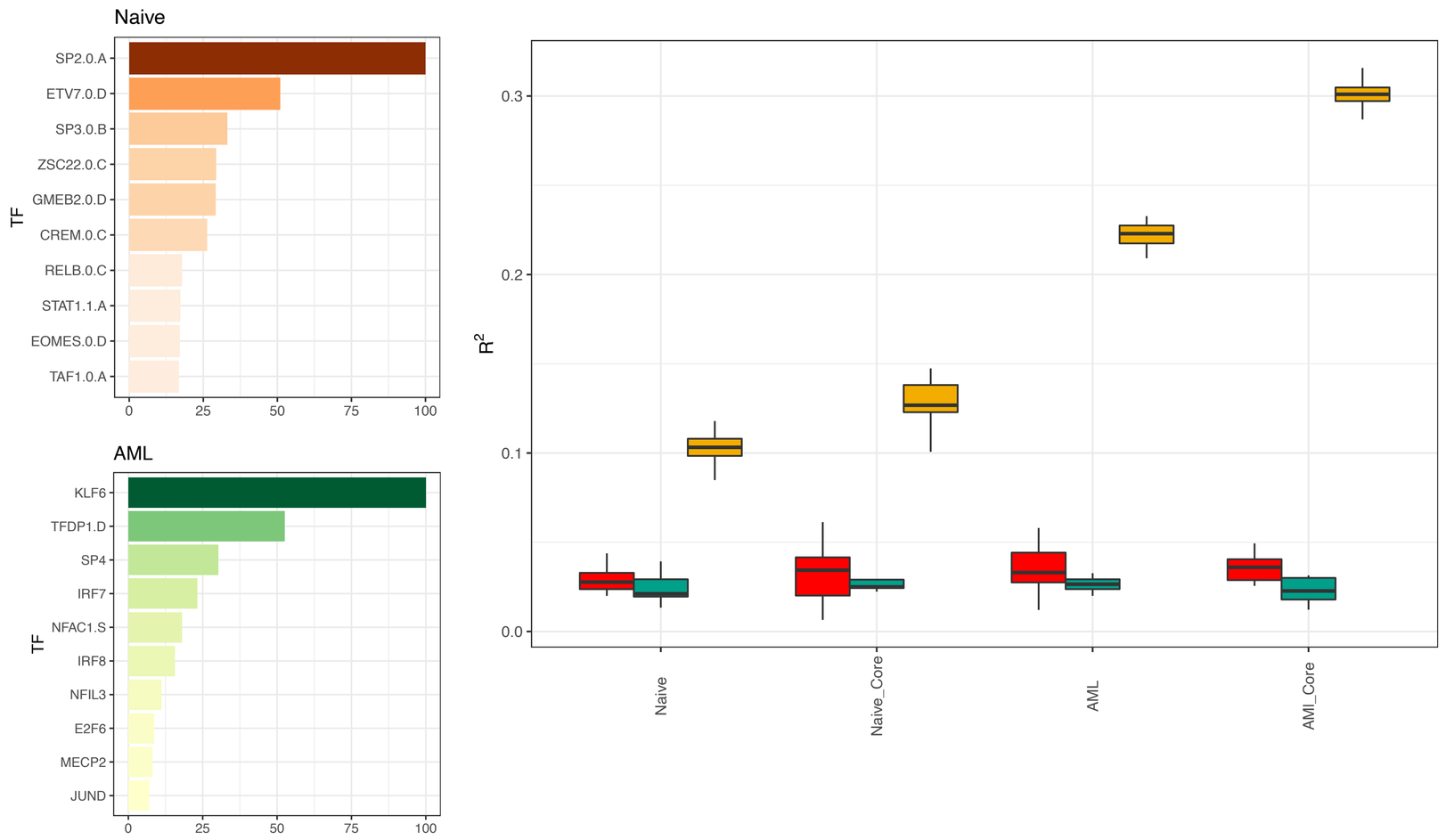


(a) A gene regulatory network constructed using 132 samples from CD4+ Naive T-cells. (b) Distribution of p-values from correlations between peaks and genes (within 250kb) CD4+ networks. If more signal in negative than in positive correlations, QC failed. (c) Number of TF-peak-gene links in real vs permuted GRN.



(a) R squared for predicting differential expression between resting and LPS stimulated naive CD4+ T-cells using real Naive network vs random network. **(b)** Optimizing TF-peak and Peak-Gene FDRs based on the prediction model. **(c)** Specificity of GRN; Naive network cannot predict DE for other CD4+ subcelltypes properly. **(d)** Comparison between different types of networks. **(e)** R-squared (y-axis) is shown for predicting differential expression between subtypes of AML (left) and resting vs LPS stimulated naive CD4+ T-cells (right) based on models trained on different cell-type specific networks (x-axis). In both cases, only the cell-type specific network is predictive. Controls include overfitting: learned on permuted differential expression values (green), and edge-specificity: learned on network with randomized edges (red).

EMBL



(a, b) TF Importance: Different Variable Importance Measures are available for Random Forest. We used debias impurity-based variable importance measure to extract important TFs. **(a)** Important TFs sorted by their scores in Naive CD4+ prediction model. **(b)** Important TFs in AML prediction models.

(c) Core GRN; GRN constructed only for small subset of important TFs, using only high quality connections. The comparison between Core and default network shows an improvement in AML and Naive dataset.

Conclusion

- Evaluation Model shows real TF-Genes connection has been captured by our GRN.
- Our GRNs are cell type specific, as evidenced by the fact that they cannot learn differential expression in for other cell types.
- For each specific cell type GRN, small number of TFs are more important that others which can be further investigated.

EMBL