Cell Systems, Volume *15*

Supplemental information

Transcriptome data are insufficient to control false discoveries in regulatory network inference

Eric Kernfeld, Rebecca Keener, Patrick Cahan, and Alexis Battle

Supplemental Information for “Transcriptome data is insufficient to control false discoveries in regulatory network inference”

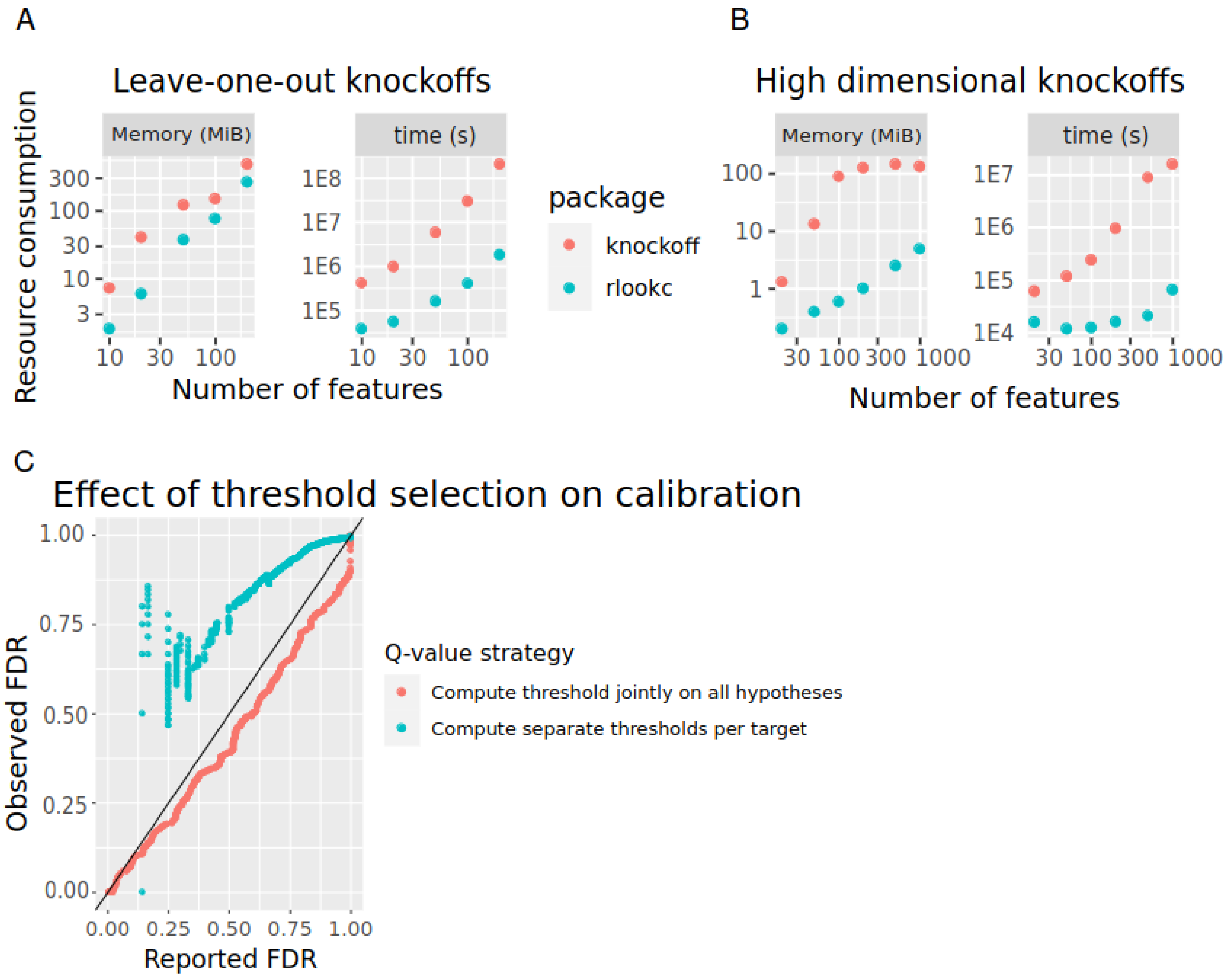

Figure S1. Knockoff construction for transcriptome-scale data. Related to Figure 1.
A) Runtime and memory consumption for leave-one-out Gaussian knockoff construction with n=1000 observations using rlookc (our method) and the reference implementation in the R package “knockoff”, with the method indicated by the color.

B) Runtime and memory consumption for high-dimensional Gaussian knockoff construction with n=10 observations using rlookc (our method) and the reference implementation in the R package “knockoff”, with the method indicated by the color.
C) Threshold selection. Reported FDR from knockoff filter (x-axis) versus observed FDR (y-axis) in a simulated variable selection problem with 805 observations, 334 features, and 334 response variables. One trend line shows FDR when separate thresholds are selected for each target and the other shows FDR when a single threshold is used across all targets, with the threshold selection method indicated by color.

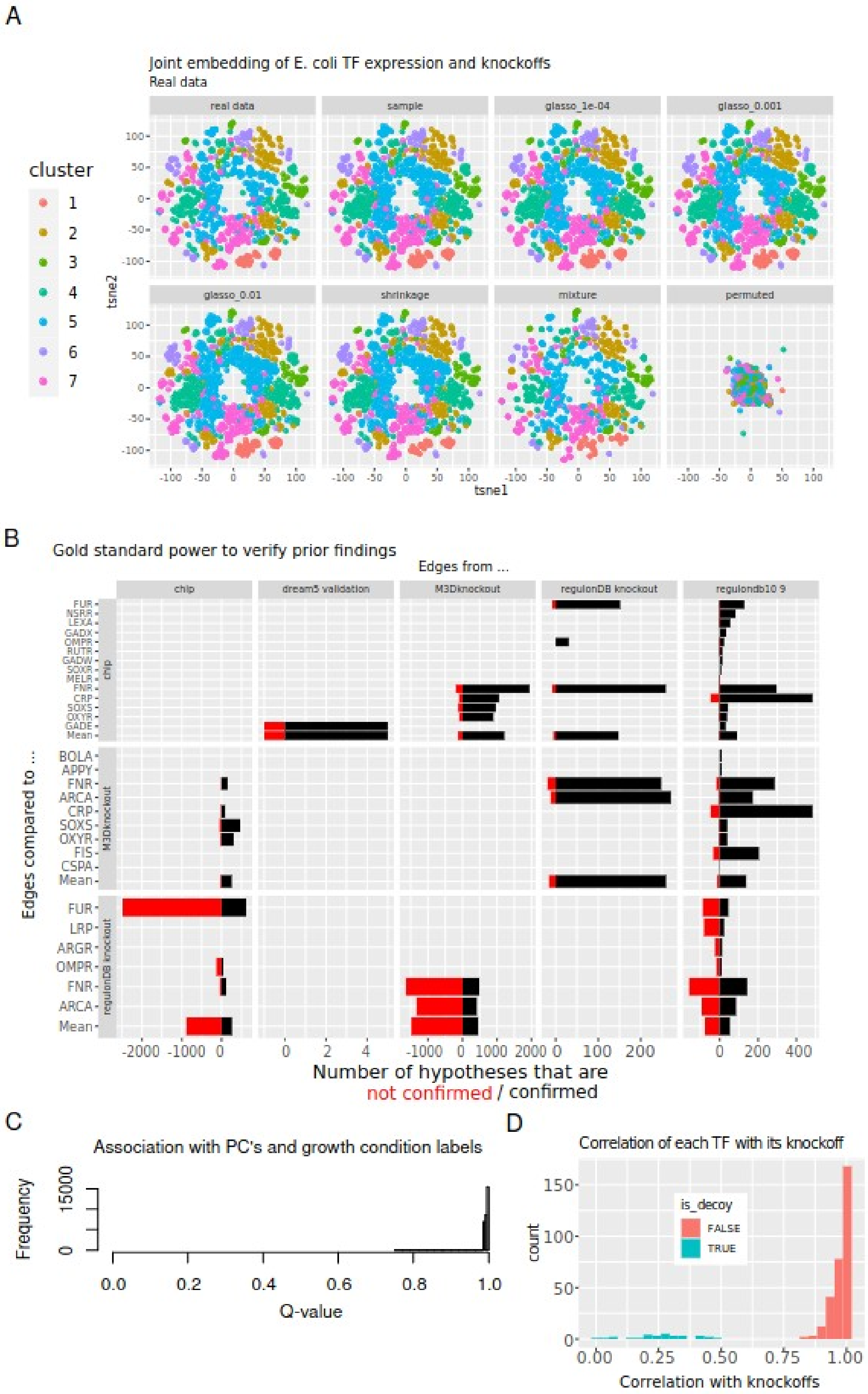
A
Joint embedding of E. coli TF expression and knockoffs
Real data
cluster
1
2
3
4
5
6
7
real data
sample
glasso_1e-04
glasso_0.001
glasso_0.01
shrinkage
mixture
permuted
tsne2
100
50
0
-50
-100
tsne1
-100 -50 0 50 100
B
Gold standard power to verify prior findings
Edges from ...
chip
dream5 validation
M3Dknockout
regulonDB knockout
regulondb10 9
Edges compared to ...
chip
FUR
NSRR
LEXA
GADX
OMPR
RUTR
GADW
SOXR
MELR
FNR
CRP
SOXS
OXYR
GADE
Mean
M3Dknockout
BOLA
APPY
FNR
ARCA
CRP
SOXS
OXYR
FIS
CSPA
Mean
regulonDB knockout
FUR
LRP
ARGR
OMPR
FNR
ARCA
Mean
-2000 -1000 0
0 2 4
-1000 0 1000 2000
0 100 200
0 200 400
Number of hypotheses that are
not confirmed / confirmed
C
Association with PC's and growth condition labels
Frequency
0
15000
0.0 0.2 0.4 0.6 0.8 1.0
Q-value
D
Correlation of each TF with its knockoff
count
0
50
100
150
is_decoy
FALSE
TRUE
0.00 0.25 0.50 0.75 1.00
Correlation with knockoffs

Figure S2. Technical characteristics of knockoff construction on the *E. coli* TF expression dataset. Related to Figure 3.

A) T-SNE embeddings of *E. coli* TF expression and corresponding knockoff features. Samples are colored based on k-means cluster assignments, which were trained on the TF features. Figure is based on n=805 microarray profiles.

B) Comparison of gold standards. Hypotheses are extracted from the gold standard labeled in the top margin and checked to see if they are supported by the gold standard labeled in the left-hand margin. Hypotheses are omitted if they cannot be checked by the gold standard on the left, for instance if it is based on ChIP or knockout data and the regulator was never ChIPped or knocked out. Color indicates confirmed versus unconfirmed hypotheses.

C) Q-values for testing of conditional independence between each TF and each of the principal components or perturbation indicators that was explicitly conditioned on in Fig. 5b. The knockoff generation method is “glasso_1E-04”. Figure is based on n=805 microarray profiles.

D) Correlation distribution between variables and their knockoff copies. Gaussian knockoffs were generated for the 805 by 332 matrix of *E. coli* TF expression using the sample covariance matrix. Color indicates whether the variable is a real gene or a decoy.