Identifying Gene Regulatory Network Rewiring using Latent Differential Graphical Models

Dechao Tian, Quanquan Gu, and Jian Ma

Supplementary Text

Data used in the Application of LDGM to TCGA Breast Cancer Datasets

Genes involved in differential network reconstruction were extracted from estrogen signaling pathway. Estrogen signaling pathway (hsa04915) was downloaded from KEGG [1] with 212 interactions connecting 82 genes/TFs. Here interactions and genes/TFs were extracted from the pathway by KEGGgraph [2]. Totally 420 gene expression samples from Luminal A and 141 samples from Basal-like were retrieved using the cBioPortal [3, 4]. Expression levels of 82 genes from Luminal-A subtype and 81 genes from Basal-like subtype are significantly not normally distributed (Shapiro-Wilk test [5], Benjamini-Hochberg-corrected FDR $<$0.05). Sample latent correlation matrices are computed as the inputs for the models.

Principle of majority to infer the source of a differential edge

Here we use a heuristic approach to determine the source given a differential edge. In other words, if we have identified a differential edge, we would like to decide which group (either Luminal A subtype or Basal-like subtype in our TCGA data application) is more likely to derive this edge. Recall that $\mathbf{\Delta} = \mathbf{\Theta}^L - \mathbf{\Theta}^B$, where $\mathbf{\Theta}^L$ and $\mathbf{\Theta}^B$ are precision matrices and represent individual networks from Luminal A subtype and Basal-like subtype, respectively. $\mathbf{\Delta}$ represents the differential network between the two subtypes. Let $\widehat{\mathbf{\Delta}}$ be an estimator of $\mathbf{\Delta}$ by LDGM. Let $\widehat{\mathbf{\Theta}}_k^L = (\widehat{\Theta}_{ij,k}^L)$ and $\widehat{\mathbf{\Theta}}_k^B = (\widehat{\Theta}_{ij,k}^B)$ be estimators of $\mathbf{\Theta}^L$ and $\mathbf{\Theta}^B$ by Glasso with a tuning parameter $\lambda = \lambda_k$. Here $\lambda_k, 1 \leq k \leq 30$, is selected such that individual networks $\widehat{\mathbf{\Theta}}_k^L$ and $\widehat{\mathbf{\Theta}}_k^B$ gradually grow from empty networks ($k = 1$) to complete networks ($k = 30$). Assume that $i - j$ is an estimated differential interaction by LDGM, i.e., $\widehat{\Delta}_{ij} \neq 0$. Then a principle of majority based on Glasso is applied to infer which subtype the differential interaction $i - j$ only exists or has a much stronger regulatory relationship. More specifically, $i - j$ is from Luminal A subtype if $\sum_{k=1}^{30} \mathbb{1}\left(|\widehat{\Theta}_{ij,k}^L| > |\widehat{\Theta}_{ij,k}^B|\right) > \sum_{k=1}^{30} \mathbb{1}\left(|\widehat{\Theta}_{ij,k}^L| < |\widehat{\Theta}_{ij,k}^B|\right)$ and from Basal-like subtype otherwise.

Note on generating regularization parameter λ

Throughout this paper, we always compared LDGM with other models on a sequence of λ to test if advantages or meaningful biological discoveries of LDGM are consistently observed along the sequence of λ. To generate the sequence of λ, we first selected λ_{max} which is the minimum value of 10^k, $|k| = 1, 2, \cdots$, such that estimated differential network is an empty network, i.e., there is no interaction in the network. Then we selected λ_{min} which is the maximum value of $\lambda_{\text{max}}/2^k$, $k = 1, 2, \cdots$, such that the estimated differential network is a complete network, i.e., every pair of genes are connected in the network. Then we generated a sequence of 30 λ from λ_{min} to λ_{max} by equal increment.

References

[1] Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* **28**(1), 27–30.

[2] Zhang, J. D. and Wiemann, S. (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics,* **25**(11), 1470–1471.

[3] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery,* **2**(5), 401–404.

[4] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling,* **6**(269), pl1–pl1.

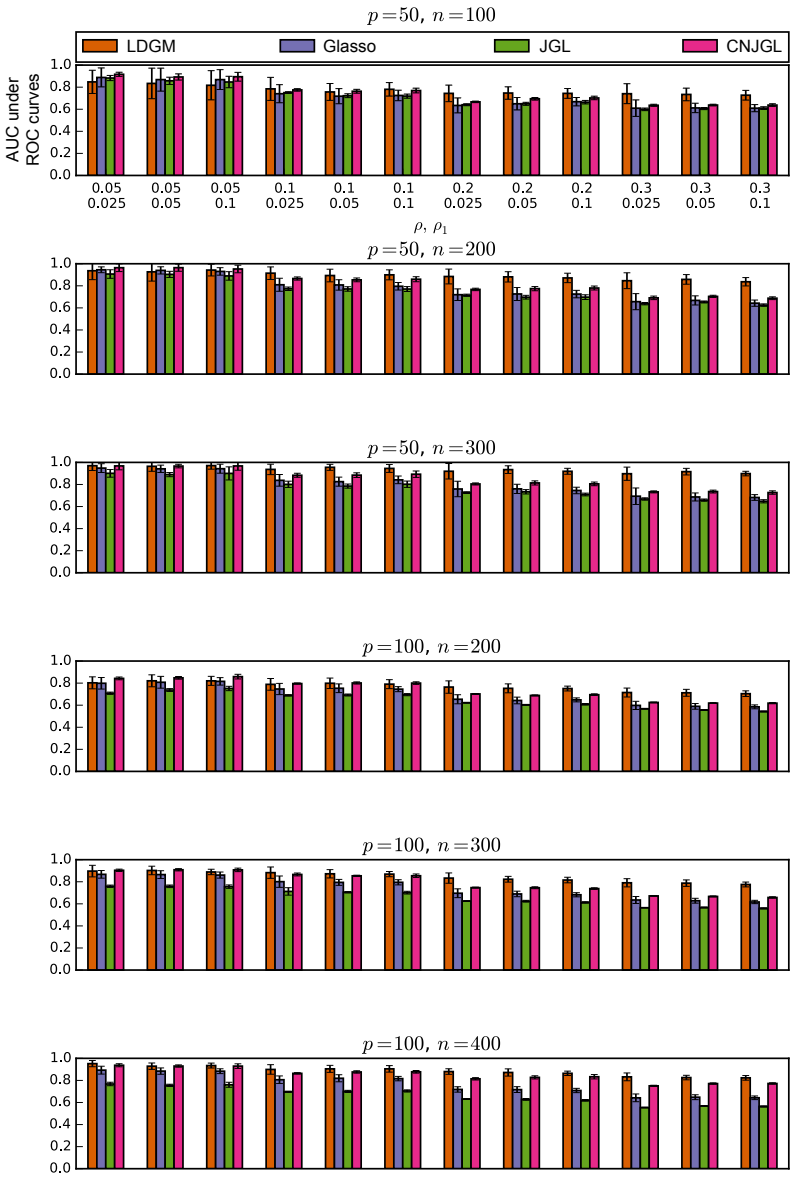
[5] Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika,* **52**(3/4), 591–611.

Supplementary Tables

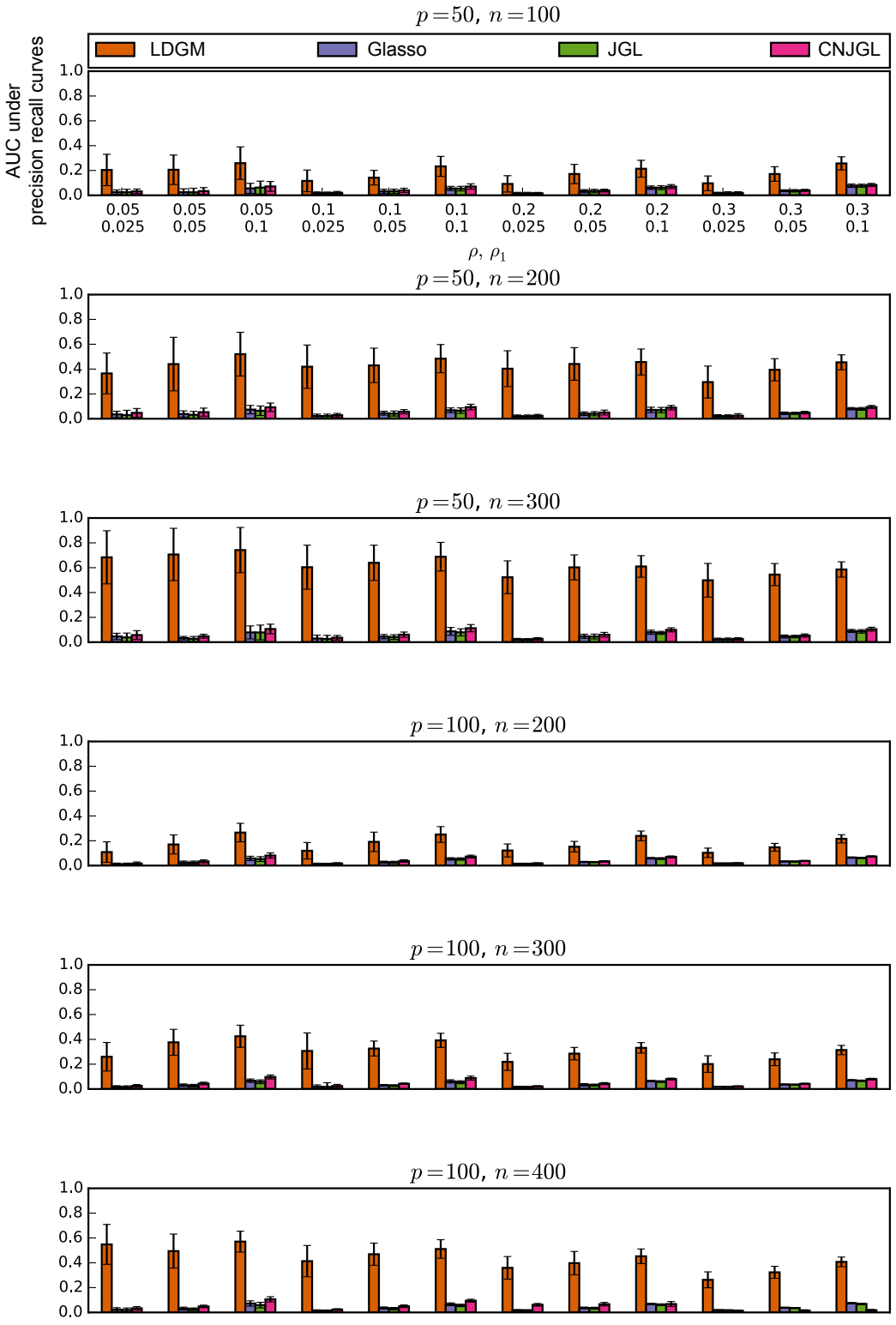
Supplementary Table 1: List of significantly enriched BioCarta pathways (FDR <0.05). Luminal A dataset contains 25 genes that have a majority ($>50\%$) differential interactions from Luminal A subtype in the reconstructed differential network shown in Fig. 5C. Basal-like dataset represents 31 genes that have a majority ($>50\%$) differential interactions from Basal-like subtype. Pathway analysis is performed using DAVID.

Subtype	Term	Genes	P value	FDR
Luminal A	Trka receptor signaling pathway	AKT1, SOS1, PIK3CA, SHC1, PIK3R1	5.99e-07	6.58e-04
	Multiple antiapoptotic pathways from IGF-1R signaling lead to BAD phosphorylation	AKT1, SOS1, PIK3CA, SHC1, PIK3R1	5.40e-06	5.94e-03
	PTEN dependent cell cycle arrest and apoptosis	AKT1, SOS1, PIK3CA, SHC1, PIK3R1	1.27e-05	1.40e-02
Basal-like	Cadmium induces DNA synthesis and proliferation in macrophages	MAPK1, HRAS, MAP2K1, JUN, PLCB1	1.29e-06	1.44e-03
	Roles of β-arrestin-dependent recruitment of Src Kinases in GPCR signaling	MAPK1, HRAS, ADCY1, MAP2K1, PLCB1	5.40e-06	6.05e-03
	Aspirin blocks signaling pathway involved in platelet activation	MAPK1, HRAS, MAP2K1, GNAI1, PLCB1	6.82e-06	7.63e-03
	Signaling pathway from G-Protein families	HRAS, ADCY1, MAP2K1, GNAI1, JUN	1.04e-05	1.17e-02
	Angiotensin II mediated activation of JNK pathway via Pyk2 dependent signaling	EGFR, MAPK1, HRAS, MAP2K1, JUN	2.99e-05	3.35e-02

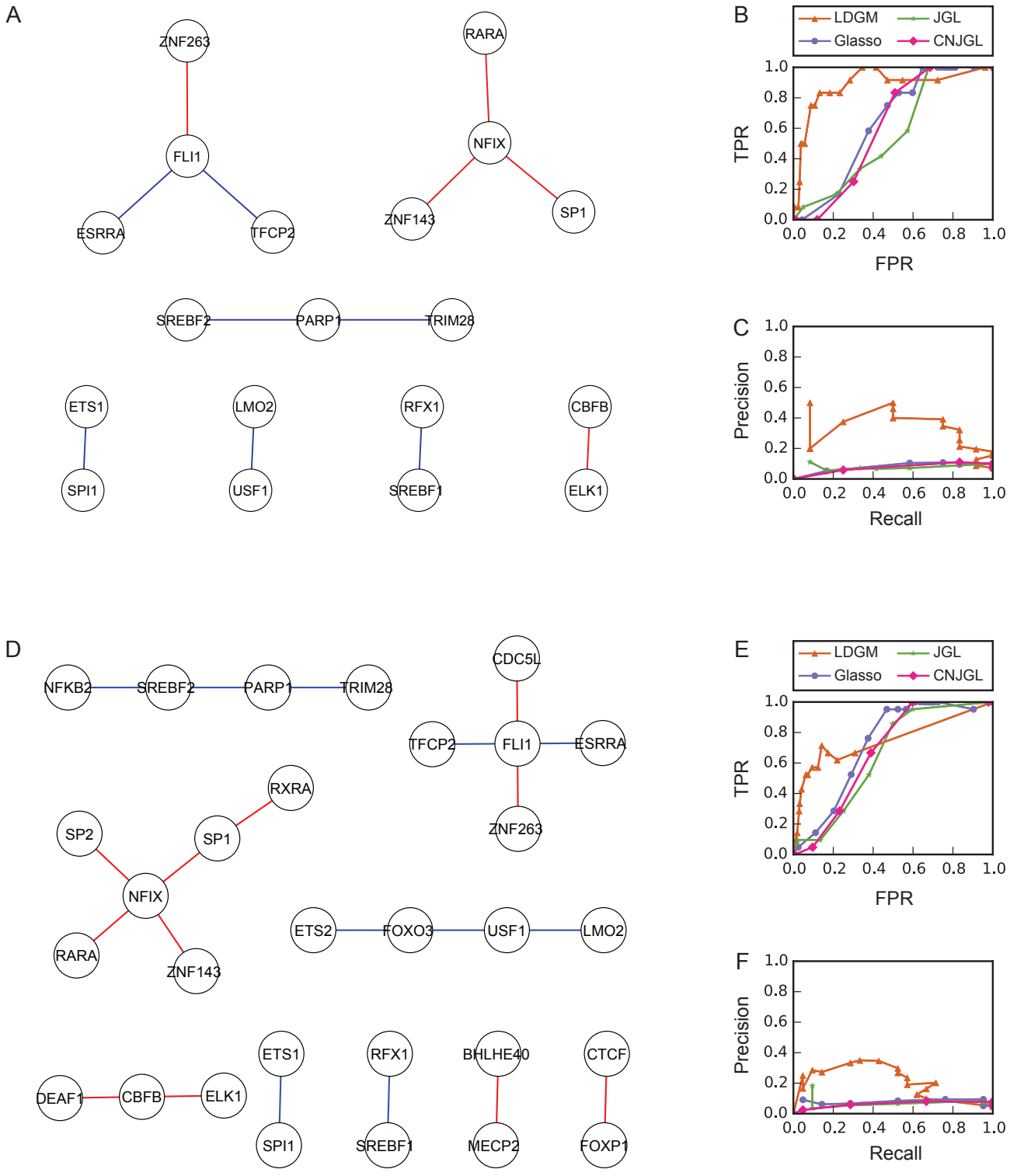
Supplementary Figures



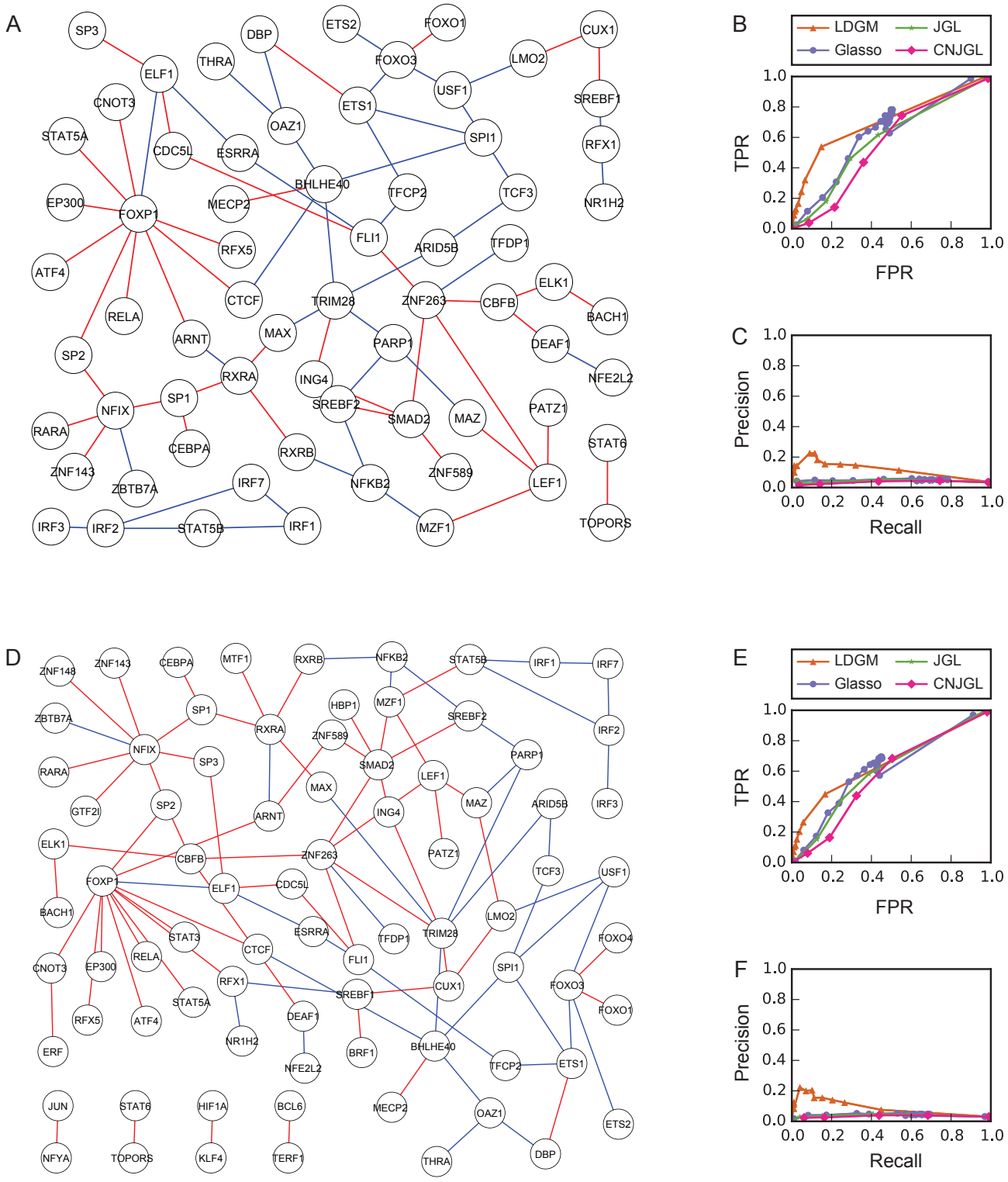
Supplementary Figure 1: AUC under ROC curves of different methods on simulated data under different p, n, ρ and ρ_1. The advantage of LDGM becomes more visible when differential networks are more dense with an increased ρ_1 regardless of p and n. Here ρ stands for the individual network density and ρ_1 is the proportion of network-specific edges. Bar height represents the AUC under an averaged ROC curve over 30 runs. Error bar represents one standard deviation of AUC under 30 replicated ROC curves.



Supplementary Figure 2: AUC under precision-recall curves of graphical models on simulated data based on different p, n, ρ and ρ_1. LDGM consistently has a much larger AUC than other models. Here ρ stands for the individual network density and ρ_1 is the proportion of network-specific edges. Bar height represents the AUC under an averaged precision-recall curve over 30 runs. Error bar represents one standard deviation of AUC under 30 replicated precision-recall curves.



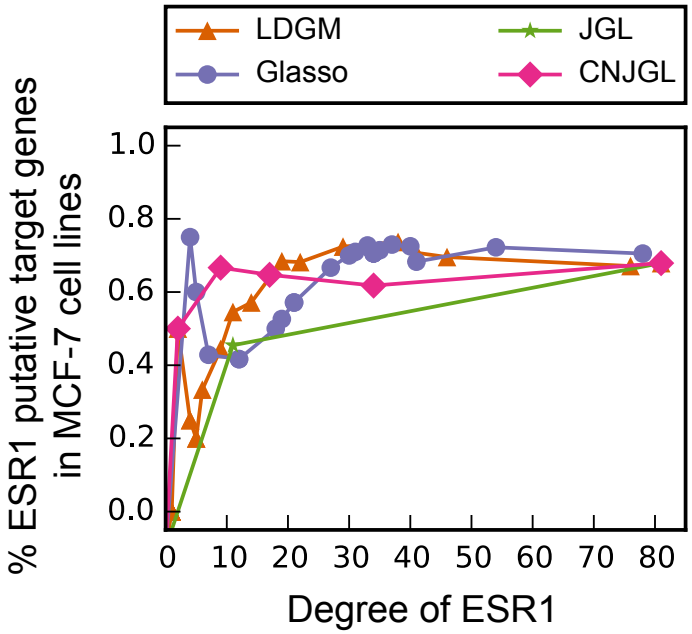
Supplementary Figure 3: Performance of different methods on the GTEx data (brain and whole blood data). **(A)** The benchmark network with 19 TFs and 12 tissue-specific interactions by setting $r_b = 0.95$ and $r_w = 0.8$. Red edges are interactions specific to brain while blue edges are specific to whole blood. **(B)** ROC curves and **(C)** Precision-recall curves to recover the benchmark network. **(D)** The benchmark network with 30 TFs and 21 tissue-specific interactions by setting $r_b = 0.9$ and $r_w = 0.75$. **(E)** ROC curves and **(F)** Precision-recall curves to recover the benchmark network.



Supplementary Figure 4: Performance of different methods on the GTEx data (brain and whole blood data). **(A)** The benchmark network with 68 TFs and 78 tissue-specific interactions by setting $r_b = 0.7$ and $r_w = 0.6$. Red edges are interactions specific to brain, while blue edges are specific to whole blood. **(B)** ROC curves and **(C)** Precision-recall curves to recover the benchmark network. **(D)** The benchmark network with 82 TFs and 98 tissue-specific interactions by setting $r_b = 0.6$ and $r_w = 0.6$. **(E)** ROC curves and **(F)** Precision-recall curves to recover the benchmark network.

Supplementary Figure 5: QQ-plot for normality test of gene expression data. Expression data are from 82 genes in estrogen signaling pathway in Luminal A subtype. Blue data points should be close to the red diagonal line if the expression of the gene is approximately normally distributed.

Supplementary Figure 6: QQ-plot for normality test of gene expression data. Expression data are from 82 genes in estrogen signaling pathway in Basal-like subtype.



Supplementary Figure 7: Validation of the estimated differential networks by different models. Totally 54 ChIP-seq experiments on ESR1 from MCF-7 cell line were downloaded from CistromeDB. A putative target gene of ESR1 in MCF-7 cell lines is defined as a gene where there is at least one ESR1 ChIP-seq peak within 5 kbp of the gene in at least 10 out of 54 ChIP-seq experiments.