1 Supplementary Methods

1.1 K-fold cross validation

The first step of the SmCCNet algorithm (Figure 1, Step I) is to determine the best penalty parameter pair through K-fold cross validation (K-fold CV, Figure S1). First, the dataset on n subjects is split into K folds, i.e. K pairs of test and training sets (Step I.1), and a selection of paired values for the two penalty parameters is considered. Then, for each training set (Train k) and each possible penalty pair (Penalty j), we compute the pseudo weights through SmCCA and subsampling, and obtain the prediction error (Step I.2). (The pseudo canonical weights are the average absolute values of $u = (w_1, w_2)$ from each sub-sampling. Since the sign of u is ambiguous, we use the absolute canonical weights $|u|$ instead of u. Both u and $-u$ satisfy Eq 3. For each penalty pair j, the prediction errors are computed for all K folds. The sum of these errors, Δ_j (Eq S1), is reported as the total prediction error for the penalty pair j (Step I.3). The pair corresponding to the smallest total prediction error is the best penalty choice (Step I.4).
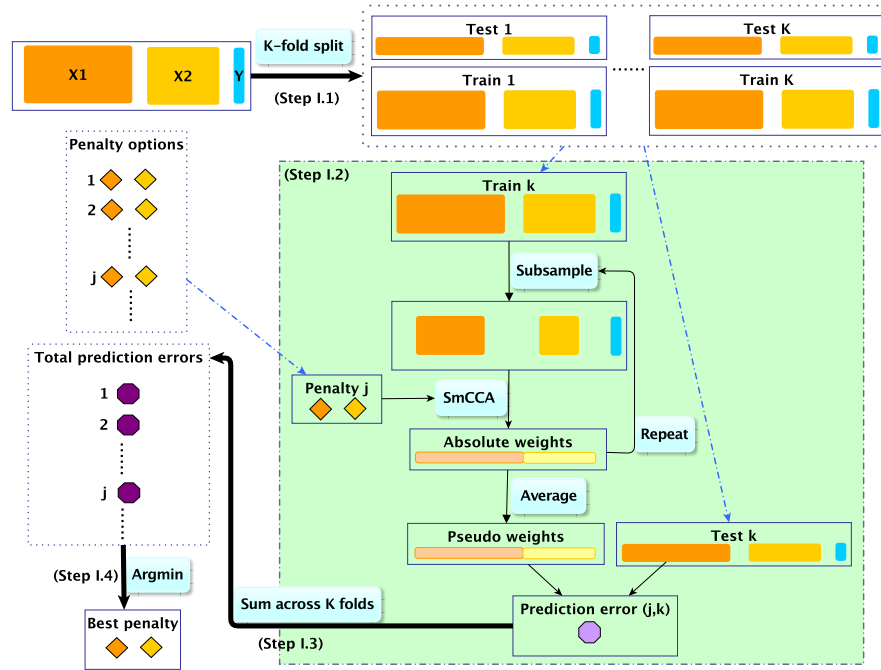

Fig. S1. Selection of penalties. Step I.1: Split subjects into K folds of test and training sets. Step I.2: Consider a list of paired penalty options. For each training data and penalty option, estimate pseudo weights through feature subsampling and SmCCA. Compute prediction error with corresponding test set. Repeat Step I.2 for all combinations of penalties and each of the K folds. Step I.3: Compute the total prediction error for each penalty option by summing across K folds. Step I.4: Define the best penalty pair to be the one that corresponds to the least total prediction error.

For the kth fold, denote test sets and training sets as $X_{test_k,1}, X_{test_k,2}$ and $X_{train_k,1}, X_{train_k,2}$, respectively. Given a penalty pair j, estimate the pseudo weights $(\nu_{train_k,1,j}, \nu_{train_k,2,j})$ based on the training data, and compute the prediction error for penalty pair j as

$$\Delta_j = \sum_{k=1}^{K} \left| \left| \nu_{train_k,1,j}^T X_{test_k,1}^T X_{test_k,2} \nu_{train_k,2,j} \right| - \left| \nu_{train_k,1,j}^T X_{train_k,1}^T X_{train_k,2} \nu_{train_k,2,j} \right| \right| \tag{S1}$$

The best penalty pair is given by $\arg\min_j \Delta_j$. The K-fold CV includes subsampling features with the same proportions as Figure 1, Step II. Suppose the number of features for X_1, X_2 are p_1, p_2, and the subsampling proportions are q_1, q_2, respectively. Then the range for penalty parameter c_s is $[1, \sqrt{p_s \times q_s}]$, for $s = 1, 2$.

1.2 Mediation framework and key parameters

Since miRNAs can bind mRNAs and promote transcript degradation or repress translation, the mediation framework assumes miRNAs $\longrightarrow$ mRNA $\longrightarrow$ Pheno (Eq 4). The key simulation parameters and settings are

1. number of functional groups, G, assuming one miRNA per group.
2. number of mRNA affected by each relevant miRNA (first arrow in Eq 4), $m_1, \cdots, m_G$.
3. number of mRNA affecting Pheno within each functional group (second arrow in Eq 4), $l_1, \cdots, l_G$. Note that $l_g \leq m_g, \forall g \in \{1, \cdots, G\}$.
4. mRNA-miRNA dependency structure. Denote the expression of miRNA g as $miRNA_g$, and the expression of gene j in group g as $mRNA_{gj}$. Then

$$mRNA_{gj} \sim N(\beta_{gj0} + \beta_{gj1} miRNA_g, \ \sigma_1^2), \tag{S2}$$

and we assume that miRNAs negatively regulate associated mRNAs, $\beta_{gj1} < 0, \ \forall g, \ \forall j \in \{1, \cdots, m_g\}$.

5. Pheno-mRNA dependency structure. Let each functional group effect on Pheno be $Pheno_g$, and denote the total effect of all groups as $Pheno$.

$$Pheno_g = \sum_{j=1}^{l_g} w_{gj} mRNA_{gj}, \ \ Pheno \sim N\left(\alpha_0 + \sum_{g=1}^{G} \alpha_g Pheno_g, \ \sigma_2^2\right). \tag{S3}$$

6. Additional Q mRNA-miRNA groups that are not associated with Pheno:

$$\text{miRNAs} \longrightarrow \text{mRNA} \nrightarrow \text{Pheno}. \tag{S4}$$

For each of the Q groups, one miRNA affects r genes. The dependency structure is implemented as in Eq S2.

1.3 Simulation background datasets

The background noise was generated based on the two real multi-omics datasets by independently permuting subjects and feature order within each omics dataset. The first multi-omics dataset came from a large recombinant inbred mouse panel that had been bred from reciprocal crosses between the Inbred Long Sleep (ILS) and Inbred Short Sleep (ISS) strains, called the ILSXISS (LXS) panel. The original long-sleep and short-sleep lines were selectively bred for a long or short duration of loss of righting reflex due to ethanol (Markel *et al.*, 1995; Williams *et al.*, 2004). A total of 175 mice (57 LXS strains with 3 replicates and 2 strains with 2 replicates) were sacrificed and had total RNA extracted from whole brain tissue. The extraction and quantitation details for the miRNA sequencing data and the mRNA microarray data can be found in (Russell *et al.*, 2018; Rudra *et al.*, 2017; Vestal *et al.*, 2018) and the data is available at NCBI/GEO https://www.ncbi.nlm.nih.gov/geo/ with identifier GSE125953. The miRNA expression estimates were derived from miRNA sequencing data using a variance stabilizing transformation (VST) in the DESeq2 R package (Anders and Huber, 2012). We then selected moderately to highly heritable protein-coding genes and miRNA features (heritability score ≥ 0.3) (Rudra *et al.*, 2017) and took the average transformed expression levels within each strain. Each strain was considered as a subject. The filtered LXS dataset contained 57 subjects, 3212 genes and 341 miRNAs. The second multi-omics dataset came from The Cancer Genome Atlas (TCGA) breast invasive carcinoma project. The high-throughput sequencing based miRNA and mRNA expression levels for 1064 subjects were downloaded using TCGAbiolinks (Colaprico *et al.*, 2015). The mRNAs were also collected at the gene level. We focused on white female patients with tumor free status and protein coding genes. After filtering out lowly expressed and low variant (little variance among subjects) features, we had 358 subjects, 2124 genes and 163 miRNAs. The sequencing data were also transformed via VST.

1.4 Simulation parameters and evaluation

The simulation datasets were generated with both LXS and TCGA background data, according to the parameter choices in Table S1. For each simulation run, mRNA/miRNA that were in the simulated networks were the 'positive' cases, and all other miRNA/mRNA were considered 'negative' cases. After the methods were run on the simulated data, we examined the similarity matrix for the positive nodes, and took the maximal connection to other nodes (except for the diagonal). If that maximal value in the similarity matrix was above a threshold (e.g., 0.1), then this was considered a true positive, otherwise it was a false negative. For the negative nodes, if the value in the similarity matrix was above a threshold (e.g., 0.1), then this was considered a false positive, otherwise it was a true negative. For a series of thresholds, the false/true positive/negatives were used to compute the true positive rates (tpr) and false positive rates (fpr), plot the ROC curve tpr$\sim$fpr, and compute the area under the curve (AUC). The median and range of AUCs are reported in Table 1 across the simulations, and for each method. The larger the values, the better the performance of the method. AUCs were compared between methods using a paired Wilcoxon test.

1.5 Competing methods: WGCNA and SsCCA

To evaluate SmCCNet, we compared its performance to two competing methods: weighted gene coexpression network analysis (WGCNA) and sparse supervised CCA (SsCCA). WGCNA is one of the most popular method for gene coexpression network discovery. It has been used to combine miRNA and mRNA data, either by creating mRNA modules using WGCNA and then relating each module to a miRNA of interest (Nunez *et al.*, 2013; Noh *et al.*, 2014), by constructing separate modules for miRNAs and mRNAs, then relating each module to the phenotype of interest (Ponsuksili *et al.*, 2013; Mamdani *et al.*, 2015; Wu *et al.*, 2016). The first approach relies on prior knowledge and the latter approach does not create a miRNA-mRNA network. A natural way to produce miRNA-mRNA networks is applying WGCNA to the mRNA and miRNA data (e.g., Miao *et al.*, 2016; Li *et al.*, 2017) and focusing on the multi-omics modules that are significantly associated with the trait of interest.

SsCCA was proposed along with SmCCA for incorporating phenotypic trait information in Witten and Tibshirani, 2009. The authors compared SsCCA to SmCCA for the case of a binary phenotype, and expressed preference of SsCCA over SmCCA. While SsCCA is computationally easier, it only uses phenotype information to prioritize genomic features. SsCCA only assigns non-zero weights to features that are moderately or highly correlated to the phenotype. It does not utilize the magnitude of strength of the association of feature with a continuous phenotype when selecting omics features. In contrast, our approach simultaneously includes phenotype in both feature selection and canonical weight calculation. It does not require final selected features to have a moderate, or high marginal correlation with the phenotype. For a fair comparison, we also add the subsampling scheme in the implementation of the SsCCA method.

2 Data extraction and quantification

2.1 COPD miRNA and mRNA expression

The original miRNA dataset consisted of 30 individuals. We removed three samples due to either low sequencing depth or missing clinical information. The original data contained 1173 miRNAs. We first removed 3 hemolysis-susceptible miRNAs (Kirschner *et al.*, 2013; Shkurnikov *et al.*, 2016), and then

Parameter	Description	Naïve (I)	Realistic (II)	MoreGroup (III)	Noisy (IV)
$l_g, \forall g$	Number of relevant mRNA per group (out of 5)	5	2	2	2
G	# of relevant miRNA	2	2	5	2
$\alpha_1, \cdots, \alpha_G$	Group contribution to Pheno	α_1=3	α_1=3	α_1=3	α_1=3
		α_2=5	α_2=5	α_2=3	α_2=5
				α_3=3	
				α_4=5	
				α_5=5	
σ_1 LXS	Residual standard deviation of mRNA	0.1	0.1	0.1	1
TCGA		1	1	1	5, 10, 15
$m_g, \forall g$	Number of targeted mRNA per group	5			
$\beta_{gj0}, \forall g$	mRNA baseline	Selected to guarantee positive expression value			
$\beta_{gj1}, \forall g, \forall j$	miRNA contribution to mRNA	-1			
$w_{gj}, \forall g, \forall j$	Weights of relevant mRNAs	$1/l_g$			
α_0	Pheno baseline	10			
σ_2	Pheno residual standard deviation	5			
Q	Number of miRNA-mRNA groups not associated with Pheno	10			
r	Numberof mRNA in each of the Q group	5			

Table S1. Simulation parameter setting. Columns 3-6 show the parameter choices for the four simulation scenarios: Naïve (I), Realistic (II), MoreGroup (III) and Noisy (IV). For TCGA, we explore increasing variances σ_1 for the Noisy (IV) scenario, which are labeled IVa, IVb, IVc in the manuscript. The top half of the table indicates parameters that vary across scenarios. The bottom half shows parameter values that are consistent across scenarios.

filtered out "absent" and low-variant miRNAs by requiring more than 5 reads in at least 13 subjects, as well as a minimum standard deviation of 1 across subjects. The number of miRNAs reduced to 414 post filtering. To normalize the expression data, we applied upper-quartile normalization and Remove Unwanted Variations with Residuals (RUVr, Risso *et al.*, 2014). The generalized linear model used in RUVr to determine residuals includes the following covariates: gender, race, age, smoking status, white blood cell percentages from the complete blood count (CBC, including neutrophils, lymphocytes, eosinophils), and forced expiratory volume during the first second expressed as a percent of predicted value (FEV1pp). Finally, the corrected sequencing counts were transformed to be homoscedastic via VST (Anders and Huber, 2010).

The full CODPGene total mRNA sequencing data contained 537 peripheral blood samples and 57387 genes (Parker *et al.*, 2017). Data processing was done on the full sample set for better statistical power. We removed 10 samples either with missing spirometry or CBC information, or with CBC total exceeding 100%. The remaining samples covered all five stages of the GOLD spirometry grades (controls: 236; GOLD 1: 48; GOLD 2: 126; GOLD 3: 80; GOLD 4: 37). We categorized the samples to three groups by GOLD status: control (GOLD = 0), mild cases (GOLD = 1 or 2), and severe cases (GOLD = 3 or 4). For the genes, we first filtered down to protein coding genes only (18790/54317). Then for each gene, we required more than 20 reads in at least 117 samples (size of the smallest case-control group), and the standard deviation to be greater than 1. Post filtering, we obtained 527 samples and 5001 genes. As with the miRNA data, we applied upper-quartile normalization and RUVr (Risso *et al.*, 2014) with the same generalized linear model. Finally, the corrected sequencing counts were transformed to be homoskedastic via VST (Anders and Huber, 2010) and the same 27 pilot subjects were extracted to match the miRNA subjects.

2.2 TCGA breast cancer miRNA and mRNA expresion

The mRNA expression were collected at the gene level. We applied upper-quartile normalization to adjust for library sizes. Genes with maximum count less than or equal to 5000 were removed. Then, we extracted protein coding genes and further removed two genes that had zero reads for 80% or more across subjects. After filtering, 7973 genes remained. The miRNA expression was first adjusted via upper-quartile normalization. miRNAs with maximum count less than 51 or with at least 80% zeros were removed, and 348 miRNAs remained. Both genes and miRNAs were transformed through VST to address heteroscedasticity.

3 Supplementary Results

3.1 COPD comparison with SsCCA using FEV1pp

We also ran the COPD data set with SsCCA since it was similar in performance to SmCCNet in the simulations. As an objective way to compare performance, we examined the number of predicted and validated miRNA-mRNA target sites as a percentage of all connections in the resulting full modules that represent a negative correlation. We also changed the edge cut-height to evaluate this percentage in trimmed modules. SsCCA generally identified a smaller number of modules and edges. Compared to SsCCA, we found that for the different options (edge cut-height and predicted/validated), SmCCNet consistently identified relatively more miRNA-mRNA target pairs, as a percentage of all negative edges (Figure S2).

3.2 COPD application with percent emphysema

For percent emphysema, the optimal mRNA and miRNA penalty parameters are (2.96, 2.12) (Figure S3a). The results include 9 connected miRNA-mRNA modules before filtering. The number of miRNA and mRNA within each module ranges from 1 to 8 and 1 to 21, respectively. The miRNA/mRNA ratio
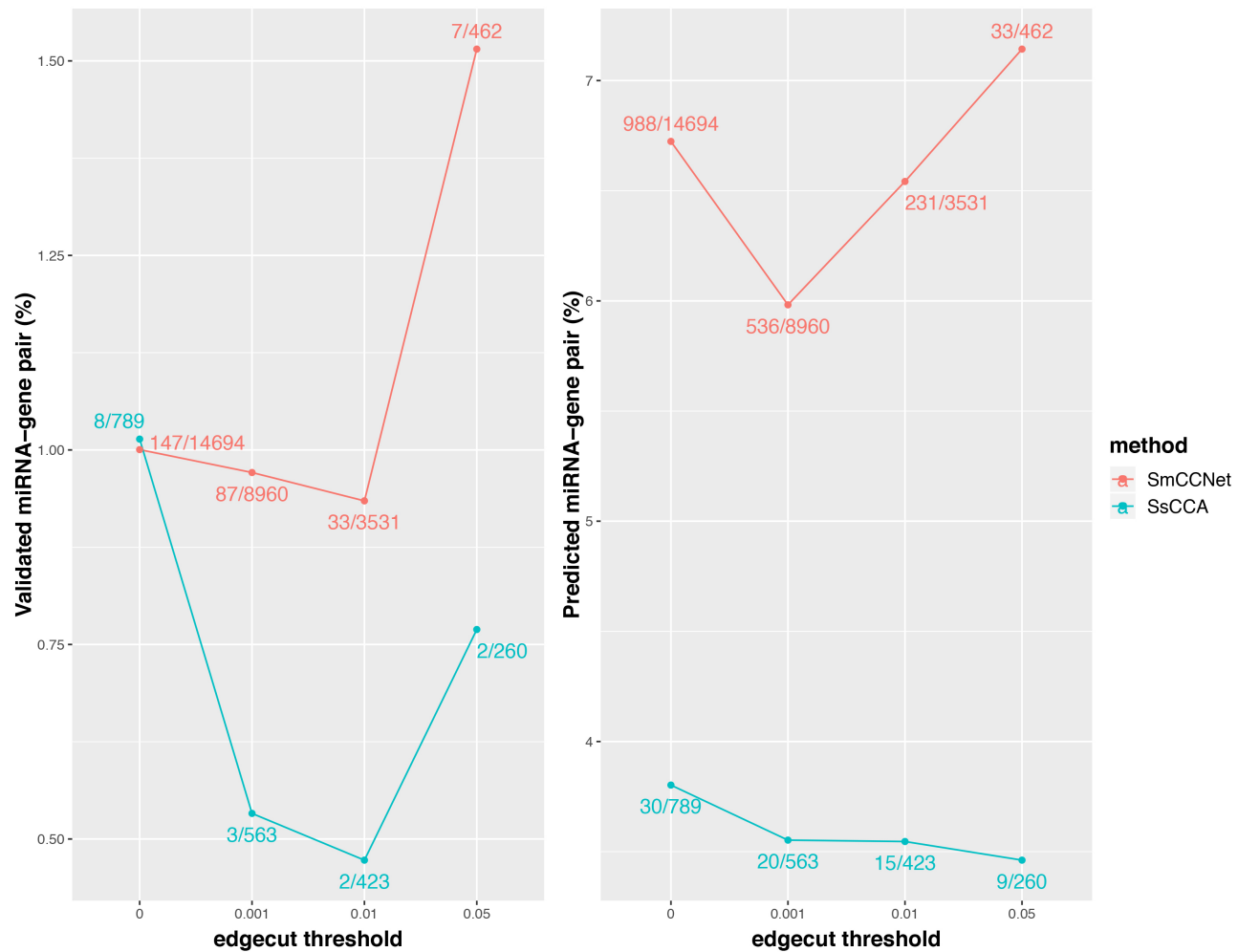
Fig. S2. Predicted and Validated miRNA-mRNA target pairs for different edge cuts. The lines indicate the percentages of predicted or validated target pairs for SmCCNet (red) or ScCCA (blue).

ranges from 0.17 to 1, indicating that there are at least as many mRNA as miRNA. There are 181 negative miRNA-mRNA connections in the 9 full modules. Out of those negative connections, 2 targets have been validated (miR-186-5p and *QRICH1*, miR-186-5p and *AASDHPPT*) and 7 additional targets have been predicted using MultiMir. All modules have very strong module-trait correlations (Figure S3b). Checking the pairwise feature-trait correlations, we notice that none of the modules contain features with weak individual correlations. The layout is rather similar to the FEV1pp pairwise correlation figure after applying an edge cut of 0.1. This suggests a lower level of noise compared to FEV1pp. After applying the 0.1 edge threshold to the percent emphysema modules, we identified the top percent emphysema-related miRNA-mRNA subnetworks: trimmed Modules 2 and trimmed Module 3 (Figure S3c), with module-trait correlation -0.83 (p-value = 2e-07) and 0.76 (p-value = 7e-06). All features in the trimmed modules exhibit moderate or strong pairwise correlation to percent emphysema (Figure S3b). The hubs are miR-6894-3p and miR-1227-3p for trimmed Modules 2 and 3, respectively. Similar to the FEV1pp-related top subnetwork (trimmed Module 2), the hubs are connected to all genes and miRNAs in the trimmed modules. Note that miR-6894-3p also has a strong connection to miR-186-5p, which is known to be up-regulated in the COPD subjects in the Li population (Ding *et al.*, 2017). Another overlap with the trimmed FEV1pp-related top subnetwork is miR-150-3p, which is known to be up-regulated with increasing emphysema severity (Christenson *et al.*, 2013). There is no pathway enrichment for genes identified in the percent emphysema networks.

3.3 TCGA application with uncensored survival time

Figure S4 shows the network results. The list of genes and miRNA names for the trimmed Module 1 is included in the supplement file TrimmedModules.xlsx.

3.4 Run Time

We evaluated the computation time for a fixed set of penalties based on an Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz Ubuntu 18.04.1 LTS operating system. The number of features is larger in COPD compared to the TCGA data, but the sample size is larger in TCGA compared to COPD. WGCNA is notably computationally efficient and not affected by sample and feature size. SmCCNet slows down with more features, while SsCCA slows down with sample size. All methods run on the order of seconds or minutes.
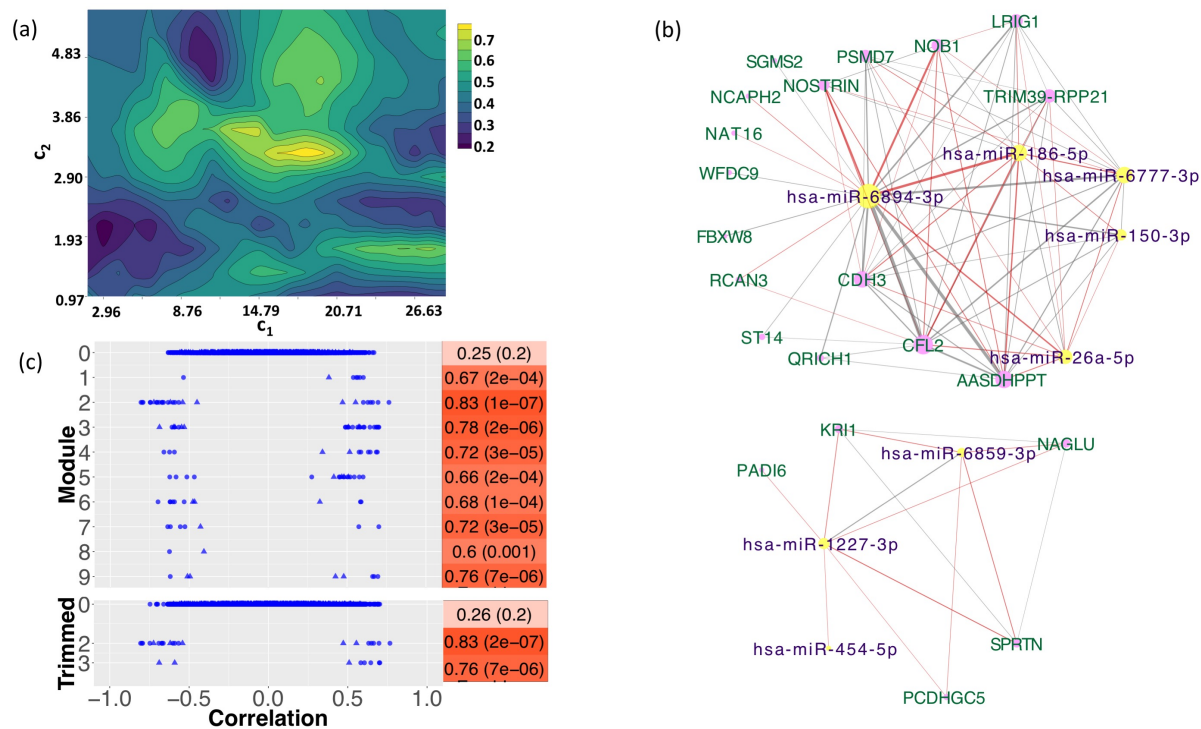
Fig. S3. Percent emphysema results. (a) Total prediction error contour. The x- and y-axes indicate LASSO penalties considered for mRNA and miRNA, respectively. Blue to yellow scale indicates increasing of total prediction error. (b) Correlations to FEV1pp for full (top) and trimmed (bottom) modules. The mRNAs are circles and the miRNAs are triangles. All features excluded in identified modules are grouped as Module 0. Heatmap indicates module-trait correlations (with p-values in parentheses). (c) Trimmed module network. Negative edge are in red; positive edges are in gray. The width of the network edges indicates the connection strength. Yellow nodes with purple labels are miRNAs; magenta nodes with green labels are mRNAs. Node size corresponds to the number of edges connections.

Method	COPD	TCGA
SmCCNet	21.67 minutes	14.24 minutes
SsCCA	8.85 minutes	14.36 minutes
WGCNA	14.11 seconds	13.09 seconds

Table S2. Computation time comparisons. The COPDgene data includes 27 subjects with 5001 mRNAs and 414 miRNAs. The TCGA cancer data includes 358 subjects with 2124 mRNAs and 163 miRNAs.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Anders, S. and Huber, W. (2012). Differential expression of rna-seq data at the gene level–the deseq package.

Christenson, S. A., Brandsma, C.-A., Campbell, J. D., Knight, D. A., Pechkovsky, D. V., Hogg, J. C., Timens, W., Postma, D. S., Lenburg, M., and Spira, A. (2013). mir-638 regulates gene expression networks associated with emphysematous lung destruction. *Genome Medicine*, **5**(12), 114.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., *et al.* (2015). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, **44**(8), e71–e71.

Ding, Y., Tian, Z., Yang, H., Yao, H., He, P., Ouyang, Y., Yao, J., Li, M., and Jin, T. (2017). Microrna expression profiles of whole blood in chronic obstructive pulmonary disease. *International Journal of Clinical and Experimental Pathology*, **10**(4), 4860–4865.

Kirschner, M. B., Edelman, J. J. B., Kao, S. C.-H., Vallely, M. P., Van Zandwijk, N., and Reid, G. (2013). The impact of hemolysis on cell-free microrna biomarkers. *Frontiers in Genetics*, **4**, 94.

Li, S., Li, B., Zheng, Y., Li, M., Shi, L., and Pu, X. (2017). Exploring functions of long noncoding rnas across multiple cancers through co-expression network. *Scientific Reports*, **7**(1), 754.

Mamdani, M., Williamson, V., McMichael, G. O., Blevins, T., Aliev, F., Adkins, A., Hack, L., Bigdeli, T., Van Der Vaart, A. D., Web, B. T., *et al.* (2015). Integrating mrna and mirna weighted gene co-expression networks with eqtls in the nucleus accumbens of subjects with alcohol dependence. *PloS One*, **10**(9), e0137671.
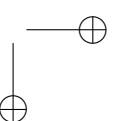
Markel, P. D., DeFries, J. C., and Johnson, T. E. (1995). Use of repeated measures in an analysis of ethanol-induced loss of righting reflex in inbred long-sleep and short-sleep mice. *Alcoholism: Clinical and Experimental Research*, **19**(2), 299–304.

Miao, X., Luo, Q., Zhao, H., and Qin, X. (2016). Ovarian transcriptomic study reveals the differential regulation of mirnas and lncrnas related to fecundity in different sheep. *Scientific Reports*, **6**, 35299 EP –.

Noh, H., Park, C., Park, S., Lee, Y. S., Cho, S. Y., and Seo, H. (2014). Prediction of mirna-mrna associations in alzheimerâĽ™s disease mice using network topology. *BMC Genomics*, **15**(1), 644.

Nunez, Y. O., Truitt, J. M., Gorini, G., Ponomareva, O. N., Blednov, Y. A., Harris, R. A., and Mayfield, R. D. (2013). Positively correlated mirna-mrna regulatory networks in mouse frontal cortex during early stages of alcohol dependence. *BMC Genomics*, **14**(1), 725.

Parker, M. M., Chase, R. P., Lamb, A., Reyes, A., Saferali, A., Yun, J. H., Himes, B. E., Silverman, E. K., Hersh, C. P., and Castaldi, P. J. (2017). Rna sequencing identifies novel non-coding rna and exon-specific effects associated with cigarette smoking. *BMC Medical Genomics*, **10**(1), 58.

Fig. S4. TCGA breast cancer results. (a) The top left figure shows the feature- and module-wise correlations with uncensored survival time. The mRNAs are circles and the miRNAs are triangles. The points are organized by the module groups. The top panel is for the full modules; the bottom panel is for the trimmed module. All features that are not included in the modules are grouped as Module 0. The heatmaps are the right-hand-side denote the module-trait correlations with p-values in parenthesis. (b) The network figure shows the trimmed Module 1. Negative edge are in red; positive edges are in gray. The width of the network edges indicates the connection strength. Yellow nodes with purple labels are miRNAs; magenta nodes are genes. The size of the node corresponds to the degree, i.e. number of edges the node connects to in the trimmed network. Due to the large number of genes (circle in the network), the gene symbols are not displayed, but available in the supplement file TrimmedModules.xlsx

Ponsuksili, S., Du, Y., Hadlich, F., Siengdee, P., Murani, E., Schwerin, M., and Wimmers, K. (2013). Correlated mrnas and mirnas from co-expression and regulatory networks affect porcine muscle and finally meat properties. *BMC Genomics*, **14**(1), 533.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, **32**(9), 896.

Rudra, P., Shi, W. J., Vestal, B., Russell, P. H., Odell, A., Dowell, R. D., Radcliffe, R. A., Saba, L. M., and Kechris, K. (2017). Model based heritability scores for high-throughput sequencing data. *BMC Bioinformatics*, **18**(1), 143.

Russell, P. H., Vestal, B., Shi, W., Rudra, P. D., Dowell, R., Radcliffe, R., Saba, L., and Kechris, K. (2018). mir-magic improves quantification accuracy for small rna-seq. *BMC Research Notes*, **11**(1), 296.

Shkurnikov, M. Y., Knyazev, E., Fomicheva, K., Mikhailenko, D., Nyushko, K., Saribekyan, E., Samatov, T., and Alekseev, B. Y. (2016). Analysis of plasma microrna associated with hemolysis. *Bulletin of Experimental Biology and Medicine*, **160**(6), 748–750.

Vestal, B., Russell, P., Radcliffe, R., Bemis, L., Saba, L., and Kechris, K. (2018). mirna-regulated transcription associated with mouse strains predisposed to hypnotic effects of ethanol. *Brain and Behavior*, **8**(6), e00989.

Williams, R. W., Bennett, B., Lu, L., Gu, J., DeFries, J. C., Carosone-Link, P. J., Rikke, B. A., Belknap, J. K., and Johnson, T. E. (2004). Genetic structure of the lxs panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. *Mammalian Genome*, **15**(8), 637–647.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**(1), 1–27.

Wu, Y. E., Parikshak, N. N., Belgard, T. G., and Geschwind, D. H. (2016). Genome-wide, integrative analysis implicates microrna dysregulation in autism spectrum disorder. *Nature Neuroscience*, **19**(11), 1463.