## Methods supplemental.

## Constructing Bayesian network for Ovarian cancer by integrating gene expression and CNA data

The TCGA Ovarian serous cystadenocarcinoma (OV) data set, consisting of the gene expression and gene-based copy number alternation (CAN) data was downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). It has been shown that CNA contributes most significantly to ovarian cancer tumorigenesis and progression. CNA alters expression level of underlying genes directly. We first defined cis-CNA genes as genes whose expression levels and their CNAs are significantly correlated. Then, we can decompose expression variance of a gene  into multiple parts as  (**Figure S\***), due to its cis-CNA, due to its regulators , and their interactions.

As searching for an optimal Bayesian network structure is a NP-hard problem, we can only include a limited number of genes with informative gene expression or CNAs as nodes in our network reconstruction procedure. We selected a set of informative genes that were expressed in the tumor tissues (the mean expression levels >5) and whose expression levels varied (the standard deviation >0.5). We also included a set of cis-CNAs (p-value <0.01 for the Spearman’s correlation between gene expression and CNA after multiple testing correction) as nodes.

The selected gene expression and gene-based CNA profiles were input into a Bayesian network reconstruction software package, RIMBANet (Reconstructing Integrative Molecular Bayesian Network) [1-4](#_ENREF_1).ABayesian network is a directed acyclic graph in which the edges of the graph are defined by conditional probabilities that characterize the distribution of states of each node given the state of its parents [5](#_ENREF_5). The network topology defines a partitioned joint probability distribution over all nodes in a network, such that the probability distribution of states of a node depends only on the states of its parent nodes: formally, a joint probability distribution  on a set of nodes  can be decomposed as , where  represents the parent set of . In our networks, each node represents expression level or CNA of a gene. These conditional probabilities reflect not only relationships between genes, but also the stochastic nature of these relationships, as well as noise in the data used to reconstruct the network.

Bayes formula allows us to determine the likelihood of a network model  given observed data  as a function of our prior belief that the model is correct and the probability of the observed data given the model: . We can set a model’s prior probability based on biological knowledge. For genes with cis-CNAs, we assume that the expression variations of these genes were directly affected by their CNAs. To represent the assumption, we set a structure prior, which is equivalent to start a searching process with an initial structure with a set of  edges instead an empty initial structure. We also assume the cis-CNAs only affected expression levels of their cis genes directly, and any trans effects on other genes were through expression variations of their cis genes. Thus, we set the prior for .

The number of possible network structures grows super-exponentially with the number of nodes, so an exhaustive search of all possible structures to find the one best supported by the data is not feasible, even for a relatively small number of nodes. We employed Monte Carlo Markov Chain (MCMC) [6](#_ENREF_6) simulation to identify potentially thousands of different plausible network structures, which are then combined to obtain a consensus network (see below). Each reconstruction begins with a null network. Small random changes are then made to the network by flipping, adding, or deleting individual edges, ultimately accepting those changes that lead to an overall improvement in the fit of the network to the data. We assess whether a change improves the network model using the Bayesian Information Criterion BIC [7](#_ENREF_7), which avoids overfitting by imposing a cost on the addition of new parameters. This is equivalent to imposing a lower prior probability  on models with larger numbers of parameters.

Searching optimal BN structures given a dataset is an NP-hard problem. We employed an MCMC method to do local search of optimal structures. As the method is stochastic, the resulting structure will be different for each run. In our process, 1,000 BN structures were reconstructed using different random seeds to start the stochastic reconstruction process. From the resulting set of 1,000 networks generated by this process, edges that appeared in greater than 30% of the networks were used to define a consensus network. A 30% cutoff threshold for edge inclusion was based on our simulation study [8](#_ENREF_8), where a 30% cutoff yields the best tradeoff between recall rate and precision. The consensus network resulting from the averaging process may not be a BN (a directed acyclic graph). To ensure the consensus network structure is a directed acyclic graph, edges in this consensus network were removed if and only if (1) the edge was involved in a loop, and (2) the edge was the most weakly supported of all edges making up the loop.

**Reference**

**1. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics* 40, 854-861 (2008).**

**2. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and genome research* 105, 363-374 (2004).**

**3. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS computational biology* 3(2007).**

**4. Zhu, J. *et al.* Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS biology* 10(2012).**

**5. Pearl, J. *Probabilistic reasoning in intelligent systems : networks of plausible inference*, xix, 552 p. (Morgan Kaufmann Publishers, San Mateo, Calif., 1988).**

**6. Madigan, D.a.Y., J. Bayesian graphical models for discrete data. *International Statistical Review* 63, 215-232 (1995).**

**7. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* 6, 461-464 (1978).**

**8. Zhu, J. *et al.* Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. *PLoS Comput Biol* 3, e69 (2007).**

**Figure**

****

**Figure S\*** Gene expression variation of a gene is caused by CNA of the gene and binding of transcription factors (TFs) to its promoter region. **A.** low gene expression due to a low copy number of the gene; **B.** low gene expression due to none TF binding; **C.** high gene expression due to TF binding at the promoter region; **D.** high gene expression due to gene copy number amplification; **E** higher gene expression due to gene copy number amplification with TF binding at the promoter region.