Zheng et al., Supplemental Text S1. Naïve Bayesian Network.

Classification is a statistical procedure to place individual items into groups based on features associated with the items. A training set consists of already labeled items: positive and negative gold standards. Classification extracts rules from the training items, and then applies these rules to predict the grouping of new items. A simple example of a classification is to place two people into "friends" or "not friends" groups based on features, such as whether they live in the same town, like the same sports team, drive the same type of car, etc.

A naïve Bayesian network is a classification method developed by applying Bayes's law and with the naïve assumption of feature independence.

The prior odds is the chance of finding a positive example, and is defined by the ratio of positive probability to negative probability:

$$O_{prior} = \frac{P(pos)}{P(neg)}.$$

The posterior odds is the chance of finding a positive example after considering *n* features $f_1, f_2 \ldots f_n$:

$$O_{posterior} = \frac{P(pos \mid f_1, f_2 \ldots f_n)}{P(neg \mid f_1, f_2 \ldots f_n)}.$$

The likelihood ratio (LR) is the relative plausibility of the null hypothesis and the alternative hypothesis. The joint LR of multiple features is defined as:

$$LR(f_1, f_2 \ldots f_n) = \frac{P(f_1, f_2 \ldots f_n \mid pos)}{P(f_1, f_2 \ldots f_n \mid neg)}.$$

The joint LR can be calculated as the product of individual LRs by assuming feature independence:

$$LR(f_1, f_2 \ldots f_n) = LR(f_1) \times LR(f_2) \times \ldots \times LR(f_n) = \frac{P(f_1 \mid pos)}{P(f_1 \mid neg)} \times \frac{P(f_2 \mid pos)}{P(f_2 \mid neg)} \times \ldots \times \frac{P(f_n \mid pos)}{P(f_n \mid neg)},$$

where individual LRs can be easily obtained based on feature values of gold standards [1]. Individual LRs represent the predictive power of each feature.

The prior odds and posterior odds can be connected according to the Bayes' law:

$$O_{posterior} = LR(f_1, f_2 \ldots f_n) \times O_{prior}.$$

The posterior odds represents the quantitative prediction from the naïve Bayesian method. In the context of the HFOnet, the posterior odds is a measure of functional association between two genes. As the prior odds is a constant, the value of the posterior odds only depends on the joint LR. Therefore, the joint LR can be used as a quantitative metric for Bayesian predictions. In the context of the HFOnet, we use the joint LR as the measure of functional association between two genes. Any gene pairs with joint LR values greater than a predetermined threshold are predicted to be functionally related, and unrelated otherwise.

We can determine the threshold for the joint LR using the same Bayesian formula $O_{posterior} = LR(f_1, f_2 \ldots f_n) \times O_{prior}$. In order to achieve posterior odds greater than 1, the joint LR must be greater than $1/O_{prior}$. The prior odds can be estimated from the training data. For example, there are 4,606 genes and 20,034 gene pairs in our positive gold standards. The prior odds can be calculated as the ratio of interacting pairs to no interacting pairs, that is $O_{prior} = \frac{20034}{4606 \times 4605/2 - 20034} = \frac{1}{528}$. Therefore, the joint LR must be greater than 528 to achieve a posterior odds greater than 1. Any gene pairs with a joint LR value greater than 528 are predicted to be functionally associated. Otherwise they are functionally unrelated.

BOR Papers in Press. Published on February 19, 2010 as DOI: 10.1095/biolreprod.109.079590

Zheng et al., Supplemental Text S2. Network properties.

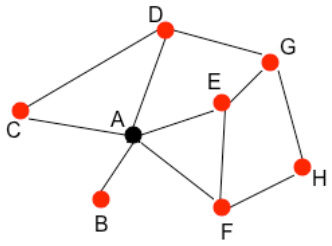
A network is a highly abstract picture of complex systems. It contains two entities, nodes and links. Nodes represent components, and links represent interactions between two components. Nodes and links together form a network (Figure a). Here we introduce a few concepts used in the paper that allow us to characterize the HFOnet network.

Degree: This is the number of directly linked neighbors for a given node. For example, node A has five links to nodes B, C, D, E, and F (Figure a). Therefore, the degree of node A is 5. Accordingly, degrees of other nodes in the network can be counted. The average degree of an undirected network with *N* nodes and *L* links is *2L/N*. There are 8 nodes and 11 links in the network of Figure a. Thus the average degree is 1.375.

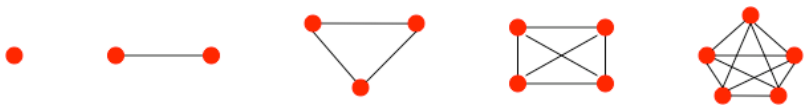
Clustering coefficient: Transitivity implies that if node A interacts with node C, and A interacts with D, then it is highly likely that D also has a direct link to C (Figure a). This is easy to understand in a friendship network: if A-C and A-D are good friends, then it is probable that C-D are also good friends. The ACD triangle implies transitivity among three nodes. Node transitivity can be quantified by the clustering coefficient with the formula $\frac{n}{k(k-1)/2}$, where *n* is the number of triangles passing through a node, *k* is the number of neighbors of the node, and $k(k-1)/2$ represents the number of all possible triangles going through the node [2]. Node A has 5 neighbors (B, C, D, E, F), thus the number of all possible triangles passing through node A is 10. Two triangles pass through node A (ACD, AEF). Therefore, the clustering coefficient of node A is 2/10 or 0.2. The average clustering coefficient of a network is the mean of the clustering coefficient for each node. It characterizes the overall tendency of nodes to form clusters in a network.

Clique: A clique is a set of nodes that every node is connected with every other node. If nodes represent people and links represent friendship, then a clique is a group of people in which everyone is everyone else's friend. The simplest clique is a single node. Figure b lists cliques in which the number of nodes ranges from 1 to 5. For cliques with more than two nodes, every node has a clustering coefficient equal to 1.

a. An undirected network



b. Cliques



BOR Papers in Press. Published on February 19, 2010 as DOI: 10.1095/biolreprod.109.079590

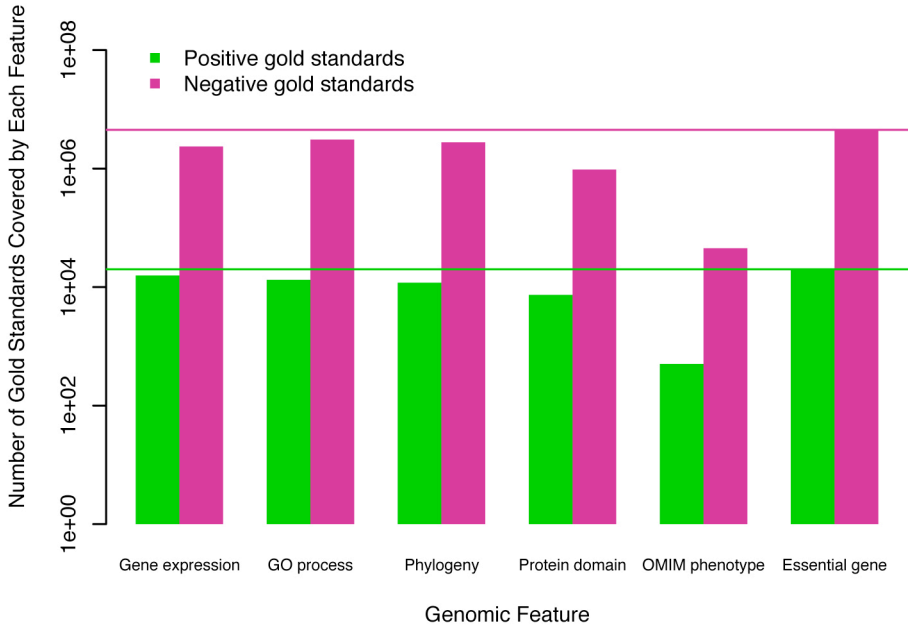
REFERENCES USED IN SUPPLEMENTAL DATA

1. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003; 302:449-453.

2. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998; 393:440-442.

BOR Papers in Press. Published on February 19, 2010 as DOI: 10.1095/biolreprod.109.079590

Zheng et al., Supplemental Figure S1. Number of gold standards covered by each genomic feature. The green and purple lines label the numbers of positive and negative gold standards, respectively.



BOR Papers in Press. Published on February 19, 2010 as DOI: 10.1095/biolreprod.109.079590

Zheng et al., Supplemental Table S1. Independence between genomic features and gold standards.

	Gene expression	GO process	Phylogeny	Protein domain	OMIM phenotype	Essential gene	Gold standard
Gene expression		0.01	0.02	0.01	0.00	0.00	0.02
Go process	0.00		0.15	0.10	0.03	0.05	0.13
Phylogeny	0.00	0.03		0.04	0.02	0.02	0.03
Protein domain	0.00	0.00	0.00		0.15	0.03	0.24
OMIM phenotype	0.00	0.00	0.00	0.00		0.00	0.08
Essential gene	0.00	0.00	0.00	0.00	0.00		0.05
Gold standard	0.00	0.01	0.00	0.00	0.00	0.00	

The top-right section records Pearson correlation coefficient, and the bottom-left section records mutual information.