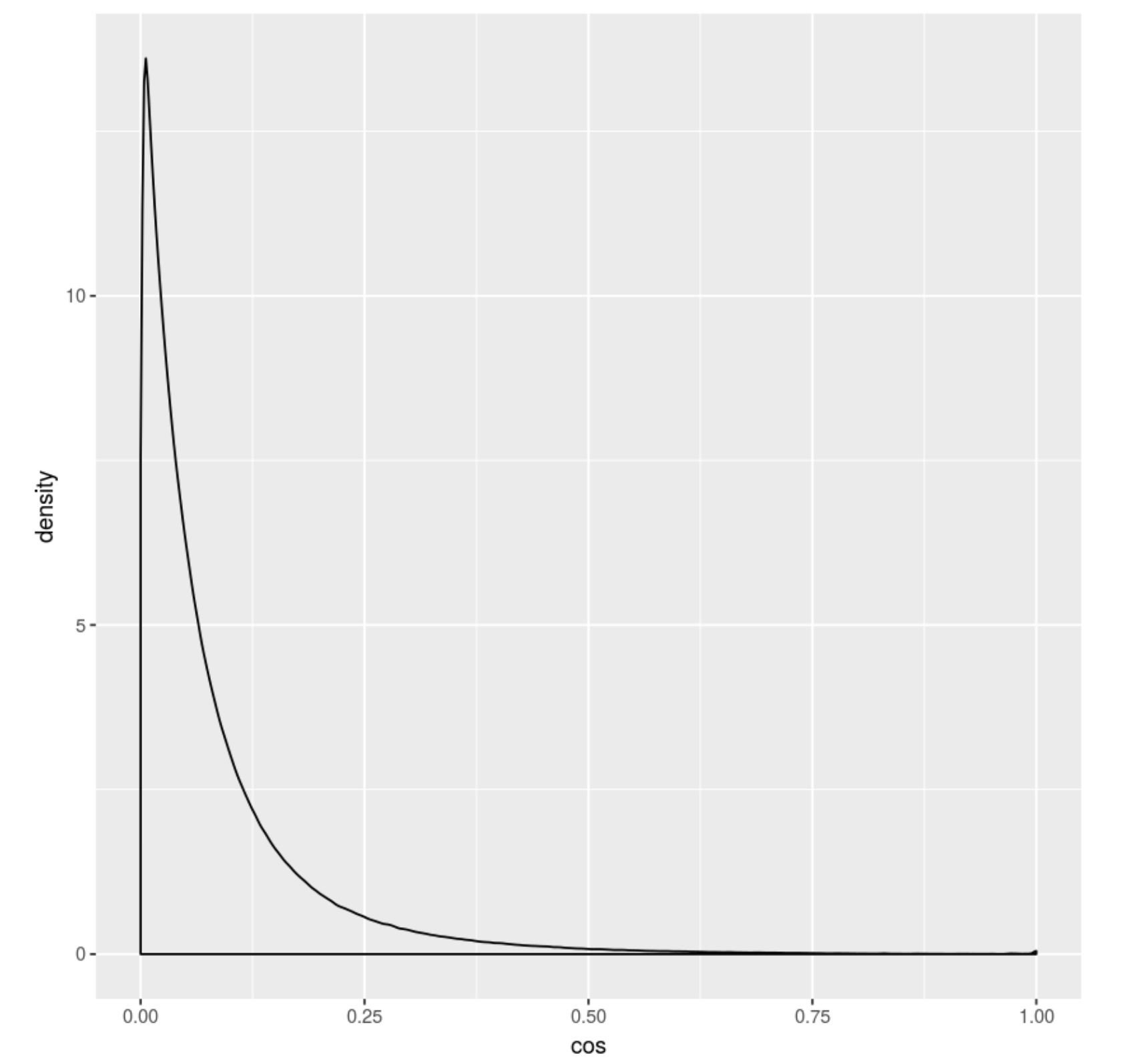
Hi all,

Just to clarify the point I was trying to make quickly today.

The distribution of cosine similarities in the database closely resembles a beta(0.514,5.470) distribution (the empirical density curve and percentiles of interest are shown below).



Quantiles:

25%

0.0194092998350546

50%

0.0499095612098125

75%

0.11037188161076

95%

0.294446060779202

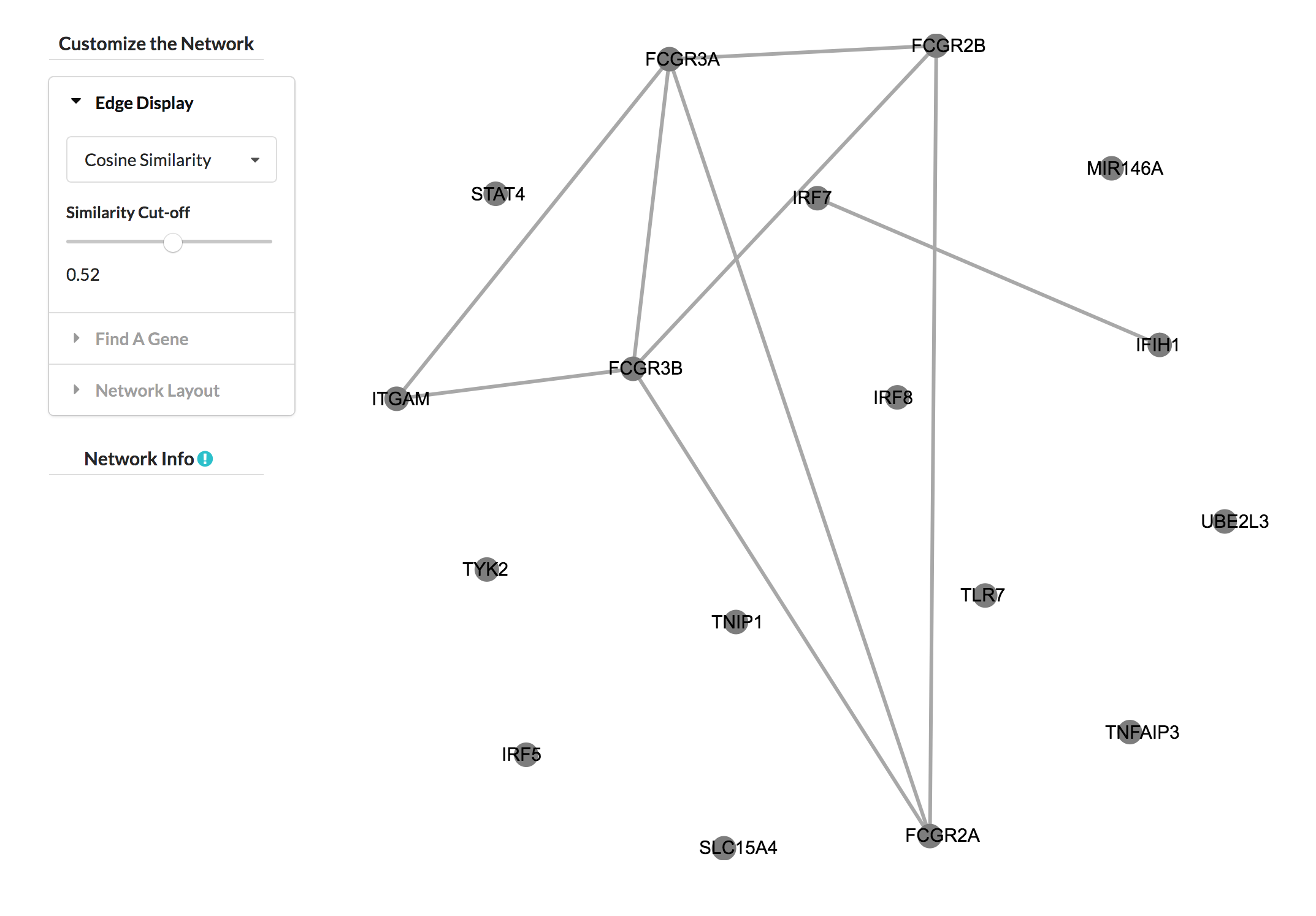
97.5%

0.389547128756657

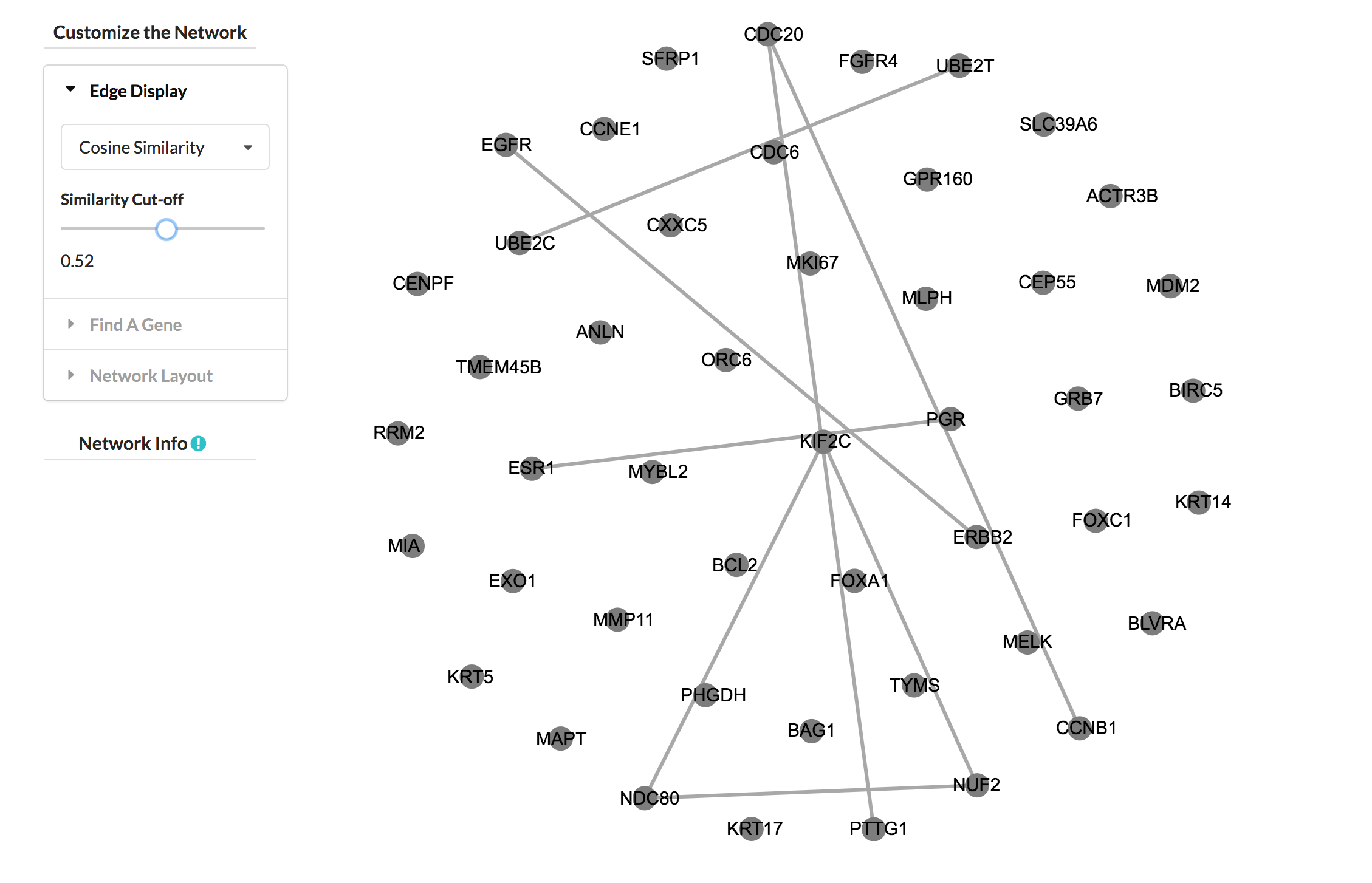
99%

0.519052089748183

Thus, using a cosine similarity cutoff of around 0.52 would result in only 1% of the total possible gene-gene pairs to be considered associated. When I looked at the example lupus genes in the Network Query with a cutoff of 0.52, I found 7 edges in the resultant network. With 18 genes in this list, there are 18 choose 2 = 153 possible edges. If these genes were just a random selection from the database, I would expect GAIL to return a network of 153\*0.01 = 1.5 edges. You could adopt a binomial(n = 153,p = 0.01) to obtain a p-value of 0.00016 associated with selecting 7 edges when the probability of an edge is 0.01. Thus, 7 edges appearing in a set of genes known to be related gives evidence to support GAIL's performance.



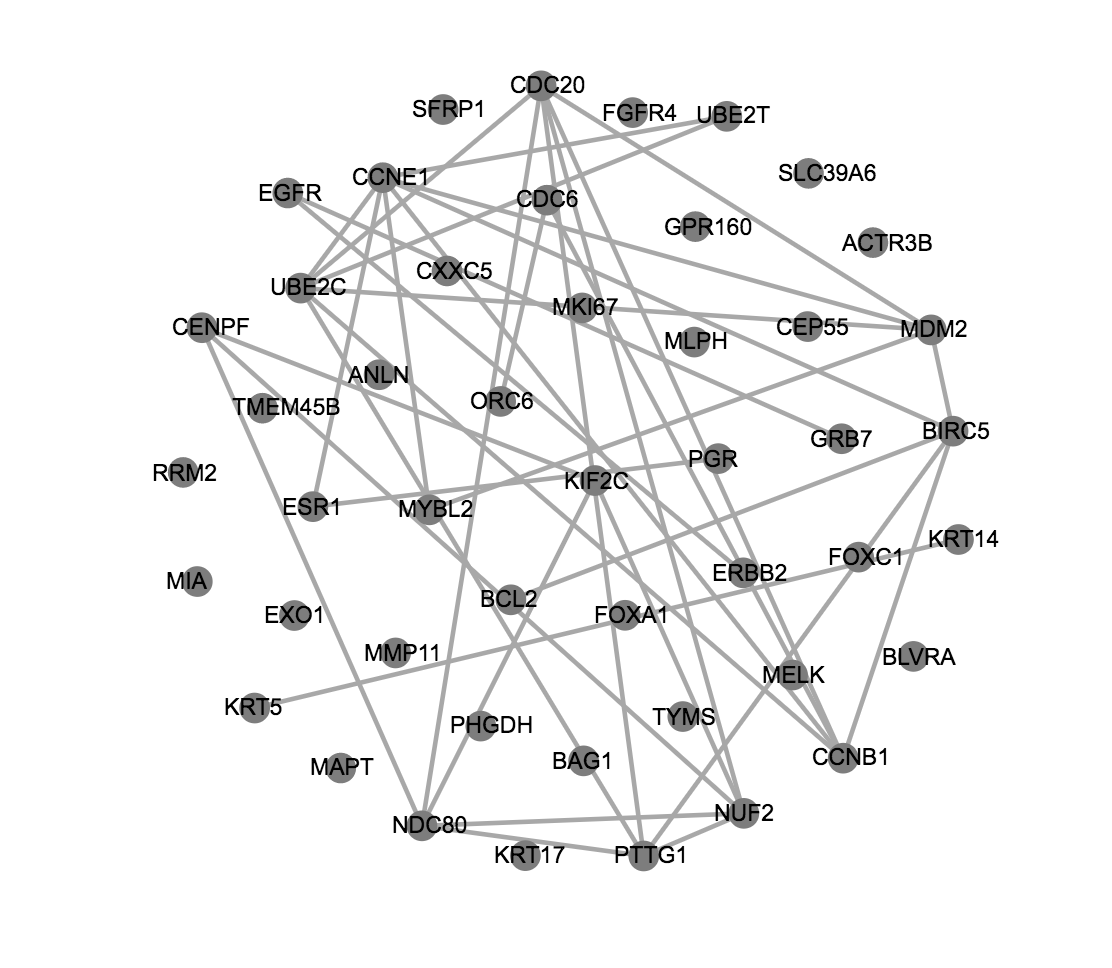
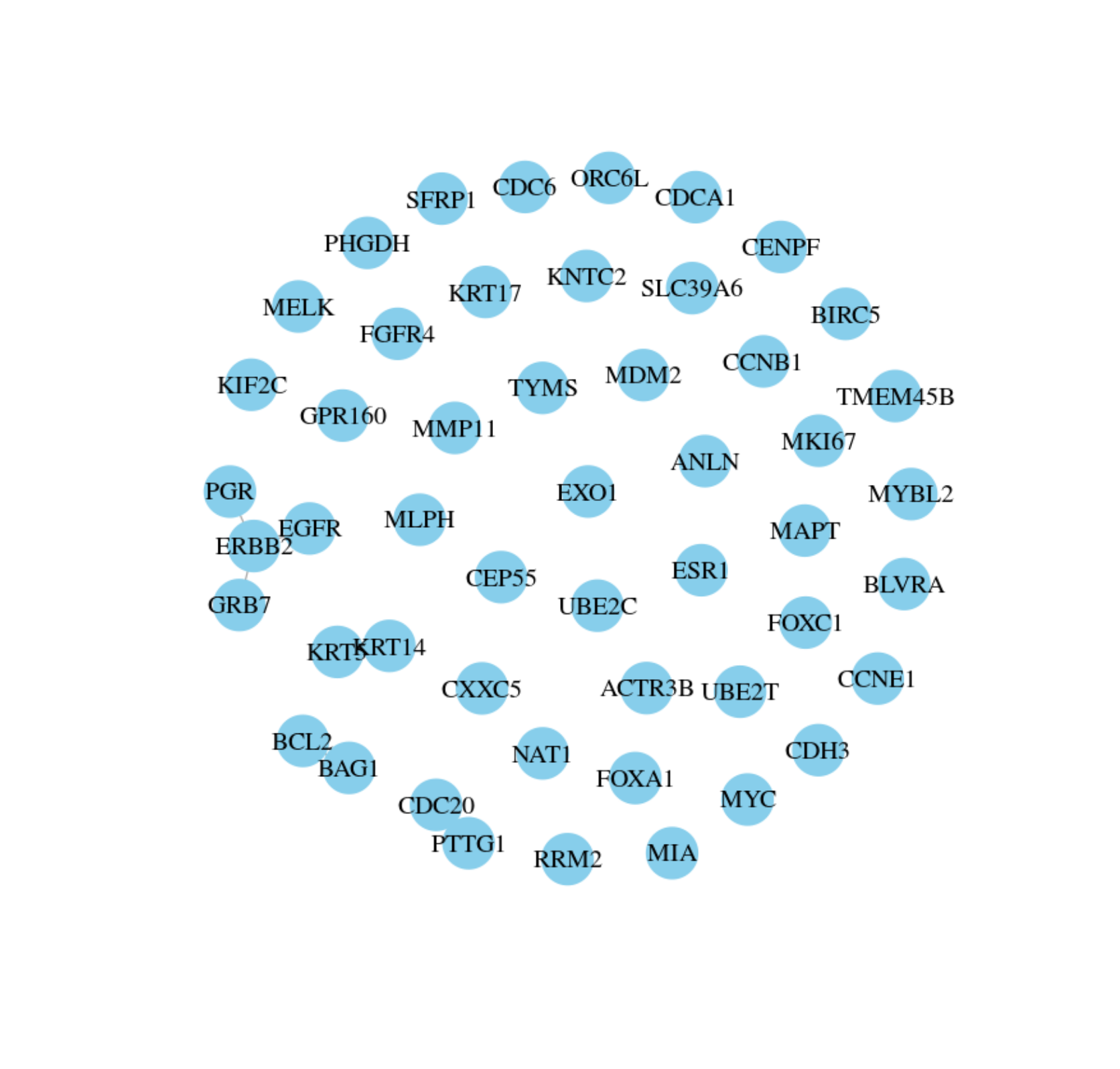
However, when I investigated the Pam50 genes at the same similarity cutoff I found the following network.



With 50 genes in the input list, there is a total of 50 choose 2 = 1225 possible edges. Here, the expected number of edges is 1225\*0.01 = 12.25, but I count only 9 edges. So, GAIL is providing no evidence that the Pam50 genes is not just a random selection of 50 genes.

The main caveat with this however is that GAIL is not able to provide edges where missing data is present, which causes the number of edges present in a given network to be almost always artificially low. It may well be that many of the edges we would expect to be present in the Pam50 network would be present if there were no missing values. I'll have to look into this further.

I have found that the GAIL networks for the Pam50 genes are in general more rich than those given by a simple co-occurrence network. You can compare the two networks below, where the blue network is a simple co-occurrence network. Both networks are built under a cosine similarity cutoff of 0.29 (roughly the 95th percentile of cosine similarities in the database).



GAIL's network much more closely reflects the clustering described in the original Pam50 paper (http://ascopubs.org/doi/full/10.1200/JCO.2008.18.1370).

I hope this helps.

Carter