Supplementary Information

A strategy to incorporate prior knowledge into correlation network cutoff selection

Benedetti et al.
Supplementary Figures

**Supplementary Figure 1: Size dependence of statistical cutoffs in the glycomics replication cohorts.**

A, C, E Correlation cutoff (0.01 FDR, Benjamini-Hochberg) as a function of the dataset sample size for the three correlation measures considered (Pearson correlation in red, exact partial correlation in purple, and shrinkage partial correlation based on GeneNet in black) in the Korčula 2010 (A), Split (C) and Vis (E) cohorts, respectively. For each dataset size, the original dataset was subsampled 1,000 times. Curves represent the mean correlation cutoff across the subsampling results for the different correlation measures. Error bars represent the corresponding 95% confidence intervals across the subsamplings. B, D, F Number of edges in the correlation network after applying a 0.01 FDR cutoff as a function of the dataset sample size in the Korčula 2010 (B), Split (D) and Vis (F) cohorts, respectively. For each dataset size, the original dataset was subsampled 1,000 times. Curves represent the mean number of edges across the subsampling results for the different correlation measures. Error bars represent the corresponding 95% confidence intervals across the subsampling. Note that for parcor, correlation values can only be estimated for a sample size greater or equal to the number of variables, in this case 50.
Supplementary Figure 2: Network quality as a function of the correlation cutoff in the glycomics replication cohorts.

A, C, E Fisher’s exact test p-values as a function of the correlation cutoff calculated in the Korčula 2010 (A), Split (C) and Vis (E) cohorts, respectively, for three correlation estimators: Pearson correlation (red), exact partial correlation (purple), GeneNet partial correlation (black). For each correlation cutoff, the original dataset was bootstrapped 1,000 times. Curves represent the mean negative log10 of the two-sided Fisher’s exact test p-values over the bootstrapping results for the different correlation measures. Error bars represent the 95% confidence intervals of the bootstrapping results. Dashed lines represent the mean of the bootstrapped statistical cutoffs for GeneNet. Interestingly, the minima of the GeneNet curve are fairly similar across cohorts: 0.24 (Korčula 2010), 0.23 (Split), 0.24 (Vis). B, D, F Mean negative log10 of the two-sided Fisher’s exact test p-value for partial correlations estimated with GeneNet in the Korčula 2010 (B), Split (D) and Vis (F) cohorts, respectively, as a function of both sample size and correlation cutoff. The red line represents the mean of the 0.01 FDR cutoff across the bootstrapping.
Supplementary Figure 3: Cutoff optimization as a function of the sample size for parcor correlation in the glycomics replication cohorts.

A, B, C, D Mean negative log10 of the two-sided Fisher’s exact test p-value for partial correlations estimated with parcor in the Korčula 2013 (A), Korčula 2010 (B), Split (C) and Vis (D) cohorts, respectively, as a function of both sample size and correlation cutoff. The red line represents the mean of the 0.01 FDR cutoff across the bootstrapping.
Supplementary Figure 4: Cutoff optimization as a function of the sample size for Pearson correlation in the glycomics replication cohorts.

A, B, C, D Mean negative log10 of the two-sided Fisher’s exact test p-value for Pearson correlations in the Korčula 2013 (A), Korčula 2010 (B), Split (C) and Vis (D) cohorts, respectively, as a function of both sample size and correlation cutoff. The red line represents the mean of the 0.01 FDR cutoff across the bootstrapping.
Supplementary Figure 5: Cutoff optimization with partial knowledge in the glycomics replication cohorts.

A, C, E Incomplete biological reference in the Korčula 2010 (A), Split (C) and Vis (E) cohorts. For each percentage, 100 different adjacency matrices were generated by randomly removing edges from the IgG glycosylation pathway. The curves in the figure represent the mean negative log10 of the two-sided Fisher's exact test p-value across 1,000 bootstrapping on each adjacency matrix.

B, D, F Incorrect biological reference in the Korčula 2010 (B), Split (D) and Vis (F) cohorts. Edges in the IgG glycosylation pathway were randomly swapped to simulate incorrect information in the biological reference. For each number of swaps, 100 adjacency matrices were generated and the mean negative log10 of the two-sided Fisher’s exact test p-value across those curves and across the 1,000 bootstrapping are shown as curves in the plot. Here, the red curve represents the mean negative log10 of the two-sided Fisher’s exact test p-value across 100 fully randomized adjacency matrices (rnd). The error bars on this curve represent the 95% confidence interval of the bootstrapping. Any signal that falls within these intervals should be regarded as noise.
Supplementary Figure 6: Cutoff optimization with coarse prior knowledge in the glycomics replication cohorts.

A, B, C Coarse biological reference in the Korčula 2010 (A), Split (B) and Vis (C) cohorts. For IgG glycomics data we know that only enzymatic reactions between glycans attached to the same IgG isoform are feasible (adjacency matrix 1, green) and, in addition, that only they can be modified by the addition of one sugar unit at a time (adjacency matrix 2, blue). The black curve corresponds to the optimization performed on the full reference (adjacency matrix 3, black) for comparison. For each correlation cutoff, the original dataset was bootstrapped 1,000 times. The curves in the figure represent the mean negative log10 of the two-sided Fisher’s exact test p-value across the 1,000 bootstrapping for the different adjacency matrices. Error bars represent the 95% confidence intervals of the bootstrapping results. In all plots, the black curve corresponds to the optimization performed on the full reference.
Supplementary Figure 7: Parcor (A) and Pearson (B) correlation cutoff optimization with partial and coarse knowledge in the metabolomics cohort.

We used biochemical reactions from the RECON database as partial prior knowledge (adjacency matrix 1, blue), as well as sub- and super-pathway annotations (adjacency matrices 2 and 3, in black and green, respectively). For each correlation cutoff, the original dataset was bootstrapped 1,000 times. Curves represent the mean negative log10 of the two-sided Fisher’s exact test p-value across the bootstrapping. Error bars represent the 95% confidence intervals of the bootstrapping results.

Supplementary Figure 8: Cutoff optimization for the transcriptomics data (see following page).

For each of the 469 pathway considered, protein-protein interaction networks from STRING were used as reference. The black curve represents the mean negative log10 of the two-sided Fisher’s exact test p-value across 100 bootstrapping, and the error bars show the corresponding 95% confidence intervals. Vertical lines indicate the mean of the statistical cutoffs, and the areas the corresponding 95% confidence intervals over the bootstrapping. Pathways are shown in alphabetical order.
Supplementary Methods

Estimation of correlation p-values

The generalized probability density function $f_0$ for Pearson and partial sample correlation coefficients $r$ can be written as

$$f_0(r; \kappa) = (1 - r^2)^{\frac{\kappa - 3}{2}} \frac{\Gamma \left[ \frac{\kappa}{2} \right]}{\pi^2 \Gamma \left[ \frac{\kappa - 1}{2} \right]}$$

(1)

where $\kappa$ indicates the degrees of freedom, and $\Gamma$ is the gamma function.

For Pearson correlation the degrees of freedom is $\kappa = N - 1$, and for parcor $\kappa = N - G + 1$ where $N$ is the sample size and $G$ is the number of variables. Of note, this demonstrates that, for regular partial correlation, $N$ cannot be smaller than $G$ if $\kappa$ is to remain positive.

For any given observed correlation value, the test statistic can be computed from the null probability density above via Fisher’s transformation. A p-value can be directly obtained from this value using Student’s t distribution. Since there is a direct dependency of the test statistic on the square root of the sample size (see equation above), increasing sample sizes will increase the test-statistic, and subsequently decrease the p-value, as observed in Figure 2 of the main manuscript.

The GeneNet approach is based on a regularized estimation of the sample covariance matrix $\Sigma$ via a shrinkage parameter $\lambda$, resulting in a positive-definite and stable covariance matrix even in the case of $N < G$. As shown by the authors, the empirical distribution of the partial correlation coefficients estimated with this approach still follows the generalized probability density function form above, with $\kappa > 0$ even for $N < G$. In this case, $\kappa$ is not a simple function of $N$ and $G$, and has to be estimated from the shrinkage covariance matrix. Once $\kappa$ has been estimated, the probability density function again allows us to compute p-values for estimated partial correlation coefficients, which this time will not directly depend on the sample size $N$. 
Supplementary References

1. Hotelling, H., New Light on the Correlation Coefficient and Its Transform, *Journal R. Stat. Soc. Ser. B* 15:2, 193-232 (1953).

2. Schafer, J., Strimmer, K., An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics* 21:6, 754–764 (2005).

3. Schäfer, J., Strimmer, K., A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat. Appl. Genet. Mol. Biol.* 4:1 (2005).

4. Efron, B., Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis, *J. Am. Stat. Assoc.* 99:465, 96-104 (2004).

5. Efron, B., Correlation and Large-Scale Simultaneous Significance Testing, *J. Am. Stat. Assoc.* 102:447, 93–103 (2007).