Supplementary Information for

Enhancing the accuracy of Network Medicine through understanding the impact of sample size in gene co-expression networks

Joaquim Aguirre-Plans[1,2,3], Bingsheng Chen[1], Susan Dina Ghiassian[4], Alex Jones[4], Viatcheslav R. Akmaev[4], Alif Saleh[4], Deisy Morselli Gysi[1,2,5], Albert-Laszlo Barabasi[1,2,5,6,*]

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, MA 02115, USA; [2]US Department of Veteran Affairs, Boston, MA 02130, USA; [3]STALICLA Discovery and Data Science Unit, Barcelona 08039, Spain; [4]Scipher Medicine Corporation, Waltham, MA 02453, USA; [5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; [6]Department of Network and Data Science, Central European University, Budapest 1051, Hungary. * Corresponding Author. **Email:** barabasi@gmail.com

This PDF file includes:

Supplementary Text

Text S1: Derivation of the test statistic for the Pearson correlation coefficient

The equation to determine the critical value of Pearson's correlation coefficient (ρ) for a given sample size (n) is derived from the t-statistic formula used for testing the significance of a correlation coefficient. The derivation of this statistic is rooted in the distribution of the correlation coefficient under the null hypothesis, which posits no correlation. For correlation coefficients under the null-hypothesis, the sample correlation is approximately normally distributed, with standard error (SE) as follows:

$$SE(\rho) = \sqrt{\frac{1-\rho^2}{n-2}} \qquad \text{Eq. S1}$$

Here, ρ is the sample correlation coefficient and n is the sample size. The t-statistic is obtained by dividing the sample correlation coefficient ρ by this standard error as follows:

$$t_\alpha = \frac{\rho}{SE(\rho)} = \frac{\rho}{\sqrt{\frac{1-\rho^2}{n-2}}} = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \qquad \text{Eq. S2}$$

Here, t_α is the t-statistic for a given level of significance (α). The t-statistic is used to test the null hypothesis that the population correlation coefficient is zero (i.e., there is no linear correlation). A t-statistic exceeding the critical value from the t-distribution in absolute terms leads to the rejection of the null hypothesis, thereby affirming the presence of a significant linear correlation. This statistical approach presumes that the data adheres to a bivariate normal distribution, implying that each variable is normally distributed and the inter-variable relationship is linear. This holds approximately in case of non-normal observed values if sample sizes are large enough.

Supplementary Figures

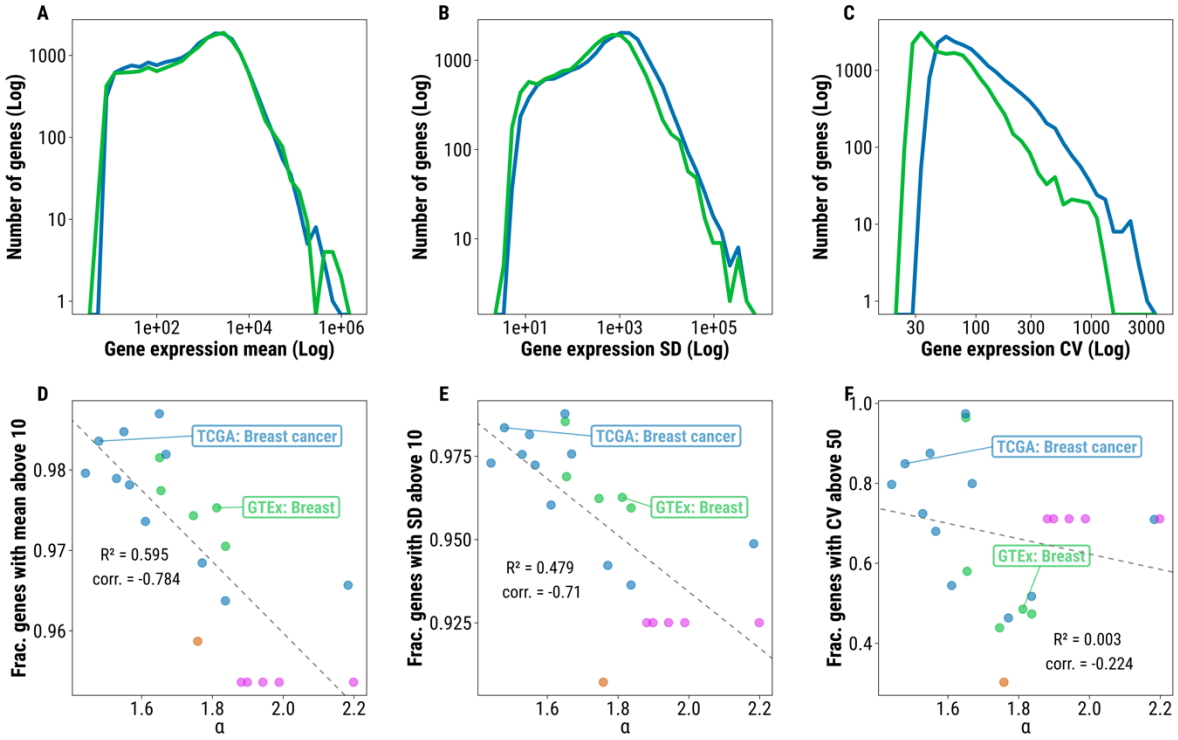


Figure S1. Relationship between gene expression variability and scaling exponent (α) in analytical model. (A), **(B)** and **(C)** display the distribution of gene expression mean, standard deviation (SD), and coefficient of variation (CV), respectively, for TCGA Breast Cancer and GTEx Breast datasets. While the gene expression mean distributions are similar, the SD and CV distributions are skewed towards higher variability in TCGA Breast Cancer. **(D)**, **(E)** and **(F)** illustrate the association between scaling exponent (α) and the proportion of genes with a given gene expression mean, SD, and CV. While the gene expression CV shows no relationship with α, the SD and mean exhibit a clear linear relationship (with a few outliers). This indicates that datasets with a higher proportion of genes having an SD or mean above 10 have a slower convergence towards significant correlations in gene co-expression networks.

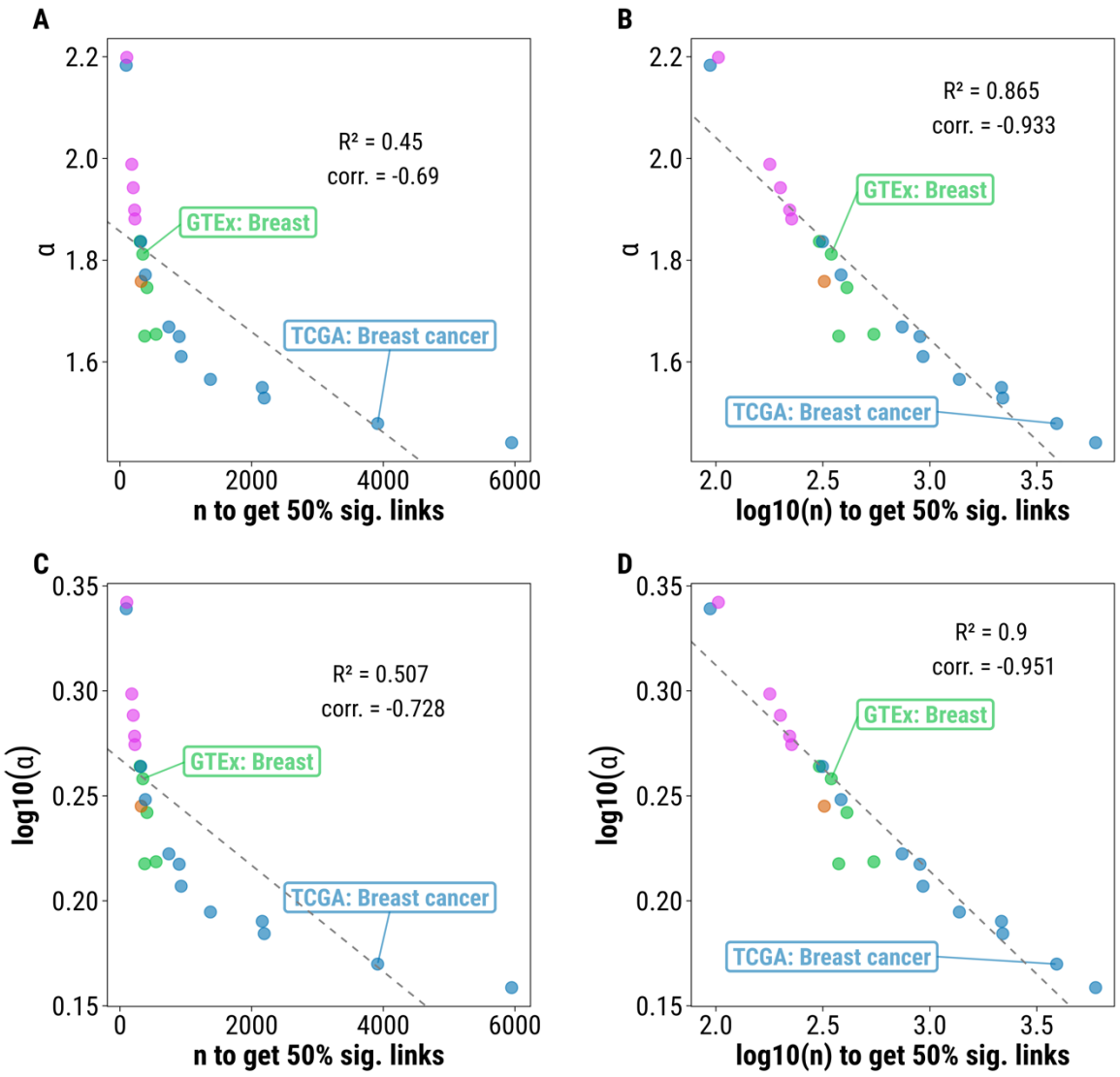
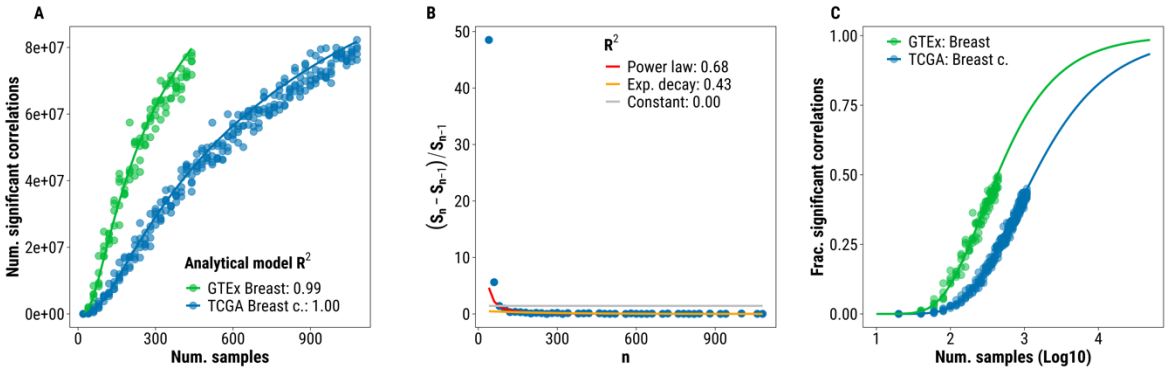


Figure S2. Relationship between the scaling exponent (α) and the number of samples predicted by the power-law model to get 50% of the links. The panels **(A)**, **(B)**, **(C)** and **(D)** show respectively the linear-linear, linear-logarithm, logarithm-linear and logarithm-logarithm relationships between the scaling exponent α and the number of samples predicted by the power-law model to get a statistical significance higher than 0.05 in 50% of the links of the co-expression networks. The analysis is performed for 22 subsets of the 4 original datasets (R. arthritis in orange, GTEx in green, TCGA in blue and GSE193677 in pink), highlighting the examples of GTEx Breast and TCGA Breast cancer. The relationship observed in the panels B (linear-logarithm) and D (logarithm-logarithm) is linear, as supported by adjusted R^2 values of

0.839 and 0.881 respectively, and indicate that the greater the α of the model is, the smaller the number of samples is required to get 50% of significant links.



D

	Power law		Exponential decay		Analytical model		
	R^2	ε	R^2	ε	α	R^2	ε
UC inflamed	0.76	1.19	0.64	1.66	2.18	0.98	0.43
GTEx: Breast	0.77	1.07	0.54	1.53	1.80	0.99	0.14
GTEx: Lung	0.70	0.81	0.44	1.03	1.72	0.99	0.33
R. arthritis	0.85	0.63	0.65	0.91	1.83	0.99	0.23
TCGA: Lung cancer	0.81	0.85	0.62	1.13	1.89	1.00	0.15
TCGA: Breast cancer	0.68	1.31	0.43	1.77	1.65	1.00	0.06

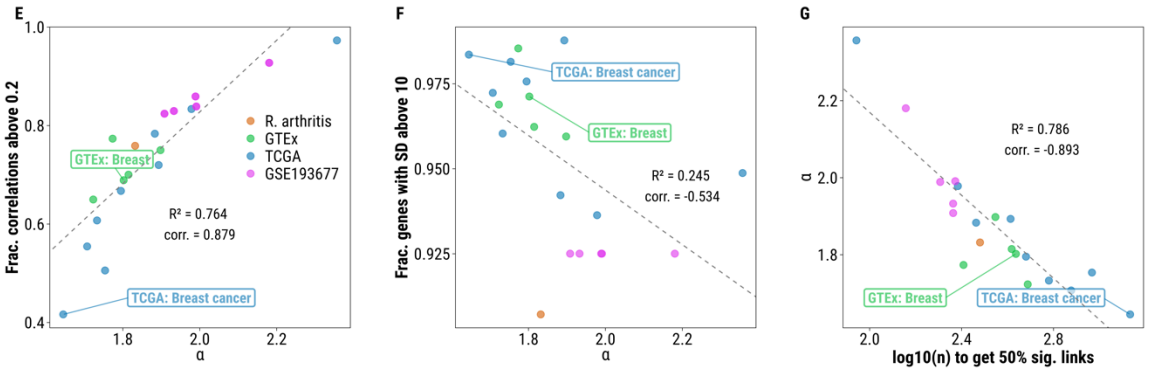


Figure S3. Reproduction of the power-law scaling relationship between sample size and gene co-expression network link significance using Spearman's correlation. (A) The scatterplot illustrates the same pattern as in Figure 2: a diminishing return pattern. The lines in the plot represent the fitting of the analytical model derived from the power-law scaling relation for two different datasets (GTEx breast mammary tissue and TCGA breast cancer), showing a highly accurate fitting (adjusted R^2 of 1 approximately). **(B)** The scatterplot shows a comparison between the fitting of the power law model (red, R^2 = 0.68), the exponential decay model (orange, R^2 = 0.43) and the constant discovery model (grey, R^2 =

0) on data from TCGA breast cancer, depicting the relationship between the link discovery rate and sample size. The fit of the power law model is very similar to the one observed using Pearson's correlation data (**Figure 2B**). **(C)** Prediction of network evolution across sample size, comparing empirical values (as points) to analytical model predictions (as lines) for GTEx breast mammary tissue and TCGA breast cancer datasets. **(D)** The table summarizes the adjusted R^2 and mean relative error (ε) values for three models: the power-law model, the exponential decay model (both depicted in **Figure S3B**), and the analytical model derived from the power-law relationship (shown in **Figures S3A** and **S3C**). These metrics are evaluated across six datasets (UC inflamed, GTEx breast, GTEx lung, R. arthritis, TCGA lung cancer, and TCGA breast cancer). The results highlight improvements in R^2 and reductions in ε when using the power-law and analytical models compared to the exponential decay model. The analytical model also includes a column for the discovery exponent (α), which demonstrates higher values for faster-converging datasets, such as GTEx breast (α = 1.80) compared to slower-converging datasets like TCGA breast cancer (α = 1.65). **(E)** Depiction of the linear relationship (adjusted R^2 of 0.764 and Pearson's correlation of 0.879) between the discovery rate (α) and the fraction of correlations above 0.2 predicted by the model for 22 subsets of 4 datasets (R. arthritis in orange, GTEx in green, TCGA in blue and GSE193677 in pink). **(F)** Scatterplot revealing the relationship (adjusted R^2 of 0.245 and Pearson's correlation of -0.534) between α and the fraction of genes with standard deviation above 10 for 22 subsets. **(G)** Scatterplot showing the linear relationship (adjusted R^2 of 0.786 and Pearson's correlation of -0.893) between α and the logarithm of the number of samples predicted by the model to achieve 50% of significant links for 16 subsets.

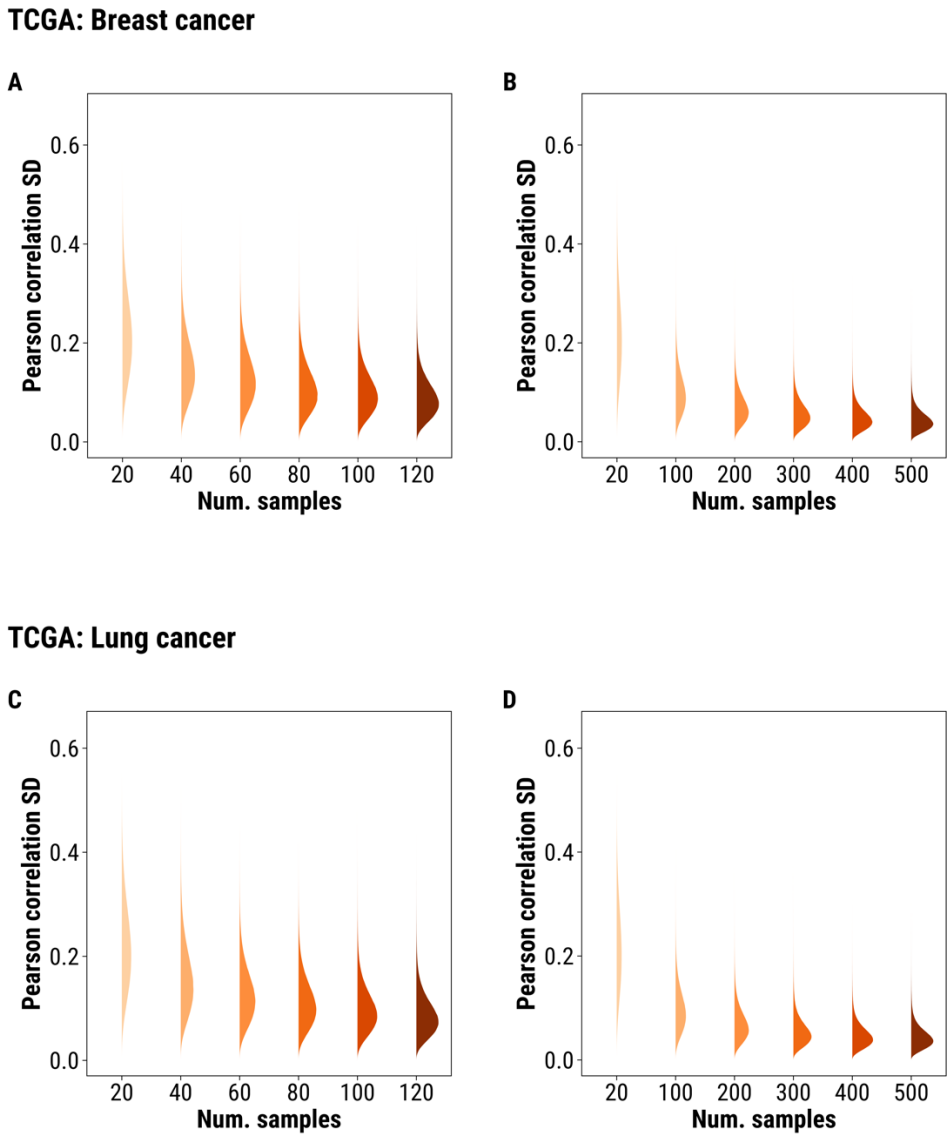


Figure S4. Assessmen of correlation variability across five distinct gene co-expression networks derived from equivalent sample sizes. Specifically, one million correlations were selected at random, and their standard deviation (SD) values were computed for each of the specified sample size ranges. Panels **(A)** and **(B)** depict networks constructed from TCGA Breast cancer, where sample sizes varied from 20 to 120 and 20 to 500, respectively. Conversely, panels **(C)** and **(D)** exhibit networks constructed from TCGA Lung cancer, with sample sizes spanning 20 to 120 and 20 to 500, respectively.

Supplementary Tables

Table S1. Statistical comparison of the co-expression of direct protein-protein interactions (PPI) with a negative set of randomly selected protein pairs with a distance greater than 3 (D > 3) in the human protein-protein interactome.

Size	*Abs. Corr. Mean (PPI)*	*Abs. Corr. Mean (D > 3)*	*Mean Diff.*	*Abs. Corr. SD (PPI)*	*Abs. Corr. SD (D > 3)*	*SD Diff.*	*Wilcox. Stat.*	*Wilcox. P-value*
20	0.48	0.348	0.131	0.223	0.195	0.027	1.14E+11	< 2.2e-16
100	0.443	0.286	0.158	0.229	0.18	0.049	1.18E+11	< 2.2e-16
200	0.422	0.265	0.157	0.232	0.179	0.053	1.18E+11	< 2.2e-16
300	0.432	0.263	0.169	0.237	0.184	0.054	1.19E+11	< 2.2e-16
400	0.429	0.259	0.17	0.236	0.182	0.054	1.20E+11	< 2.2e-16
500	0.418	0.249	0.169	0.237	0.18	0.057	1.19E+11	< 2.2e-16

Table S2. Statistical comparison of the co-expression of protein pairs of different distances in the human protein-protein interactome.

Size	*KW Stat.*	*KW p-value*	*Comparison*	*Abs. Corr. Mean Diff.*	*Dunn Stat.*	*Dunn p-value unadj.*	*Dunn p-value adj.*
20	9.94E+04	< 2.2e-16	1 - 2	0.025	40.29	0	0
			1 - 3	0.083	134.23	0	0
			1 - 4	0.132	278.58	0	0
			1 - 5	0.238	150.17	0	0
			1 - 6	0.285	16.43	1.14E-60	5.71E-60
			2 - 3	0.058	84.72	0	0
			2 - 4	0.106	188.63	0	0
			2 - 5	0.213	132.51	0	0
			2 - 6	0.26	15.08	2.33E-51	9.32E-51
			3 - 4	0.048	81.92	0	0
			3 - 5	0.155	95.42	0	0
			3 - 6	0.202	11.70	1.35E-31	4.05E-31
			4 - 5	0.107	66.93	0	0
			4 - 6	0.154	8.85	8.41E-19	1.68E-18

			5 - 6	0.047	2.78	5.49E-03	5.49E-03
100	1.38E+05	< 2.2e-16	1 - 2	0.027	39.55	0	0
			1 - 3	0.097	148.75	0	0
			1 - 4	0.159	324.35	0	0
			1 - 5	0.28	181.35	0	0
			1 - 6	0.36	21.71	1.82E-104	9.08E-104
			2 - 3	0.07	98.17	0	0
			2 - 4	0.131	227.11	0	0
			2 - 5	0.253	163.32	0	0
			2 - 6	0.333	20.37	2.83E-92	1.13E-91
			3 - 4	0.062	103.03	0	0
			3 - 5	0.183	120.28	0	0
			3 - 6	0.263	16.46	7.53E-61	2.26E-60
			4 - 5	0.121	84.46	0	0
			4 - 6	0.202	12.88	5.54E-38	1.11E-37
			5 - 6	0.081	5.21	1.91E-07	1.91E-07
200	1.33E+05	< 2.2e-16	1 - 2	0.028	39.98	0	0
			1 - 3	0.096	144.47	0	0
			1 - 4	0.158	319.91	0	0
			1 - 5	0.266	174.86	0	0
			1 - 6	0.336	20.66	7.50E-95	3.75E-94
			2 - 3	0.068	94.02	0	0
			2 - 4	0.13	223.01	0	0
			2 - 5	0.238	156.80	0	0
			2 - 6	0.308	19.32	3.86E-83	1.54E-82
			3 - 4	0.062	103.90	0	0
			3 - 5	0.17	115.58	0	0
			3 - 6	0.24	15.56	1.26E-54	3.79E-54
			4 - 5	0.108	79.28	0	0
			4 - 6	0.178	11.96	5.62E-33	1.12E-32
			5 - 6	0.07	4.76	1.96E-06	1.96E-06
300	1.46E+05	< 2.2e-16	1 - 2	0.029	40.14	0	0
			1 - 3	0.102	149.44	0	0
			1 - 4	0.17	337.46	0	0
			1 - 5	0.278	178.62	0	0
			1 - 6	0.352	21.22	6.37E-100	3.19E-99
			2 - 3	0.073	98.28	0	0
			2 - 4	0.142	237.32	0	0

2 - 5	0.25	160.43	0	0
2 - 6	0.323	19.87	7.71E-88	3.08E-87
3 - 4	0.069	112.61	0	0
3 - 5	0.176	117.36	0	0
3 - 6	0.25	15.95	3.05E-57	9.15E-57
4 - 5	0.108	77.78	0	0
4 - 6	0.182	12.04	2.17E-33	4.34E-33
5 - 6	0.074	4.97	6.67E-07	6.67E-07