Supplementary Information

Constructing the synthetic dataset

In order to formulate our dataset containing signals A-E, two matrices, representing synthetic transcriptomics and synthetic proteomics datasets were generated each with dimensions 9000 x 50. 3,000 of the 9,000 rows arise from the multi-level relationships (signals A-E). The remaining 6,000 signals result from unrelated random relationships. Each signal, and subsequent mRNA-protein relationship (mRNA X related to protein Y) is defined as interactions between four elements/factors (mRNA X, Protein Y, mRNA Y and Protein X). The synthetic data contained 300 example relationships of each type of signal. For each of these relationships, as mentioned above, the synthetic mRNA and synthetic protein datasets contained 2 expression profiles each.

We explored and utilized specific computational methods to generate the synthetic data that would contain the underlying structure that characterized biological relationships. To achieve this, we utilized R package *MASS* and the utility "*mvrnorm*"(Ripley *et al.*, 2011). The *mvrnorm* function allows us to generate random variables from specified multivariate distribution with a specific positive definite matrix as the covariance of the variables generated. Within the constraints of having a positive definite covariance matrix, we generate the 4 vectors (mRNA X, mRNA Y, Protein X, Protein Y) with pairwise high or low correlation as required by the aforementioned signals. The value of high correlation is chosen from a uniform distribution of correlation values ranging from 0.6 to 1 (pvalue <0.00001). Similarly, a low correlation value is chosen from a uniform distribution ranging from -0.2 to 0.2 (pvalue ~ 0.16). This allowed us to computationally generate examples of each of the proteogenomic relationships described as described and visualized in Figure 1 (b) to include in the synthetic dataset.

Supplementary Section 1. Details of methods by which the synthetic dataset was constructed

NCI-60 preprocessing and data matching

The protein dataset identified 8114 unique protein IDs while the microarray analysis employed ~53000 investigative probe sets. After appropriate annotation and matching, both datasets were filtered down to ~6300 genes across 57 cell lines. This experimental data is a well-established resource for researchers to investigate a variety of facets of cancer and it is unique in that a matched or paired transcriptomics and proteomic data exists.

The original data are in the form of two matrices, downloaded from the Cellminer database(Shankavaram *et al.*, 2009) and the NCI-60 proteome resource(Gholami *et al.*, 2013). For the proteomics data (LFQ: Label Free Quantification), each International Protein Index (IPI) associated with a proteome profile was translated to a corresponding protein (HUGO) name. We used the "Majority Accession" IPI index provided by the Gholami *et al.* in their supplementary information (data table 3 located at wzw.tum.de/proteomics/nci60)(Gholami *et al.*, 2013), to map the profile to a single gene symbol, rather than translate all the mapped proteins. The majority accession number is the IPI index of the protein, which has the maximum number of identified peptides in the mapped protein group. Since multiple IPI indices often map to the same symbol, the identifiers were rendered non-unique. Therefore, we extracted the IPI index that provided the highest average signal to represent the protein. We redacted from the proteome dataset all IPI indices for which the identified peptides did not map to protein groups according to the supplementary table provided. The transcriptome dataset was treated in a similar fashion.

In order to avoid artifacts arising from a majority of protein abundances registered as zero (0), we require that considered proteins must have non-zero values in more than half of the 57 cell lines. This median-related frequency measure ensured data quality and reduced the downstream impact of outliers. The resulting proteome data contained 2944 proteins across 57 cell lines. The transcriptome repository was pruned to the same size after matching expressed mRNA transcripts to corresponding proteins. As a result, each gene in the resulting transcriptome data contained the protein abundance for the corresponding protein and vice versa.

Supplementary Section 2. Details of extraction and preprocessing of the NCI-60 proteogenomic dataset

Standard methods to capture biological relevance in a synthetic proteogenomic dataset

WGCNA (Weighted Correlation Network Analysis)

The *WGCNA* pipeline constructs correlation networks from the input data and extracts modules from these networks of highly correlated biological elements.

We employed the *WGCNA* pipeline individually on both of our artificial datasets (synthetic mRNA and synthetic protein). We utilized the 1-step automatic network construction and module detection pipeline (outlined at https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-02-networkConstr-auto.pdf)(Langfelder and Horvath, 2008). We utilized the default values for all parameters for the "*blockwisemodules*" construction function of the *WGCNA* R package, except for soft thresholding power and minimum module size admissible. The soft thresholding power was set at size (6) since that was the lowest power that resulted in the maximum scale free topology fit for both the synthetic datasets. The minimum module size admissible was set as two (2), as larger minimum module sizes were rendering the method ineffective in distinguishing any clusters and we wished to identify any and all possible modules that may capture proteogenomic signal. For both of the synthetic mRNA and synthetic protein dataset, the *WGCNA* pipeline results in cluster assignments for the synthetic mRNAs and proteins respectively. We investigated a) whether we were seeing mRNAs or proteins that were part of the same injected relationship group together, b) how many different types of injected signals were we able to isolate and cluster together and lastly c) could we integrate the individual results of employing *WGCNA* on the synthetic mRNA and synthetic protein dataset and unearth the dynamics of the relationships that are contained in these two datasets. Supplementary Section 3 (containing results) highlights the key findings.

iCluster (Integrative Clustering of Multiple Genomic Data Types)

As previously mentioned, the *iCluster* methodology is traditionally utilized to cluster or subtype the samples for which there are multiple types of genomic data. The R package for *iCluster* offered a function to perform integrative clustering. This utility required the datasets to contain different types of measurements for the same set of samples, and the resulting output would provide appropriate clusters for the samples only. Hence, the results from *iCluster* were not conducive to comparison or analysis in the context of the results obtained from PGC, that focuses on unearthing mRNA-protein relationships and the associated regulatory mechanisms.

Supplementary Section 3. Applicability of various standard methods to perform biological exploration on transcriptomic and proteomic datasets.

Results of employing standard exploratory methods on a synthetic proteogenomic dataset

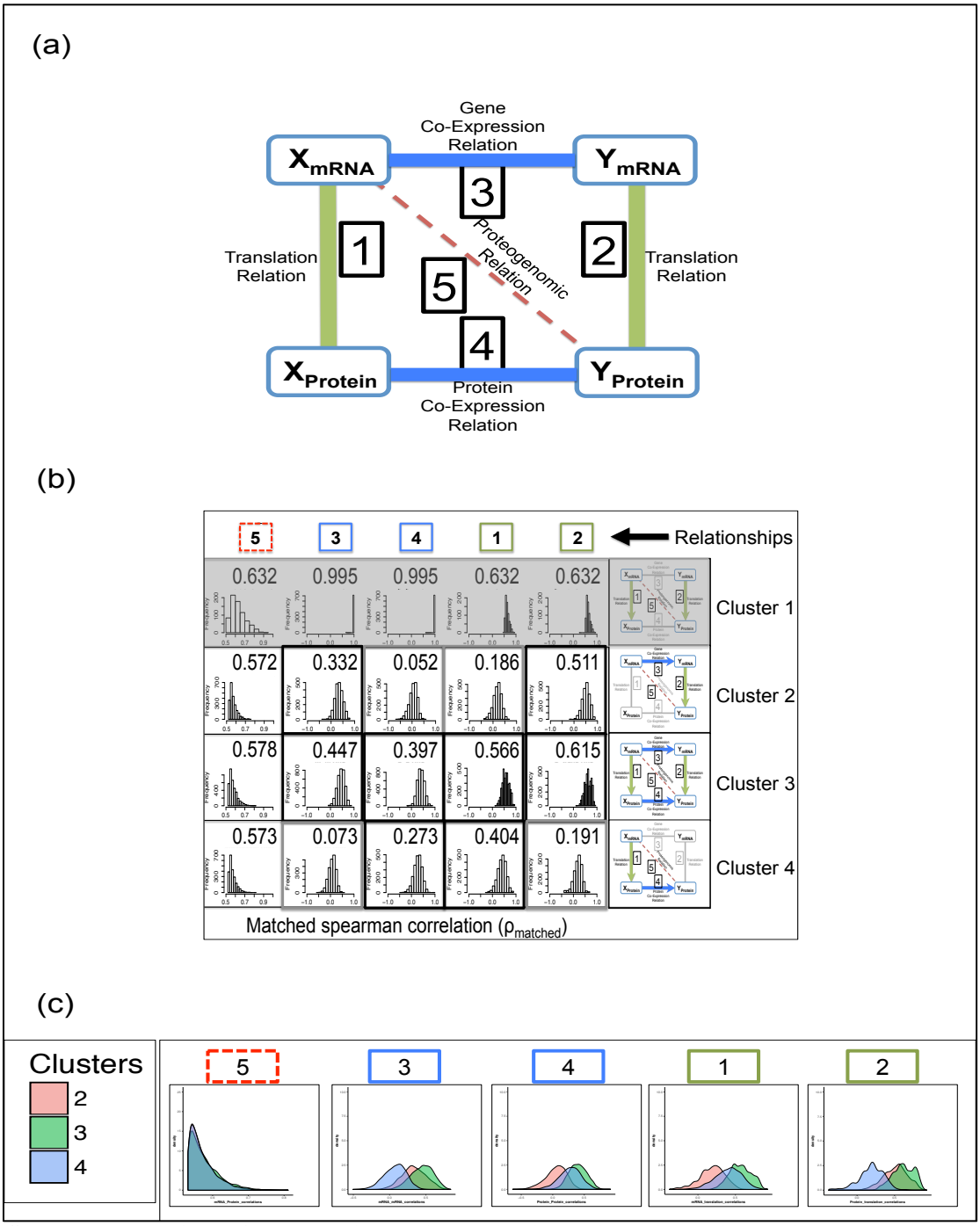
WGCNA (Weighted Correlation Network Analysis)

In order to compare the results of our workflow with traditional methods for extracting clusters with specific proteogenomic regulatory relation-ships, we employed the WGCNA pipeline on the synthetic dataset. WGCNA is a widely used gene co-expression network generation and analysis tool. It constructs correlation networks based on hierarchical clustering and aims to identify densely co-expressed gene modules. We employed WGCNA independently on the synthetic transcriptomics and proteomics dataset. The parameters and details of the pipeline are described above. A majority of elements from either dataset were not clustered (4728 and 3570 elements out of 9000 in each were not grouped into a cluster according to WGCNA), thus we focused our subsequent analyses on the elements that were effectively clustered. We attempted to gauge the utility of WGCNA in the context of identifying and annotating proteogenomic relationships of relevance.

WGCNA results were not comprehensively able to characterize this synthetic dataset in an integrative proteogenomic context. WGCNA was unable to identify all the relevant proteogenomic relationships, and while it unearthed partial pairs of the proteogenomic relationship components, they were not grouped according to any type of embedded biological rationale. Additionally, since WGCNA is a non-integrative methodology, the only way to gauge proteogenomic trends from the results were to integrate the findings from the synthetic mRNA and synthetic protein datasets. Due to the nature of the resulting clusters from either dataset, the overall results are not conducive to being integrated or aiding the discovery of embedded biological relevance. While WGCNA was able to find highly correlated pairs within each dataset, it was unable to group pairs that were part of similar regulatory signaling. Thus not only did WGCNA result in a high number of clusters (995 for synthetic transcriptome and 1521 for synthetic protein), majority of clusters contained only 2-3 mRNA/proteins on average. These clusters also contained unrelated mRNAs and proteins, which diluted the consistency and distinguishing power of the results. **The analysis of the synthetic transcriptome only uncovered (but was unable to group) pairs of mRNA factors originating from signals A, C and D (Figure 1 (b)), since these showcase high mRNA-mRNA correlation. Similarly for the synthetic protein analysis, only pairs of protein factors from signals B and E (high protein-protein correlation) were identified.**
The maximum size for the clusters was fourteen (14) elements and the minimum was two (2). Due to the high number of clusters and subsequent low number of members in each cluster, it was clear that no grouping was effectively performed which clustered the various types of signals together.

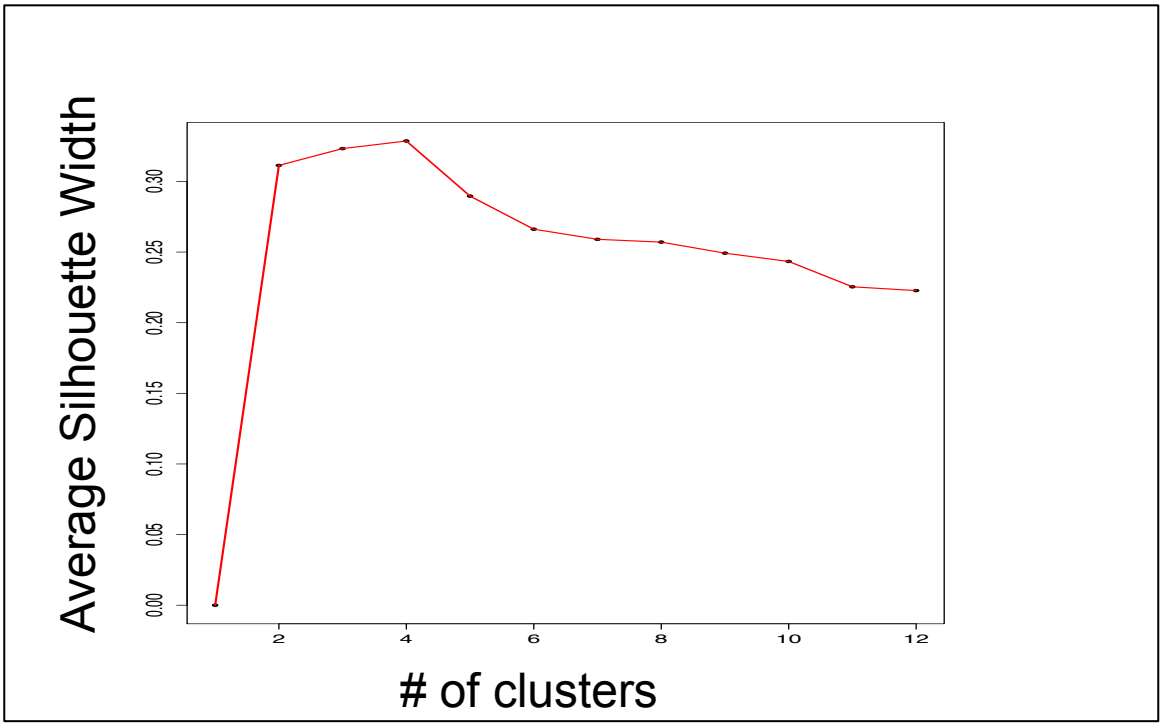
In summary, WGCNA capabilities are limited when trying to isolate relevant proteogenomic relationships due to a) processing a single dimension of the dataset and b) not being able to identify integrative proteogenomic signals effectively.

Supplementary Section 4. Results summation of employing standard analytical methods on transcriptomic and proteomic data to identify non-translational proteogenomic relationships.



Supplementary Figure 1. (a) Potential regulatory circuit (as Figure 1(a)) with numbered edges.

(b) Summary of cluster features - Each row represents a cluster reported by unsupervised clustering of non-translational relationships in the NCI-60 dataset. First five columns represent each of feature edges, the title of each distribution reports its mean. The sixth column provides the schematic from (a), highlighting the edges that expressed high matched Spearman correlation ($\rho_{matched}$). (c) Distribution of each feature across the (non- trivial) proteogenomic relationship clusters (before EM filteration).



Supplementary Figure 2. Choice of number of underlying clusters in the NCI-60 proteogenomic data: The average silhouette width of the clusters is computed as the number of clusters is changed. The average silhouette width is a measurement for how well the data is clustered; the maximum value is indicative of the number of clusters in the data.

Protein	mRNA
EEF1A2	ACAD10
TIMELESS	AKAP1
PAK4	AKT2
ATP1B1	ANLN
UBE3C	ARPC1A
PPM1F	ATP5J2
CHORDC1	ATP5L
S100A16	BANF1
C6orf130	BYSL
PLS1	C10orf32
LOC646214	C11orf10
STIP1	C11orf10
ARFGAP2	C20orf11
TUBB2A	CCDC90A
GCN1L1	CDK5RAP3
DOCK7	CENPF
AHNAK2	CHMP2A
GFM2	CIAO1
VPRBP	CIAO1
SLC25A12	COMMD1
ATP2B4	COPA
CTSA	COPA
LAMB1	COPA
LAMC1	COPA
SLC25A12	COPS6
TMED3	COQ6
AGTRAP	COX7A2
GGCX	COX7A2
ARFGAP2	CS
FASN	CS
CPT2	CYFIP1
CTSZ	DEGS1
TPM2	DEGS1
EEF1A2	EVPL
NT5C2	EVPL
GGCX	EXOSC4
COASY	FAM82A2
HADHB	FBXO22
DPP9	FKBP8
TMX4	FYCO1
WDFY1	FYCO1
EGFR	GABARAP
PCYT2	GINS1
PCYT2	GINS2
SSR1	GMPS
GLMN	HNRNPA1
GCN1L1	HSD17B8
TANC1	HSP90AA1
ACAP2	HSPA6
SOAT1	INPP5F
TIMELESS	L2HGDH
C12orf10	LEMD3
TIMELESS	LEO1
TIMELESS	LETM1
DHPS	LYRM4
SLC29A1	LZTFL1
PAICS	MAVS
KDM5C	MOGS
EFTUD2	MPP2
HDDC3	NDUFA10
TSPAN14	NDUFB4
SLC1A4	NHP2L1
DIAPH1	NLN
HDDC3	NLN
LRBA	NLN
TSTD1	NLN
PPID	NUP54
PDE12	OXA1L
CAD	PABPN1
CHORDC1	PABPN1
FEN1	PABPN1
SETD3	PABPN1
TIMELESS	PABPN1
ERAP1	PFDN1
ECE1	PIGK
KDM5C	POLR2C
MRI1	POLR2G
CORO1B	PPFIA1
SURF4	PPIA
NAGLU	PRKAG1
ITGB4	PROX1
ECHS1	PSEN1
YIPF5	PSMD8
C6orf130	PTCD3
MAT2A	QSOX2
NUDT3	RAD23A
PCYT2	RFC3
GTPBP1	RPL11
FKBP4	RPL35
STMN1	RPL5
PA2G4	RPLP0
HSPA14	RPS12
FKBP4	RPS15A
ARFGAP2	RPS24
FKBP4	RPS24
ISCA2	RPS27
TRAP1	RPS27A
C6orf108	RPS28
FKBP4	RPS3
HMGB2	RPS3
UBE2N	RPS3
ATIC	RPS3A
FKBP4	RPS3A
GLMN	RPS3A
HMGB2	RPS9
MARCKS	RRAGA
CDK6	SF3A3
SLC1A4	SH3BP1
TIMELESS	SLC25A19
PCBD1	SMARCA5
BRP44L	SNRPF
MPDU1	SNX1
NASP	SRP19
C6orf130	SRRT
CAD	SRRT
HMGB3	SRRT
TRRAP	SRRT
YARS	STK4
LSM14A	SYMPK
RASA1	TGOLN2
MTHFD1	TIMM9
PPM1F	TXN2
SLC1A4	TXN2
TIMELESS	U2AF2
TRIM28	UBE2G2
PPM1F	UBE2S
ARFGAP2	UQCRC2
MRI1	UXT
PHF6	UXT
FEN1	WDR5
MAT2A	WDR5
TIMELESS	WDR5
SLC1A4	YTHDF2

Supplementary Table 1. Member relationships of Cluster 2, as a result of employing PGC on NCI-60 data. This cluster showcases mechanistic tendency towards gene co-expression followed by translation.

Protein	mRNA	Protein	mRNA	Protein	mRNA	Protein	mRNA	Protein	mRNA
UBE2R2	ABCB7	ADAR	GM2A	NUDT9	PSME1	GMPR2	DCUN1D5	GRWD1	TERF2IP
NCOR2	ABCE1	TES	GNPDA1	NLE1	PSME2	ZNF638	DCXR	POLD2	THUMPD1
UBE2A	ABCE1	C19orf70	GNS	TMED10	PTGR1	SACS	DHCR24	YRDC	THYN1
RPS6	ACLY	ASPDH	GPX8	IMPDH1	PYCR1	RABEP1	DHX15	MYO1D	TIMM50
RPL28	ACO1	RNGTT	GSS	INPP5F	PYCR1	RPS24	DTD1	NEFM	TM9SF2
USP34	ACYP1	GABARAP	GSTP1	C19orf70	RAB13	GTF3C3	DUT	GABARAP	TMEM214
RABEP1	ADSL	RABEP1	HADHA	NEFM	RAB1B	AKT2	EDC4	IQGAP2	TMEM214
PBXIP1	AGA	ASMTL	HDAC1	BYSL	RAB34	SH3BP1	EHD4	WTAP	TMEM214
RABEP1	AHCY	TTC4	HDAC1	MYO1D	RAB5C	WNK1	EIF2A	RNASEH2C	TMPO
PSMD2	AHNAK	MYOM1	HDLBP	SIN3A	RCC2	TMUB1	ELOVL5	CHMP2B	TMX2
GATAD2B	AIFM1	WTAP	HEXB	TMEM41A	RDH11	PPM1B	ENO2	TMEM147	TOR1A
GTF3C1	AIFM1	GTF3C1	HK2	FRYL	RECQL	SNF8	ENO2	B4GALT5	TPBG
ITCH	AIP	UBE2K	HNRNPF	SLC33A1	RECQL	SPAG7	ENO2	ASPDH	TPM2
ATP6V1F	AK3	ARHGAP5	HSPA13	GTF3C1	RFC4	RPS24	EPB41L2	HMGB1	TRAP1
RRP9	AKT1	HBXIP	HSPB11	RABEP1	RFC4	RPS24	EPB41L3	SMG1	TRRAP
QRSL1	ALDH3A2	PABPN1	HSPB11	USP34	RFC4	DIAPH2	ESD	RPS6KA5	TSPAN14
HSPA8	ANP32B	SLC12A9	ICAM1	ZNF638	RFC4	TSPAN10	ETFA	MAP1LC3B	TUBA1A
FUCA2	APOA1BP	MPV17	KIAA2013	CDK4	SCAMP1	SLC38A1	FASN	GABARAP	UBA1
GABARAP	APOO	POLR1A	KIAA2013	ATL2	SCPEP1	BIRC6	FDPS	DDX5	UBXN1
C11orf10	ARHGAP1	UBE2R2	KIAA2013	MYOM1	SCPEP1	C20orf11	FTSJD2	BIRC6	UCHL5
TMEM41A	ASPH	FLOT1	KPNA3	MYO1D	SDCBP	RPS27	FUBP3	NDUFS2	UPP1
EDEM3	ATG7	MCAM	L1CAM	TMEM41A	SEC23A	COX7A2L	G6PD	MYO18A	USP24
RABEP1	ATIC	BCS1L	LANCL2	GABARAP	SEC61A1	TMEM41A	G6PD	ASMTL	USP7
MYO1D	ATP6V1B2	USP34	LIG3	POLR1A	SELK	SLC38A1	GCDH	UBE2A	USP7
CRTAP	B2M	ZNF638	LIG3	TOMM5	SERPINB6	CBX5	GGCT	WNK1	USP7
MCCC1	BAX	RABEP1	LRPPRC	MYO1D	SFT2D2	LRRC16A	GGCT	RPL28	UTRN
ANKRD28	BRP44L	UBE2R2	LRRC47	B4GALT5	SH3GL1	GTF3C3	GGCX	USP34	VRK1
USP34	C1orf31	NCOR2	LUC7L3	AKT2	SLC29A1	VPS4A	GLMN	C6orf203	WDR1
MRPS23	CAPN2	UCK2	LUC7L3	C9orf86	SLC29A1	RPL28	NPLOC4	WNK1	ZFAND1
RPL27A	CFL2	WNK1	LUC7L3	SMEK2	SLC4A7	RPL19	NUCB2	TMEM41A	PLS3
RPL5	CFL2	KIAA1715	MARCKS	C19orf52	SLC9A3R2	MYO18A	NUDT3	FRYL	POLE
RPL28	CHMP6	NCOR2	MAT2A	NCOR2	SMARCC1	YRDC	NUP210	SLC38A1	POLR2I
RPL26	CHORDC1	GTF3C1	MCM7	AKT2	SMC3	THRAP3	OGDH	RPS6KA5	PON2
C6orf203	CLIC1	MYO1D	ME2	BLOC1S2	SNAP29	RPL26	OXSR1	GABARAP	PPIB
ATL2	CLPTM1	PTCD3	MGST3	PTPN12	SORT1	B4GALT5	P4HA2	TBC1D8B	PPIH
PRCP	CNP	SMNDC1	MINPP1	LCLAT1	SPRYD4	DDX49	PACSIN2	SDHD	PPP4R1
RPS6KA5	CNP	FRYL	MRI1	GCSH	SPTLC2	ASMTL	PAICS	ZNF830	PRDX1
SLC33A1	CNP	ADAR	NAGK	BIRC6	SRM	GABARAPL	PAICS	SLC33A1	PREP
RPS24	CORO1C	DYNLT1	NAGK	GTF3C1	SRM	GTF3C1	PARP1	DNAJB4	PROSC
RPS3	CRK	HSPA8	NAP1L1	BLOC1S2	STRN4	HSDL1	PCYOX1	RABGGTB	PRPSAP2
TMF1	CTSB	KIF15	NAP1L1	VPS13D	STX7	MAP1S	PFAS	GINS1	PSMD9
NIPBL	DARS	FRYL	NASP	CENPF	SUPV3L1	TMEM41A	PFN2	NLE1	PSME1
WDR77	DARS	UFM1	NCBP1	RPL28	SWAP70	C17orf49	PGM2		
GTF3C3	DCPS	CHMP2A	NCL	RNF181	TELO2	RTF1	PKN1		

Supplementary Table 2. Member relationships of Cluster 4, as a result of employing PGC on NCI-60 data. This cluster showcases mechanistic tendency towards translation followed by protein co-expression.

References

Gholami,A.M. *et al*. (2013) Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Rep*., **4**, 609–620.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Ripley,R. *et al*. (2011) MASS: support functions and datasets for Venables and Ripley's MASS. *R Packag. version*, 170.

Shankavaram,U.T. *et al*. (2009) CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*, **10**, 277.