**Intro**

**Motivation**

Here, I present a plotted database on proteomics in CCLs. Some time ago I plotted a number of proteins of interest from the database. Since coding for plots is quite time-consuming (unlike statistics which in R is mostly one liners), I decided to automate the process for any protein in the database. So, it is rather unintended but I do hope that you will find it useful.

Public databases are often a black box. Anyway, we have the right to know what is in them (even if there is not as much as we hope for). For every protein there, I provide us with a booklet of plots that contains information that can be extracted from the database. As of now, I have covered mitosis, DDR, immunity, and membrane-related proteins, around 3000 proteins altogether. I am depositing them on the Helleday server. I do hope that you find it useful.

**Publication**

I used the database from this publication of 2022. It is produced with a high-throughput method using Data Independent Acquisition (DIA)-MS with a workflow that enables high throughput and minimal instrument downtime – what should be close to clinical conditions, with spectral library using DIA-NN. The database covers almost one thousand CCLs across 28 tissues. The total number of proteins is 8500, but more than one supporting peptide was found for ca. 6700 proteins. They are presented in the paper as two individual databases. For reference, I call them DB1 for the smaller one and DB2 for the bigger one. I mainly focused on DB1 due to fewer false positive but if needed, I can generate similar plots from DB2 for more insights.

**Descriptives**

**Names**

I use the gene symbol throughout all the plots, even if the protein may have a different name, adding the protein name, gene name and Uniprot ID. Also, information from Cansar is included, indicating whether the protein is an enzyme, ligandable, or chemically ligandable. If not in the Cansar database, all values are marked as NA.

**Histograms and density plots**

I show here histograms and density plots, comparing quantity of a protein of interest against two common proteins: GAPDH (present in 100% of CCLs, with the second-highest median quantity after an H4 histone) and ACBD3 protein (identified in approximately 98% of CCLs but in small quantities). They are consistently positioned on the x-axis for easy visual comparison.

Each protein quantity is percentile ranked relative to other protein quantities in the cell to find the range, median and mean. For instance, for **DHFR** the median percentile is around 40, while the minimum is around 1, and the maximum is around 90. If you need to know in which CCLs a protein of interest is found in the smallest or largest quantities tell me and I will provide you with a list.

DHFR was detected in 55% of CCLs. When a protein is present in not all CCLs, it indicates that, under equal conditions, the protein was only found in a subset of CCLs while remaining undetected in the others. The observed variations could stem from quantities too low for detection rather than complete absence. This is also true for other proteomic databases, not all proteins are found in all CCLs or samples. Naturally, proteins with higher abundance are more reliably detected. Less detected proteins may be required in very small quantities or could be inducible under specific conditions.

Important to know, the median Pearson correlation between gene expression and protein amount is 0.42 for the database.

**SHMT2** is found in all CCLs, exhibits a high quantity with a relatively narrow distribution. It means that there is only minimal variation when whatever groups are compared at the baseline at which the measurements were taken.

**BRCA2** has a broad distribution, and is identified in 85% of CCLs, albeit with only one supporting peptide, therefore in DB2.

**ITGA3**, a cell surface adhesion molecule, has a broad distribution. It frequently showed up as an antagonist to mitotic proteins in the analysis of differential protein expression. A subset of proteins consistently displays this behavior across the entire database.

**Violin plots and barplots**

Violin plots and barplots show the distribution of amount and occurrences of proteins across tissues, with a statistical analysis of those. ANOVA assesses the statistical difference in protein quantity across tissues. Chi-square shows whether there are statistical differences in the number of CCLs where the protein is identified across tissues. (ANOVA compares protein quantity, chi square compares the number of protein occurrences)

**KIF2A** is present in almost all tissues but it is underrepresented in the head and neck cancers, and the chi square result is significant. Its amount is notably greater in blood and the peripheral nervous system, with a highly significant ANOVA result. Most often it is obvious from the plots which tissues are different, but if you ever need p values for individual differences, let me know and I will run a posthoc analysis.

**KIF1A** is identified in few tissues, with a huge prevalence in lung cancers hence the highly significant ANOVA result.

**PCNA** protein is identified in all tissues therefore the chi-square is not attributed. Here, ANOVA revealed significant differences, with notably higher levels in blood cancers compared to other tissues.

**TOMM70**, a mitochondrial membrane protein, is also omnipresent but in different quantities. It exhibited the highest levels in breast, large intestine, and pancreas cancer cell lines, although I have not found anything in the literature on why it is so or much on its role in general.

Cell lines from blood, peripheral nervous system, and skin cancers show the greatest numbers of distinctions in protein occurrences.

**Mutations and kinases**

**Mutations**

For mutation plots, I used two DepMap mutation databases – a database of all mutations on the DNA level and database of mutations that are deemed damaging. The plots here serve as a reference, as they only show those mutations for which proteins are present, as in for instance in **BRCA2** or **TP53**. For damaging mutations, I include the names of CCLs where they are present. It’s important to note that not all mutations with a functional role are categorized as damaging, as the database designates damaging mutations as those found in binding or active sites.

The mutation plots can also show tissue preference. **PDS5B** which interacts with the cohesin complex, exhibits mutations in blood cancers, the large intestine, and endometrium but not in the lung or skin though it is nicely represented there.

Tankyrase 1 binding protein **TNKS1BP1**, which is involved in DNA damage response, has numerous mutations, but literature on its mutations is scarce and emphasizes overexpression instead. Could its high molecular weight be concealing its role in cancer?

Some mutated proteins are not outstanding at the DNA level in the mutation database but are noticeable on the protein level. A hepatoma-derived growth factor **HDGFL2** binds to methylated histones. There is no information in the literature about its mutations in cancer. For comparison, **KAT7**, a histone acetyltransferase with a similar MW, shows no mutations. **THOC1**, involved in RNA export, displays numerous mutations. For comparison, **THOC5**, a component of the same RNA export subcomplex, shows no mutations.

There might be additional context to consider, like hypermutation or gene replication timing, or a need for a broader database analysis. However, this is what stands out when a mutation occurs, and the protein is present.

**Kinases**

The kinase database comes from the publication of 2022. It includes data on affinities of 303 Ser/Thr kinases to almost 90000 phosphosites. The database was generated with a combinatorial peptide library that systematically substitutes each amino acid at nine positions surrounding a central phosphoacceptor.

The top plot shows top 10 kinases for each phosphosite. The bottom plot shows only the kinases with the score greater than 98.5% which means that the kinase scores better than 98.5% of all profiled kinases in the database for that phosphosite. For **CASP2** the top scoring kinase is IKBKE which regulates inflammatory responses.

**ABCB1** translocates drugs and phospholipids across the membrane, and is obviously strongly regulated by phosphorylation, with two prominent phosphosites. The top kinases are phosphorylase kinase and protein kinase C both of which are calcium dependent.

Personally, I did not expect that there would be so many kinases with high scores for each phosphosite. If there are a number of kinases for the same phosphosite it may mean that other kinase-determining factors may be in play: scaffolds, localization, coexpression. Or may indicate that these specific phosphorylation sites are points of convergence for multiple signaling pathways.

The greatest factor in affinity regulation is the negative selection of substrates through electrostatic filtering, hydrophobic residues – to avoid phosphorylation by irrelevant kinases.

A **TP53** kinase plot shows lots of regulations and interactions. If you are interested in the top output and it is one of those proteins where it is jumbled, tell me and I will make a plot with a different threshold.

For **MTHFD1**, the top scoring kinase is DAPK2, a positive regulator of programmed cell death.

For **MTHFD1L** the top scoring kinase is p38, aka MAPK14. If it is true and they have possibilities to interact, MTHFD1L can be regulated by MAP kinase signal transduction pathway.

For **MTHFD2** there is no kinase that scores over 98.5%.

**Gene set enrichment analysis on differentially expressed proteins**

I performed an analysis for differentially expressed proteins: between each protein being absent/low and being high. Then I performed a gene set enrichment analysis on the output to identify biological pathways enriched with differentially expressed proteins. It helps understand which proteins work together in a coordinated manner.

The plots show pathways that are downregulated when the protein is absent or low. This happens when the protein actively participates in these pathways or where the identified pathways go together with other processes involving the protein.

When **CENPC** is absent or low, downregulated pathways are chromosome, centromeric region, chromatin and DNA binding; and also RNA splicing and response to DNA damage. Upregulated are cell adhesion and cadherin binding. Note that this makeup is common for many mitotic/DNA damage proteins.

When **RAB5B** is absent or low, the downregulated pathways are transmembrane transporter activity, calcium ion binding, endoplasmic membrane; upregulated pathways are histone binding, chromosome organization, mitotic cell cycle process. This makeup is characteristic for membrane-related proteins and basically reverse to CENPC or other proteins participating in progressing of mitotic cycle.

**GPR180** is a receptor effecting brown adipocyte function and glucose homeostasis. When it is absent or low, there is downregulation in glycoprotein and lipid metabolic processes, glycosylation, and hexosyltransferase activity. This suggests its involvement in or dependence on glycosylation for its function, though I have not found any mention in the literature on this. When GPR180 is absent or low, the upregulated pathways are DNA repair and binding, kinase activity, ubiquitination, and translation.

When **MTHFD2** is absent or low, the downregulated pathways are mitochondrial membrane, methylation, transferase activity for one-carbon groups, as expected. But also among the downregulated pathways are nuclear chromosome as it is documented that MTHFD2 is a nuclear protein (paper here), RNA processing, also documented (paper here), and also histone and chromatin binding, double strand DNA binding, not documented yet but it may be probable it can do that too. Upregulated processes are cell adhesion, GTPase regulator activity, and actin binding.

When **MTHFD1** is absent or low, downregulated pathways are ubiquitination, kinase activity, GTPase regulator activity (which is upregulated for MTHFD2), mitotic cell cycle, and tubulin binding. Unlike the other MTHFD proteins it may be regulated by ubiquitination. Upregulated pathways are mitochondrial inner membrane and cell adhesion. It is probable that the plots give information on whether a protein of interest is regulated by ubiquitination or phosphorylation or both or none. The downside is that if ubiquitinated, the ubiquitination pathways dominate the output.

**TNIP1** is involved in autoimmunity by regulating nuclear factor kappa-B activation. Downregulated processes are regulation of immune response, antigen binding, etc. But also DNA repair, response to DNA damage stimulus, and mitotic cell cycle. It is not known whether it has a direct role in those, I have not found anything in literature.

**STING1** is a transmembrane protein that regulates the innate immune response to infections. Downregulated pathways are defense response, antigen processing and presentation, etc. But also, hydrolase activity acting on O-glycosyl bonds is downregulated, I have not found any literature on why and how STING1 can be involved in that.

**CGAS** is a key sensor for foreign DNA, triggers immune responses upon detecting cytoplasmic double-stranded DNA. Downregulated processes include signaling receptor activity, defense response to other organisms, and adaptive immune response. But also there is quite a number of downregulated pathways associated with rRNA binding and processing. It suggests a potential direct involvement though I have not found any literature on that.

**SOD1** is an antioxidant. However, when it is absent/low, downregulated processes are microtubule binding, tubulin binding, spindle, kinetochore, and mitotic cell cycle. In this case, the interaction with tubulin is documented and is known to be impaired in amyotrophic lateral sclerosis when SOD1 is mutant. So, this particular association is known, but if it were not known yet, it anyway shows here in the output of protein interactions.

Limitations include the need to check p-values, I included those for the topmost and the lowest pathway bars. Also check for the number of samples where the protein is found, the results may be not exactly accurate if there are very few samples in the database. Also, the algorithm tends to favor pathways with a larger number of components, potentially biasing results. If you feel that there might be something else interesting beyond what is plotted, tell me and I will generate an excel file where I will list all the pathways.

I also plotted pathway analyses on correlation coefficients, in two flavors: overrepresentation and by correlation values. Those are broader.

**Networks and cooccurrence**

**Networks**

In the network graphs, the focus is on the networks of proteins rather than isolated correlations. They show a web of connections, identifying whether a protein functions as a hub. Some of these connections form true protein complexes but mostly they indicate highly connected and co-regulated protein networks. I include separate plots on individual correlations.

**CAD** participates in de novo synthesis of pyrimidine nucleotides. It forms a very highly correlated network with PAICS, GART, ATIC, but also with components of the CCT chaperonin complex. The plots show networks where all associations are stronger than the one indicated over the plot. The plotting starts with the topmost identified correlations and decreases in steps of 0.5 between plots. Moving down a step, MTHFD1 is also involved in the network.

**PSAT1** is phosphoserine aminotransferase, it forms a network with alanyl, glycyl, isoleucyl, and tyrosyl tRNA synthetases. The next step adds more proteins involved in translation, as well as purine synthesis participants.

**RAD23A** is a nucleotide excision repair protein, but its strongest interactions are with prefoldin components, proteasome genes, and unexpectedly with PARK7 (oxidative stress sensor) and myotropin.

C**ooccurrence**

The analysis is based on the Bayes’ theorem which estimates the likelihood of an event by incorporating prior beliefs. Here, we estimate a theoretical probability of finding these two proteins together from their frequencies in the overall data distribution. Then we compare the theoretical probability with the actual protein distribution, i.e. whether there is more or fewer occurrences than the predicted values; from these we calculate p values and odds ratios.

**KNSTRN** (kinetochore localized astrin (SPAG5) binding protein) occurs together with regulator of checkpoint kinase 1 CLSPN, centrin3 CETN3, centriolar assembly protein SASS6, regulator of mitotic spindles SPAG5, BUB1B, PARP2 etc. KNSTRN is identified in 325 CCLs, while CLSPN is identified in 20 CCLs. The expected cooccurrence is circa 7 CCLs, but the observed one is 18 – almost in all CCLs where CLSPN is found.

The data is sorted by the odds ratio, the highest odds do not mean the exactly smallest p value, but rather one of the smallest. It can happen that there is a long tail of odds more than 1 as in here, and they are not shown in a plot. If you are interested in those beyond what is plotted, tell me and I will give you a full list of those.

Negative cooccurrence is when proteins are encountered together more seldom than expected. Here there is a thioredoxin reductase, RAB1B, RAB2A.

The cooccurrence analysis comes in three flavors: for all the data, and separately for blood and solid cancers, as some proteins prevail in solid cancers, and some in blood cancers.

**ORC1**. Highest odds are actin related protein 5 ACTR5 involved in double-strand break repair, DNA clamp unloader ATAD5, single stranded DNA binding protein 3 SSBP3, another origin recognition complex subunit ORC3, components of HAUS augmin-like protein complex, which takes part in cytokinesis and mitosis.

Absolute numbers of proteins may be not high, e.g. two more observed than expected, but it gives an idea of possible interactions.

**PARP9** demonstrates cooccurrence with interferon induced proteins (IFI44, IFI27, IFIH1), interferon regulatory factor, toll like receptor. A recent study identified PARP9 as a noncanonical sensor for RNA virus (paper here). This particular association is known, but if it were not known yet, it anyway shows here in the output of protein interactions.

The cooccurrence analysis is complementary to correlations and network analysis, and works for those proteins that are not identified in all CCLs. Proteins are sometimes found together because they are in the same tissue, sometimes in a same process, or at a distance of two or one handshakes, or sometimes they truly interact. It is not known solely from cooccurrence, but if proteins are found together, they might collaborate in some way.

**DEGS tables and volcano plots**

The next section shows differential expression of proteins when a protein in question is either absent/low or high. You have already seen the pathway analysis part of it. It also comes in three flavors: for all cancers, blood cancers, solid cancers.

When **AURKA** is low or absent, CCNB1, CCNA2, SPAG5 etc. are downregulated: the output is dominated by mitotic proteins. For **AURKB** it is similar, but the downregulated part is also enriched in kinesins. Those that are upregulated are on the right. These are keratins, galectins, plakins, Ca binding proteins: a diverse set of membrane proteins. The makeup is common for most mitotic proteins.

**SMC5** is in SMC5-SMC6 complex which is involved in repair of DNA double-strand breaks by homologous recombination; NSMCE components (also components of SMC5-SMC6 complex) go down with it, and other cell cycle proteins. Upregulated are several S100 proteins which may be biomarkers for low activity of the SMC5-SMC6 complex.

Limitations – check for total number of occurrences of a protein in question. Results are more reliable with more occurrences.

With **TIPIN** (control of DNA replication and the maintenance of replication fork stability) in solid cancers the proteins that go down are TIMELESS, DNA polymerase alpha (POLA1, POLA2), DNA primase PRIM1; but also – TUBB2B, KIF1A. Others may be less expected: TLE5 (a transcriptional corepressor), ACYP1.

The previous tables are sorted based on logfold change. However, logfold changes are not always big, as many proteins have a very narrow distribution range which you have already seen. Also, an individual protein may not be the most influential player on its own. Instead, its significance often is in its collaborative interaction and exhibits increases or decreases together with other proteins in a concerted manner. To explore those high veracity changes, I also made for same plots for both blood and solid cancers tables that are sorted by p values.

**SHMT1** (and volcano plot) – only here we see that other proteins involved in one carbon metabolism AHCY (S-adenosylhomocysteinase) and ADSS2 (adenylosuccinate synthetase) go down with it highly consistently. Also decreases are ADI1 (involved in methionine salvage), PCYT2 (synthesis of phosphatidylethanolamine). These may show its importance in phospholipid metabolism.

**DHFR** goes down with PCNA, RAD23A, POLE4, LIG1; but also with nucleotide kinases UCK2, DCK, AK6.

For **MUC1**, which is a membrane protein and not released in mucus, a number of cell adhesion related proteins go down with it with a very high significance, and also with a big logfold change. With this significance, it is a very distinctive makeup of its cooperation with other proteins.

I can make similar analysis and plots for gene expression or for CRISPR gene effect.

**Conclusion**

In a conclusion, more often than not, data wants to tell us something even what it tells is often far from straightforward.

Also, this isn’t big data; it is more medium-sized and imbalanced, as there are nine times more proteins than samples. The quality and size of available databases does not match the tools that we have, much more can be done with bigger databases. In this case, while AI/NN/DL can be applied it is more reliable to do that with more data. Sadly, there are lots of small databases but very few sufficiently big ones.

Anyway, even this medium sized database tells us a fair amount of information, and we have the right to know it.

One question: do you want them all? Just tell me so and I can plot all the proteins or all the proteins from the both databases, too.

#If yes – I will put them all together in a single huge folder, they will be there in an alphabetical order. Also, I can write code for plotting CRISPR or gene expression databases in a similar fashion.

*GAPDH (glyceraldehyde-3-phosphate dehydrogenase)*

*ACBD3 protein (acyl-coenzyme A binding domain containing 3)*

*PHGDH phosphoglycerate dehydrogenase*

*PCNA Proliferating cell nuclear antigen*

*chi-square (kaiskwer)*

*TNKS1BP1 tankyrase 1 binding protein 1, a member of the poly(ADP‐ribose) polymerase (PARP) superfamily*

*HDGFL2 672 aa – This gene encodes a member of the hepatoma-derived growth factor (HDGF) family. The protein binds to methyl-lysine-containing histones, H3 and H4.*

*A combinatorial peptide library that systematically substitutes each of 22 amino acids (20 natural amino acids plus phosphorylated Thr and phosphorylated Tyr) at nine positions surrounding a central phosphoacceptor position containing equivalent amounts of Ser and Thr.*

*Purified recombinant kinase preparations. Substrate recognition motifs were determined through a positional scanning peptide array. In summary – a motif-based approach to study signaling.*

*IKBKE (Inhibitor Of Nuclear Factor Kappa B Kinase Subunit Epsilon)*

*ABCB1 (ATP binding cassette subfamily B member 1)*

*(generally applicable gene set enrichment)*

**Correlations, GSEA and overrepresentation on correlation coefficients**

There are also correlations per se, and two more flavors of pathway analysis.

**Correlations**

**POLD1** ubiquitous, correlates strongly to fatty acid synthase, why? Then with a smaller strength to ATP citrate lyase, adenosylhomocysteinase, and phosphatidylethanolamine binding protein 1.

**POLDIP2** – correlated to mitochondrial ribosomes – why?

Limitations – when there is a small number of paired observations; I removed those that have fewer than 100 pairs, so everything in the plots has more than 100 paired obs

**Overrepresentation**

pathway analysis on overrepresentation of correlation coefficients – i.e if the most of correlated prots belong to the same pathway. Only those with more than 100 paired obs.

**TRAFD1** – negative feedback regulator that controls excessive innate immune response. SH3 domain binding – proteins that interact with other proteins; proteasome complex, negative regulation of immune response, antigen processing and presentation; and also – microtubule cytoskeleton involved in mitosis, kinase regulator activity. Of all pathways – most broad, accounts for mostly the processes that go together with the one where the prot participates.

**GSEA**

pathway analysis on correlation coefficients, taking the coefficient size/value into account

**DYNC1LI2**. Not much is known. Here it is associated with regulation of actin filament, cytoskeletal protein binding, microtubule motor activity. But also – with phospholipid and phosphatidylinositol binding, clathrin coat, endosome transport, and proteasome.

*GAPDH (glyceraldehyde-3-phosphate dehydrogenase)*

*ACBD3 protein (acyl-coenzyme A binding domain containing 3)*

*PHGDH phosphoglycerate dehydrogenase*

*PCNA Proliferating cell nuclear antigen*

*chi-square (kaiskwer)*

*TNKS1BP1 tankyrase 1 binding protein 1, a member of the poly(ADP‐ribose) polymerase (PARP) superfamily*

*HDGFL2 672 aa – This gene encodes a member of the hepatoma-derived growth factor (HDGF) family. The protein binds to methyl-lysine-containing histones, H3 and H4.*

*A combinatorial peptide library that systematically substitutes each of 22 amino acids (20 natural amino acids plus phosphorylated Thr and phosphorylated Tyr) at nine positions surrounding a central phosphoacceptor position containing equivalent amounts of Ser and Thr.*

*Purified recombinant kinase preparations. Substrate recognition motifs were determined through a positional scanning peptide array. In summary – a motif-based approach to study signaling.*

*IKBKE (Inhibitor Of Nuclear Factor Kappa B Kinase Subunit Epsilon)*

*ABCB1 (ATP binding cassette subfamily B member 1)*

*(generally applicable gene set enrichment)*

**Histogram extras**

MTHFD1 is present in a substantial quantity with a broad distribution, and observed in 100% of CCLs.

MAD2L2, also only one supporting peptide, is present in a limited quantity with a broad distribution, and is found in 15% CCLs.

CAV1, a main component of caveolae plasma membranes, exhibits similar characteristics. Both show high variability between CCLs, as seen from the broad distribution.

PHGDH, responsible for early steps in L-serine synthesis, also demonstrates a broad distribution. However, it infrequently experiences significant decreases or increases in the analysis of differential protein expression.

**Violin plots extras**

**TP63** is found in few tissues and exhibits an insignificant ANOVA result. The chi square is however highly significant, with prevalence in the head and neck as well as in the esophagus cancers.

The total analyses shows that “Cell lines from hematopoietic and lymphoid, peripheral nervous system, and skin cell types show the greatest numbers of distinctions”.

For **BAD**, both protein concentrations and protein presence are higher in cervix, thyroid and prostate cancers (both ANOVA and chi square are significant) while FXR2 concentrations are high in bone, peripheral nervous system, and prostate.

For **BCL2** (from DB2), ANOVA yielded an insignificant result. It is significantly more often found in blood cancers which is seen from the very significant chi square.

AURKA is present in few tissues, while AURKB is much more prevalent.

ADPRS – both conc and presence low in head and neck

CDK2 – high in skin

FXR1 – high in esophagus

**Mutations extras**

The analysis allows for a visual assessment of a protein’s tolerance to mutations or being a cancer driver. For instance, **BUB3** which functions in spindle checkpoint signaling without any observed mutations, and the antiapoptotic protein **BAX** which is a cancer driver.

The mutation plots can also show tissue preference. **NUP98**, a nucleoporin, shows overrepresentation of mutations in the large intestine and skin.

**GAGE extras**

**KDM1A** is a component of several histone deacetylase complexes, though it silences genes by functioning as a histone demethylase. When it is absent or low, the downregulated pathways are histone modification, N-acetyltransferase activity, methyltransferase activity, and also mitotic cell cycle, microtubule binding. Upregulated pathways are cell surface, cell adhesion, calcium ion binding.

**SLC45A2** is a transporter protein that mediates melanin synthesis. Downregulated are transporter activity, lipid metabolism, pigmentation. Upregulated are DNA repair, DNA replication, kinase binding. (*upregulated – when a protein in question is low, the cell is busy with these*)

**MTHFD1L** is also located in mitochondria as is obvious from the downregulated pathways. It is also associated with RNA processing, though it is not clear whether directly or not, I have not found any literature on that. Unlike those for MTHFD2, downregulated processes are also cell junction and transmembrane receptor proteins. Upregulated pathways are GTPase regulator activity, ubiquitination, microtubule binding, and kinase activity.

**Network extras**

**EGFR** 65 is associated with COL17A1 (type XVII collagen which unlike most collagens is a transmembrane protein); then with CDH3 (cadherin3, calcium-dependent cell-cell adhesion protein); then with keratins, laminins and plectin.

**USP7** 100 (ubiquitin specific peptidase 7) – with BUB3, RCC2, GMPS (guanine monophosphate synthase), GSPT1 (G1 to S phase transition 1). It may indicate deubiquitination targets of this protein.

**FAF1** participates in NFKB signaling where it serves as a ubiquitin-binding protein. It mediates programmed cell death. Also, it takes part in the progression of DNA replication forks. However, unlike other proteins in NFKB signaling, the network suggests its substantial involvement in mitotic processes and in translation. Both FAF1 and RAD23A are then highly involved in protein networks with association strengths between 0.5 and 0.55.

**Cooccurrence extras**

**TNFRSF8** (TNF receptor superfamily member 8). It prevails in blood cancers. It goes together with ADPRH (removal of mono-ADP-ribose attached to arginine residues on proteins) which is a prognosis-related biomarker in glioma, but not much more is known about the protein. Then there are CD300A (a member of the CD300 glycoprotein family of cell surface proteins found on leukocytes involved in immune response signaling pathways); FCER2 (a B-cell specific antigen); SLAMF1 (signaling lymphocytic activation molecule family member 1); IRF5 (interferon regulatory factor 5); NCF2 (neutrophil cytosolic factor 2); RUNX3 – a very specific transcription factor; ENTPD1 – a plasma membrane protein that hydrolyzes extracellular ATP and ADP to AMP, little is known about the protein, its inhibition may confer anticancer benefits. All of them also prevail in blood cancers. Negative cooccurrence is biased for solid cancers.

**POLA1** is found in roughly half of solid cancers. The part with greatest odds is enriched for DDR components: RMI2 (homologous recombination-dependent DNA repair), MND1 and PSMC3IP (stimulates RAD51-mediated homologous strand assimilation), RNF138 (double-strand break repair via homologous recombination), SSBP3. These are interspersed with proteins involved in chromatin modification (SMARCAD1, KDM4B); RNA modification (CELF4, ELAVL4); and cell cycle progression (HAUS, CHEK2).

**RAD18** is involved in postreplication repair of UV-damaged DNA. Encountered together with MIS18A kinetochore protein A, CHEK1, BUB1B, SKA3, CEP55 (centrosomal protein 55), CDC20, cell division cycle associated 7 CDCA7. But also: CUEDC2 (negative regulation of cytokine production involved in inflammatory response), PALD1 (protein tyrosine phosphatase), and very little is known about these proteins. Negative cooccurrence is with CTSS (cathepsin S, degradation of antigenic proteins).

**Volcano plot extras**

**FAS** is a member of the TNF-receptor superfamily and plays a role in programmed cell death. This is in blood, when FAS is absent or low, the other proteins absent or low with it are HLA components, Il4I1, IRF, TRAF1, CD molecules.

**IL27RA** (DB2, interleukin 27 receptor subunit alpha) in solid cancers. Most prominent interactions are with TPD52 (tumor protein D52) – a very little studied protein; MIEN1 (migration and invasion enhancer 1) – also a very little studied protein; with HOOK1 (promotes vesicle trafficking), other immune related proteins (grancalcin, platelet activating factor).

When **CDCA5** (cell division cycle associated 5) is low or absent, cyclins, AURKB, SKA3 are low, too; the downregulated proteins are in general related to centrosome, spindle, kinetochore; there are several other less known mitotic proteins and other less known proteins like huntingtin interacting protein 1 related HIP1R or MAZ that are probably worth to find out what they are doing together.

Epstein-Barr virus induced 3 **EBI3** forms interleukin 27 and regulates T cell and inflammatory responses; downregulated with it in blood are TNF receptor associated proteins, FAS, IL4I1. Among the upregulated are cyclins and CKAP2.

**CD74** regulates antigen presentation for immune response; downregulated with it in blood are HLA components, other CD molecules, IRF4, NFKB2.

**TFAP4** (transcription factor AP-4). Those that are downregulated with it are BRD2 (associates with transcription complexes and with acetylated chromatin during mitosis), CDCA5, UNG. But also other transcription factors: TCF12, POU2F1, NFYB, NRF1; several chromatin remodeling factors. It may be a possible interplay or cooperation.

**RAB5A** goes down with a number of other RAB proteins, with TMED10 and TMED9 traffickers, ribophorins and clathrin heavy chain. In the upregulated there are many mitosis related proteins. These changes are small but highly significant.

**KIF2C** goes down with DLGAP5 which participates in centrosome localization; kinetochore assembly; and mitotic spindle organization, a number of other kinesins, condensin complex components, DNA polymerases, also with ubiquitin conjugating enzymes with high consistency.

Phone no of Pradeep is 0703668097. Call him and say to him that you are colleague from Sanjiv and looking for an apartment.

regards