

# Selected initial steps for global analysis of proteomic data

John P. Wilson  
Written as a Postdoc in the  
Darryl J. Pappin Laboratory of Protein Analysis and Mass Spectrometry  
Cold Spring Harbor Laboratory  
john@protifi.com

This document describes one of many possible initial approaches to glean some insight into the “meaning” of protein lists generated in proteomic studies. The tools highlighted are general and are applicable to protein lists from genomic studies, those of published literature, etc. *These steps are by no means comprehensive* and have been selected because of their usefulness, usability and availability (free). An extensive list of tools for such analyses is maintained at [www.pathguide.org](http://www.pathguide.org). Recommended URLs: [www.pathwaycommons.org](http://www.pathwaycommons.org); [cancer.cellmap.org/cellmap/cytoscape.do](http://cancer.cellmap.org/cellmap/cytoscape.do); [www.cytoscape.org](http://www.cytoscape.org); [locate.imb.uq.edu.au/](http://locate.imb.uq.edu.au/). Highly recommended reading: [www.g-sin.com/pipelines/resource/3326\\_Bader\\_slides\\_pathway\\_data\\_\\_analysis.pdf](http://www.g-sin.com/pipelines/resource/3326_Bader_slides_pathway_data__analysis.pdf) (by Gary Bader and Quaid Morris at the University of Toronto); [www.slideshare.net/micheldumontier/network-biology-from-lists-to-underpinnings-of-molecular-behaviour](http://www.slideshare.net/micheldumontier/network-biology-from-lists-to-underpinnings-of-molecular-behaviour) (by Michel Dumontier at Carleton University). Email if you’ve pressing questions or should you spot an error (or better way to do something) below. Especially let me know if you find a particularly useful tool!

## Data analysis steps

- 1) Search data against necessary databases. Required IDs – and thus databases – depend on the selected tools. Typically, Uniprot (Swissprot/TREMBL), formerly IPI accession number, gene symbol, ENSEMBL IDs, etc. are inputs.

We will use a Uniprot database.

**Note 1:** while ID mapping services (accession number convertors) are available, they are not error free and will often return missing values or multiple assignments for the same (single initial) accession number. For statistical analysis, multiple mappings are especially problematic.

**Note 2:** protein accession numbers change with version. If possible, download the database of the tool you will be using and search against it.

- 2) Filter/select data to analyze (e.g. positives not found in negatives, fold change up or down, proteins found in group 1 but not 2, etc.). In the case below, we will look at all proteins two or more standard distributions up or down regulated based on iTRAQ values. Export the selected accession numbers to a text file.
- 3) Go to <http://david.abcc.ncifcrf.gov/> (DAVID). Click “Start analysis” in the upper left corner. Input or upload the list from step 2 and select the identifier (Uniprot e.g.) as well as “gene list.” Click “Submit list.”

- 4) Use the tools to look for trends such as enriched GO terms and pathways. Pathways can be visualized by expanding the pathways selection and clicking on the “chart” button. On the table which then appears, click a pathway under the “term” header. Look for similarities in pathways, subcellular location, catalytic function etc; these insights will be useful later.
  - 5) If your analysis program does not recognize the database accession numbers you have, go to <http://www.ebi.ac.uk/Tools/picr/> and translate. E.g. IPI lists can be translated into Swissprot/TREMBL accession numbers. Export such a list as an Excel file. For Uniprot to be recognized by String, open Excel and generate a list of accession numbers without the isoform extensions (-1, -2, etc.) by using the command “LEFT(cell,6)” where “cell” is the cell with the translated Uniprot accession numbers. Fill this down and copy the full list of converted, truncated accession numbers with synonyms. Copy this and save it as a new file.
  - 6) Go to <http://string-db.org/> and input the list under multiple names. Select the appropriate species and click “Go!”
  - 7) Verify that String has correctly converted accession numbers into gene symbols and click continue.
  - 8) Click “confidence” for the view and then advanced. Adjust the required confidence score to establish a reasonable number of networks. Usually between 0.5 – 0.8 is good; recommended start value is 0.65. You may want to take screen shots of each network to determine the correct value by later visual inspection. Note that the network exhibits switch points and at some point, very fine adjustments (e.g.  $\pm 0.001$ ) will result in a completely redrawn network.
  - 9) After finding a suitable confidence value, move unconnected gene symbols out of the network. Next, adjust the network manually for visual clarity. Color schemes can be changed using the clustering menu (top left). **Note:** any refresh of the page will ruin manual adjustment.
- This is the most important step and relies on the researcher’s sense of biology and similarity. Examine proteins and their function carefully to correctly adjust the network and assign function. The output from DAVID is also often useful to assist in network identification.
- 10) Capture the screen, in tiles if needed, and stitch together (Photoshop montage feature etc.). Annotate (Illustrator etc.). An example for upregulated genes from the course sample follows.
  - 11) Compare the results of DAVID with String. A well executed analysis will result in obvious similarities.
  - 12) Finally, to find the statistical chance that the proteins identified (up or down regulated) are similar to other experiments or to published literature, use the hypergeometric distribution with some assumption for the number of active genes.

The hypergeometric distribution calculates the chances of a given number of marked samples being taken from a finite population of marked and unmarked entities without replacement. In the

classical example, if  $x$  marbles are taken from an urn containing  $y$  white and  $z$  black marbles, it calculates the likelihood of choosing  $n$  black marbles within the removed  $x$  marbles. Applied to proteomics by example, assume proteomic study 1 identified 361 proteins and proteomic study 2 identified 260 with the studies sharing 52 common proteins. If we assume that there are 74,016 actively transcribed genes (the number of IPI entries in an IPI release), the urn would contain a total of 74,016 marbles of which 361 would be black. If 260 are chosen at random, the chance of those 260 containing 52 black marbles can be calculated in Excel with the formula =HYPGEOMDIST(52,361,260,74016), which returns 1.08 E -67. The chance is dependent on the assumed number of active genes (given the same number of shared proteins, as the population size increases, the chance that the shared proteins are random decreases). For example, assuming 5381 proteins (rather than 74,016) were actively transcribed, the chance rises to a still unlikely 1.60 E -13.

Your dataset can be compared to other genes found to be upregulated in cisplatin resistance at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1586025/table/T2/> .

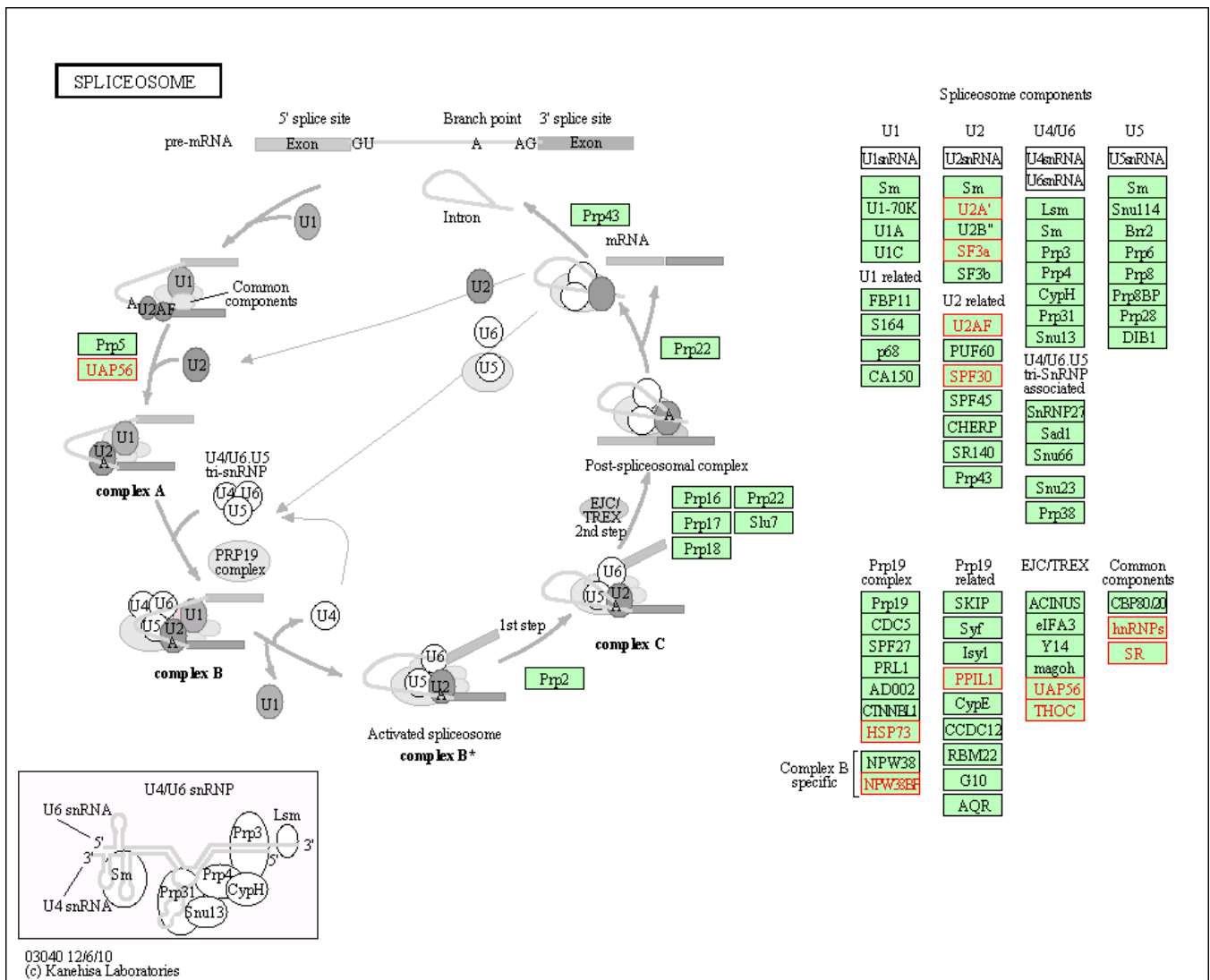
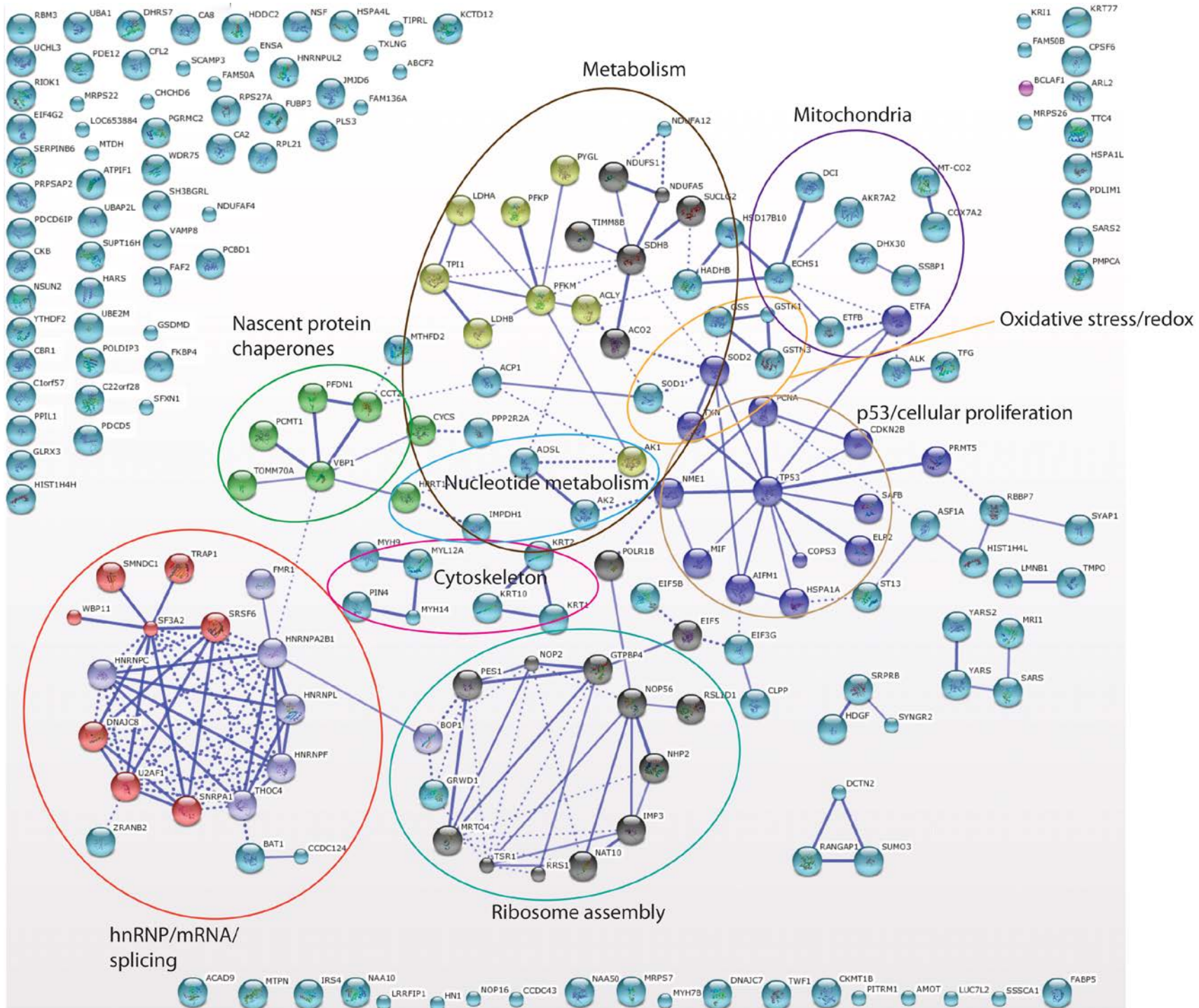
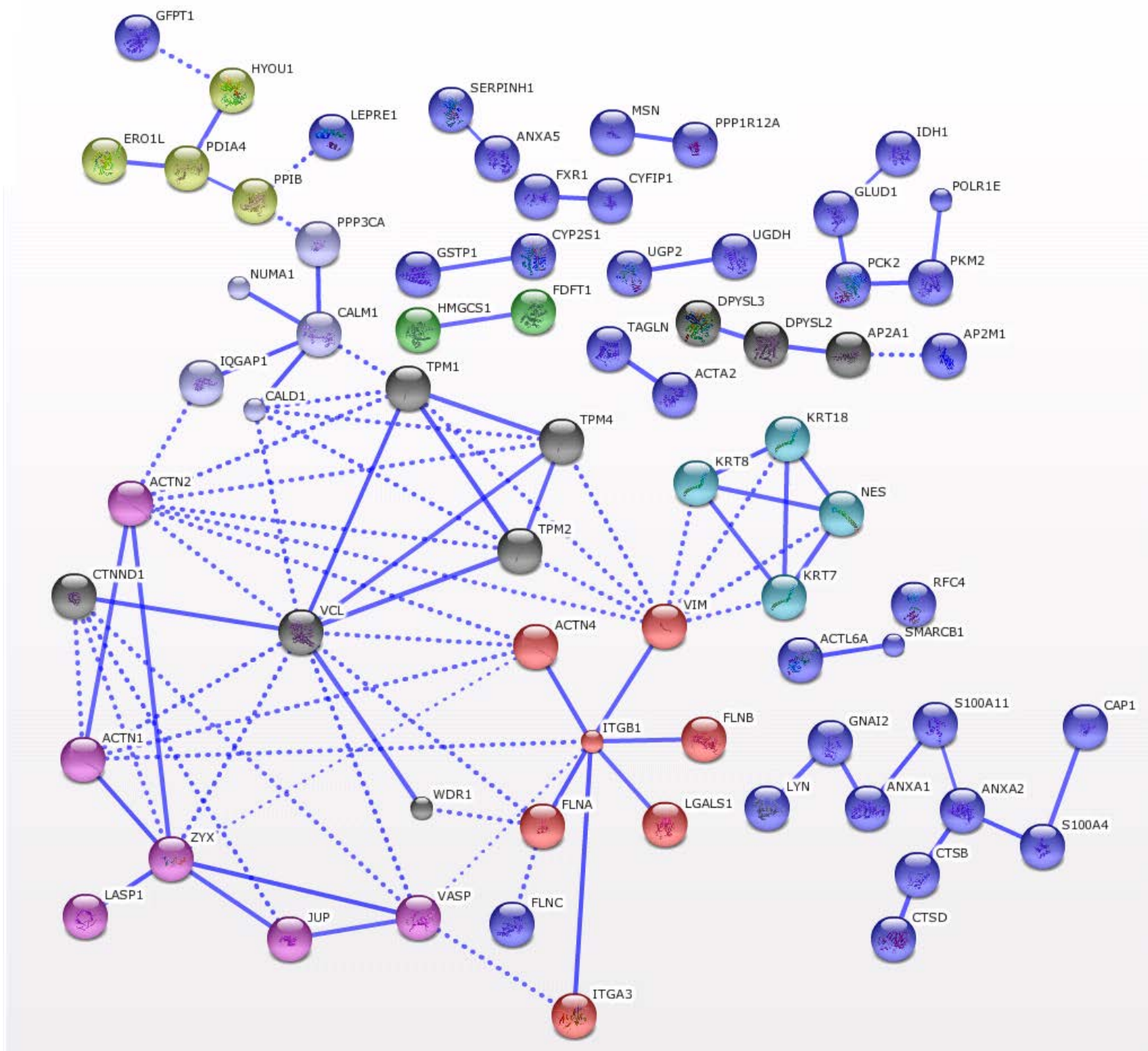


Illustration 1 (above): the spliceosome pathway is enriched in the genes upregulated in OVCAR5 over PA1 cells. Detected upregulated proteins are shown in orange.

Illustration 2 (page 5): genes upregulated in OVCAR5 cisplatin resistant cells compared to PA1 cells; course data, after manual adjustment. Confidence level 0.65. Multiple networks not found through the Kegg pathway analysis performed by Apropos are clearly visible. Please note that additional work not done below would be required before publication, for example ensuring that proteins within the annotated fields fall physically outside the designations, that all gene symbols within a designation really are correctly described by that designation, ensuring that gene symbol labels are not overwritten by oval designations, possibly deleting all unnetworked proteins, potentially including a key (String conveniently provides this), etc.

Illustration 3 (page 6): genes downregulated in OVCAR5 cells compared to PA1 cells. The networks are related to the cytoskeleton. Confidence level 0.8.





## Example of statistical comparison using the hypergeometric distribution

### Data:

Detected proteins upregulated in OVCAR5 cells (177): CBR1, KCTD12, TP53, HSPA4L, MIF, GSTM3, CKB, IRS4, MYH9, ATP1F1, LOC728641, LDHA, HSPA1B, ETFA, CA8, GTPBP4, THOC4, FKBP4, PCBD1, IMPDH1, BOP1, Q5T5C8\_HUMAN, SARS, KRT10, NDUFA12, RANGAP1, COX2, PDCD5, SSSCA1, HSPA1L, SDHB, SYAP1, KRT2, CXorf15, TXN, NSF, IMP3, SCAMP3, RPL21P19, SUPT16H, SYNGR2, HPRT1, AIFM1, YARS, ACLY, HNRNPC, C1orf57, SOD2, CHCHD6, NDUFS1, CCDC124, ABCF2, DNAJC7, UBXD8, SUMO3, TTC4, NOP56, SFXN1, NDUFA5, RBBP7, SF3A2, SSBP1, MYH14, ACAD9, PPP2R2A, TPI1, SUCLG2, PRMT5, PGRMC2, PFDN1, PPIL1, MRPS7, C22orf28, PES1, YARS2, RIOK1, LMNB1, SUMO2, ACO2, DNAJC8, MTHFD2, KRT77, UBE1, PLS3, GSDMD, VAMP8, LDHB, U2AF1, FAM50A, SOD1, FAM136A, SMNDC1, ETFB, PRPSAP2, B7Z5B6\_HUMAN, YTHDF2, HDGF, PCNA, WDR75, ST13, SNRPA1, SERPINB6, AKR7A2, MRPS22, TIMM8B, PFKM, NDUFAF4, FMR1, UBE2M, CLPP, CCDC43, UBC, RBM3, SRPRB, SH3BGRL, PYGL, HSPC111, ARL2, GSTK1, MYL12A, PCMT1, UBAP2L, HARS, SFRS6, EPPK1, MTPN, VBP1, PDLIM1, NAT10, HIST4H4, GLRX3, GRWD1, PFKP, NAA10, ECHS1, GSS, PMPCA, CA2, MTDH, WBP11, RSL1D1, EIF5, HNRNPUL2, DCI, DCTN2, LRRFIP1, RRS1, PDCD6IP, ADSL, AK1, HSD17B10, KRT1, MRPS26, TSR1, NHP2, TOMM70A, HNRNPF, EIF4G2, TRAP1, HNRNPL, ACP1, TFG, HADHB, COPS3, CYCS, SFRS13A, LOC727922, CCT2, NSUN2, EIF3G, UCHL3, COX7A2, EIF5B, FUBP3, MRTO4, ASF1A, CFL2.

Detected proteins downregulated in OVCAR5 cells (168): TMEM97, CAP1, MPP6, CALD1, AKAP8L, FDFT1, TPM2, TMEM93, ZYX, PODXL, UGP2, CRABP1, CNN2, ACTN1, TXNL1, FLNC, PKM2, FXR1, NUMA1, MAN2A1, PDIA4, CTNND1, CORO1B, PPIB, PON2, CHD5, RRAS2, SMARCB1, HYOU1, S100A4, ANXA2, UBLCP1, FERMT2, MSN, AP2M1, MAP1B, PPP3CA, ACTA2, ANLN, CKAP4, PAWR, HDGF2, CTSD, SH3BGRL3, RCN2, ESYT1, ALDH16A1, ERO1L, DPYSL3, KRT18, PVRL2, LASP1, GLG1, KRT8, PPM1B, SLC25A13, GALNT2, WDR1, RANBP3, CLIC4, GARS, COTL1, SERPINH1, CTSB, VCL, PTK7, C20orf4, SNX27, FSCN1, ANXA5, FLNA, ANXA1, ERLIN2, SURF4, VAT1L, PEA15, ANXA3, UGGT1, TPD52L2, HIST1H1B, BCAT1, GFPT2, ACTL6A, GNAI2, HSDL2, SERPINB9, PLD3, COPG2, COL18A1, LEPRE1, TPM1, TAGLN, TPM4, VASP, FLNB, SLC7A5, RSU1, ARRB2, LGALS1, UGDH, S100A11, PDF, NLN, ENOPH1, NES, MLEC, INA, POLR1E, PAPSS1, DDOST, TOR1AIP1, IMPA1, AKAP12, DPYSL2, DHCR24, ACTN2, PGD, ERLIN1, GSTP1, CLDN6, JUP, SLC2A1, VIM, SLC38A2, LIN28A, KRT7, AP2A1, GLUD1, APOE, CYP2S1, ITGB1, CYFIP1, IQGAP1, CALU, DDX42, MAP7D1, ALDH2, SLC29A1, CD2BP2, PPP1R12A, CALM2, LYN, GFPT1, CNDP2, NXF1, SLC2A3, IDH1, CALM1, RPS27L, HMGCS1, CNN3, RFC4, BAX, DSTN, PCK2, ATP5B, AHNAK, ACTN4, SLC4A7, DYNLL2, RCN1, SETD3, FKBP5, IGF2BP2, ITGA3, SCRNI1, TPP2, CAV1.

Comparison will be to: Zhang P, Zhang Z, Zhou X, Qiu W, Chen F, Chen W. **Identification of genes associated with cisplatin resistance in human oral squamous cell carcinoma cell line.** BMC Cancer. 2006 Sep 15;6:224.

Upregulated (38) and downregulated (25) genes from this study. Red bolded are shared down-regulated symbols. Green bolded are shared upregulated. In common were five downregulated and one upregulated protein.

Classification	Gene	Change Fold <sup>a</sup>			
<b>Metabolism</b>	AKR1C3	-3.96	PDE8B	2.14	
	ALDH3B1	-2.14	IL13RA1	3.03	
	GPI	-2	ARF6	2.14	
	NNMT	-2	GCA	7.06	
	MAOB	-2	PIP5K1A	2	
	NAGA	2.14	ITPR1	2.64	
	FLJ12443	2.14	<b>Oncogene</b>	FYN	2.83
	SLC27A2	2.3	FGFR3	3.03	
	<b>SLC2A3</b>	2.83	RAB31	2.83	
	<b>GARS</b>	2	EMS1	-2.14	
QPRT	2.14	<b>Others</b>	CHST2	-2.83	
<b>Cell cycle</b>	CCND3	-2.83	C1S	-2.83	
	CCND1	4.29	H2AFO	-2.3	
	ASNS	2.83	PPL	-2.3	
<b>Transcript factor</b>	TRIM29	-2	TGM2	-2.3	
	ZFP36	-2.83	KIAA0992	-2.14	
	CREM	2.83	SELENBP1	-2	
	ETV5	4.29	H2BFB	-2	
	ID1	2	DKFZp564J0323	-2	
	ID3	4	<b>PON2</b>	-2	
	CAMTA2	2	FUSIP1	2	
			<b>DDOST</b>	2	
<b>Transport</b>	MUC1	-2.14	ODAG	2.14	
	ATP1B1	-2	RECQL	2.14	
	TCIRG1	-2.3	DMD	2.14	
	COX8	-2.46	G1P2	2.3	
	KPNB2	2.3	RAFTLIN	2.3	
<b>Signal transduction</b>	MAP2K6	-2.3	SEMA3F	2.3	
	IGFBP3	-5.66	N33	2.3	
	<b>PDLIM1</b>	2	TUBA3	2.3	
	IGFBP7	2.3	<b>KRT7</b>	2.46	
			<a href="#">BC008967</a>	3.25	
		ITGA1	3.48		



Probabilities associated with the shared 1 or 5 proteins of up- and down-regulated genes given various assumptions of the number of active genes. The shared five downregulated proteins are statistically significant. Note that as the number of active genes increases, even a single shared hit can become statistically relevant. An example Excel formula for the highlighted value is displayed below the table. Note that in a real comparison, extensive caution is necessary to verify that the same gene symbols were being compared (one gene or protein can have many gene symbols) and were consistently present (e.g. the designations BC008967 or Q5T5C8\_HUMAN wouldn't exactly work).

Active genes	Probabilities	
	Up	Down
50000	0.118060	2.01E-08
40000	0.142825	6.03E-08
30000	0.180313	2.47E-07
20000	0.242415	1.78E-06
10000	0.348239	4.82E-05
5000	0.355408	0.001104

=HYPGEOMDIST(5,25,168,5000)