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. Recommended URLs: www.pathwaycommons.org; cancer.cellmap.org/cellmap/cytoscape.do; www.cytoscape.org; locate.imb.uq.edu.au/. Highly recommended reading: 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 (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. Not that the network exhibits switch points and at some point, very fine adjustments (e.g. ±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/ .

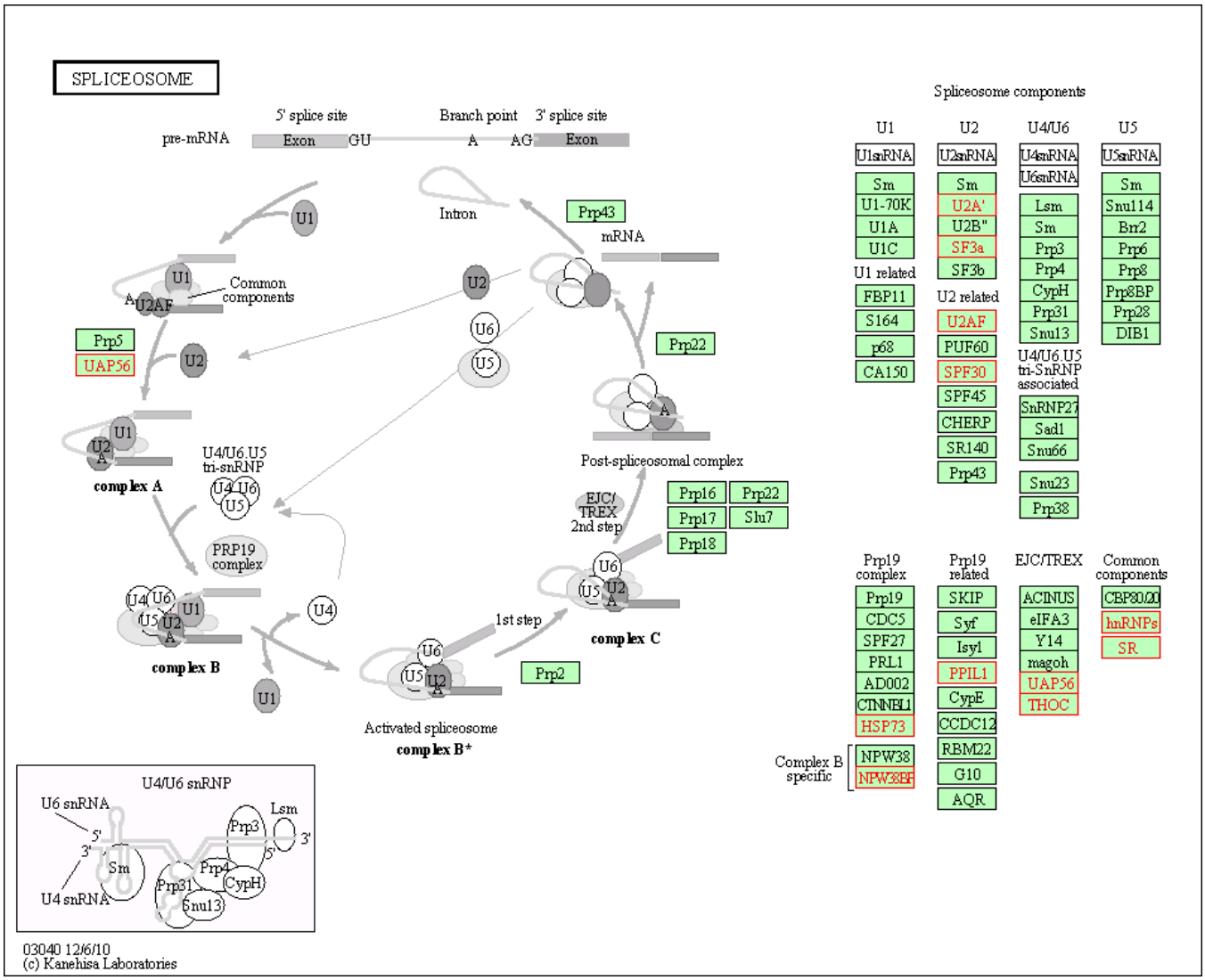
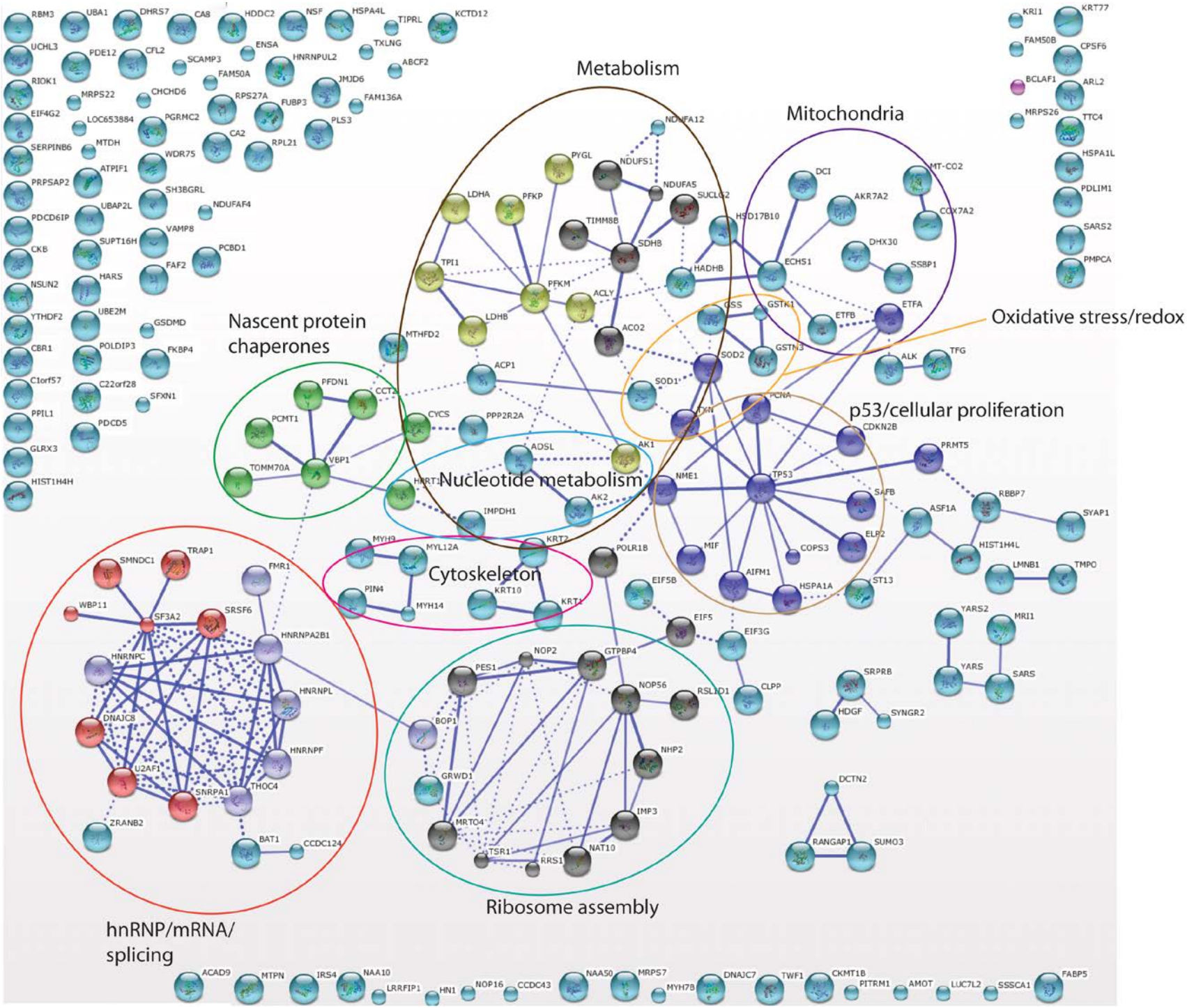
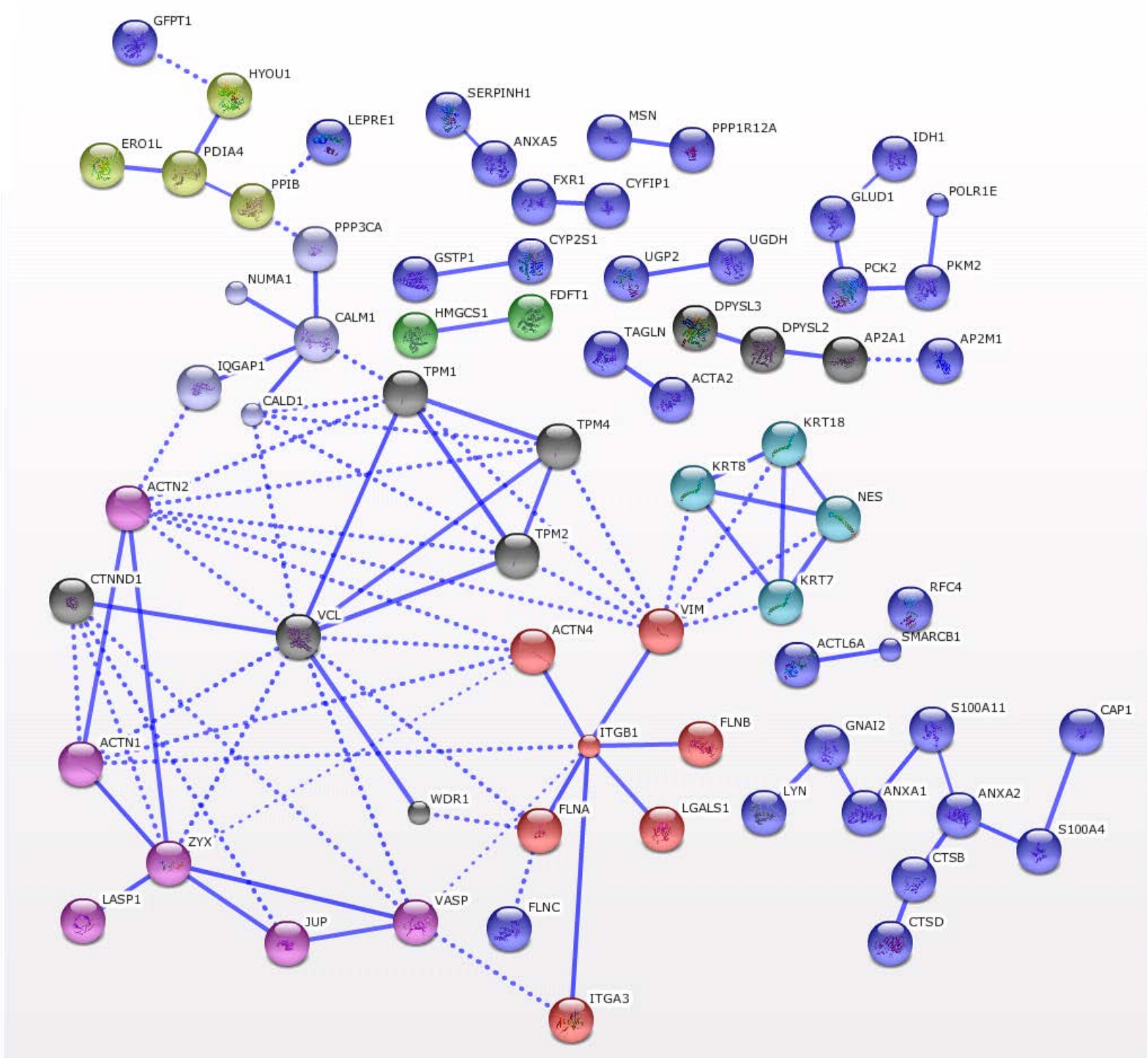


Illustration 1 (above): the spliceosome pathway is enriched in the genes upregulated in OVCAR5 over PA1 cells. Detected upregulated proteins are shown in orange.

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.


Metabolism
Mitochondria
Oxidative stress/redox
Nascent protein chaperones
p53/cellular proliferation
Nucleotide metabolism
Cytoskeleton
hnRNP/mRNA/ splicing
Ribosome assembly
RBM3
UBA1
DHRS7
CA8
HDDC2
NSF
HSPA4L
TIPRL
KCTD12
UCHL3
PDE12
CFL2
SCAMP3
ENSA
HNRNPUL2
TXLNG
ABCF2
RIOK1
MRPS22
CHCHD6
FAM50A
RPS27A
FUBP3
JMJD6
FAM136A
EIF4G2
LOC653884
PGRMC2
CA2
RPL21
PLS3
SERPINB6
MTDH
WDR75
ATPIF1
PRPSAP2
SH3BGRL
NDUFAF4
PDCD6IP
UBAP2L
VAMP8
CKB
SUPT16H
PCBD1
FAF2
NSUN2
HARS
YTHDF2
UBE2M
GSDMD
CBR1
POLDIP3
FKBP4
C1orf57
C22orf28
SFXN1
PPIL1
PDCD5
GLRX3
HIST1H4H
KRI1
KRT77
FAM50B
CPSF6
BCLAF1
ARL2
MRPS26
TTC4
HSPA1L
PDLIM1
SARS2
PMPCA
NDUFA12
PYGL
NDUFS1
NDUFA5
SUCLG2
LDHA
PFKP
TIMM8B
SDHB
HSD17B10
DCI
MT-CO2
AKR7A2
COX7A2
DHX30
SSBP1
TPI1
PFKM
HADHB
ECHS1
ACLY
LDHB
GSS
GSTK1
ETFB
ETFA
MTHFD2
ACO2
GSTM3
SOD2
ALK
TFG
ACP1
SOD1
PFDN1
CCT2
PCMT1
CYCS
PPP2R2A
VBP1
TOMM70A
PCNA
CDKN2B
PRMT5
ADSL
AK1
HPRT1
AK2
IMPDH1
NME1
TP53
SAFB
RBBP7
ASF1A
SYAP1
ELP2
HIST1H4L
MIF
COPS3
AIFM1
HSPA1A
ST13
LMNB1
TMPO
POLR1B
EIF5B
EIF5
EIF3G
MYH9
MYL12A
KRT2
PIN4
MYH14
KRT10
KRT1
YARS2
MRI1
YARS
SARS
SRPRB
HDGF
SYNGR2
CLPP
NOP2
GTPBP4
PES1
NOP56
RSL1D1
BOP1
NHP2
GRWD1
IMP3
MRTO4
TSR1
NAT10
RRS1
DCTN2
RANGAP1
SUMO3
SMNDC1
TRAP1
FMR1
WBP11
SF3A2
SRSF6
HNRNPA2B1
HNRNPC
HNRNPL
DNAJC8
HNRNPF
U2AF1
THOC4
SNRPA1
ZRANB2
BAT1
CCDC124
ACAD9
MTPN
IRS4
NAA10
LRRFIP1
HN1
NOP16
CCDC43
NAA50
MRPS7
MYH7B
DNAJC7
TWF1
CKMT1B
PITRM1
AMOT
LUC7L2
SSSCA1
FABP5


GFPT1
HYOU1
SERPINH1
LEPRE1
MSN
PPP1R12A
ERO1L
PDIA4
ANXA5
IDH1
FXR1
CYFIP1
GLUD1
POLR1E
PPIB
PPP3CA
CYP2S1
UGP2
UGDH
GSTP1
PCK2
PKM2
NUMA1
FDFT1
DPYSL3
HMGCS1
CALM1
TAGLN
DPYSL2
AP2A1
AP2M1
IQGAP1
TPM1
ACTA2
CALD1
TPM4
KRT18
KRT8
ACTN2
NES
TPM2
CTNND1
KRT7
RFC4
VIM
VCL
ACTN4
SMARCB1
ACTL6A
S100A11
CAP1
ITGB1
FLNB
GNAI2
ACTN1
LYN
ANXA1
ANXA2
WDR1
FLNA
LGALS1
S100A4
ZYX
CTSB
LASP1
VASP
FLNC
JUP
CTSD
ITGA3

6

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, ATPIF1, 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, SCRN1, 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 [a]
Metabolism	AKR1C3	-3.96
	ALDH3B1	-2.14
	GPI	-2
	NNMT	-2
	MAOB	-2
	NAGA	2.14
	FLJ12443	2.14
	SLC27A2	2.3
	SLC2A3	2.83
	GARS	2
	QPRT	2.14
Cell cycle	CCND3	-2.83
	CCND1	4.29
	ASNS	2.83
Transcript factor	TRIM29	-2
	ZFP36	-2.83
	CREM	2.83
	ETV5	4.29
	ID1	2
	ID3	4
	CAMTA2	2
Transport	MUC1	-2.14
	ATP1B1	-2
	TCIRG1	-2.3
	COX8	-2.46
	KPNB2	2.3
Signal transduction	MAP2K6	-2.3
	IGFBP3	-5.66
	PDLIM1	2
	IGFBP7	2.3
	PDE8B	2.14
	IL13RA1	3.03
	ARF6	2.14
	GCA	7.06
	PIP5K1A	2
	ITPR1	2.64
Oncogene	FYN	2.83
	FGFR3	3.03
	RAB31	2.83
	EMS1	-2.14
Others	CHST2	-2.83
	C1S	-2.83
	H2AFO	-2.3
	PPL	-2.3
	TGM2	-2.3
	KIAA0992	-2.14
	SELENBP1	-2
	H2BFB	-2
	DKFZp564J0323	-2
	PON2	-2
	FUSIP1	2
	DDOST	2
	ODAG	2.14
	RECQL	2.14
	DMD	2.14
	G1P2	2.3
	RAFTLIN	2.3
	SEMA3F	2.3
	N33	2.3
	TUBA3	2.3
	KRT7	2.46
	BC008967	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).

	Probabilities	
Active genes	**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)