Supplementary information

Ying-Ying Xu, Fan Yang, Yang Zhang, and Hong-Bin Shen, Bioimaging based detection of mislocalized proteins in human cancers by semi-supervised learning.

CONTENTS

Supplementary text: evaluating metrics

In this study, seven multi-label metrics were used to evaluate the performance of the classifier model. Suppose there are L classes. Let $\hat{Y}_{t_j} = [\hat{y}_1^j, \hat{y}_2^j, \mathrm{L}\ , \hat{y}_L^j]$ denotes the predicted label vector of the j-th test sample t_j, while $Y_{t_j} = [y_1^j, y_2^j, \mathrm{L}\ , y_L^j]$ is the corresponding real vector. The five metrics are defined below:

1) Subset accuracy

$$Subset\_accuracy = \frac{1}{q}\sum_{j=1}^{q} \Phi(\hat{Y}_{t_j} = Y_{t_j}) \tag{S1}$$

where $\Phi(g) = \begin{cases} 1, & g \text{ is true} \\ 0, & otherwise \end{cases}$.

Subset accuracy is the fraction of samples whose predicted label set is the same as the true label set. This metric is severe and ignores the much difficulty against single-label learning. Yet it is direct-viewing, and can reflect the performance of the classification.

2) Accuracy

$$Accuracy = \frac{1}{q}\sum_{j=1}^{q} point(\hat{Y}_{t_j}) \tag{S2}$$

Each test sample prediction can be scored by:

$$point(\hat{Y}_{t_j}) = \frac{\sum_{l=1}^{L} \Phi(y_l^j = 1, \hat{y}_l^j = 1)}{\sum_{l=1}^{L} \Phi(y_l^j = 1, or\ \hat{y}_l^j = 1)} \tag{S3}$$

Accuracy is more lenient to errors than subset accuracy because if not all the predicted labels of a sample are correct, then subset accuracy gives 0, but accuracy gives a value between 0 and 1,

reflecting the degree of partial correctness.

3) Recall

For a class l,

$$Recall(l) = \frac{1}{\sum_{j=1}^{q} \Phi\left(y_l^j = 1\right)} \sum_{t_j \in \{t_j | y_l^j = 1\}} point(\hat{Y}_{t_j}) \tag{S4}$$

Then the uniform recall of the total testing samples is computed as:

$$Recall = \frac{1}{L} \sum_{l=1}^{L} Recall(l) \tag{S5}$$

4) Precision

We can obtain precision in a similar way:

$$Precision(l) = \frac{1}{\sum_{j=1}^{q} \Phi\left(y_l^j = 1\right)} \sum_{t_j \in \{t_j | y_l^j = 1\}} point(\hat{Y}_{t_j}) \tag{S6}$$

$$Precision = \frac{1}{L} \sum_{l=1}^{L} Precision(l) \tag{S7}$$

The above two metrics are extensions of the classic definitions to measure recall and precision of each class in traditional single-label learning. Recall is the fraction of true labels that are correctly predicted, while precision is the fraction of predicted labels that are correctly predicted.

5) Label accuracy

For a class l,

$$Label\_accuracy(l) = \frac{1}{q} \sum_{j=1}^{q} \Phi\left(y_l^j = y_l^j\right) \tag{S8}$$

$$Average\_label\_accuracy = \frac{1}{L} \sum_{l=1}^{L} Label\_accuracy(l) \tag{S9}$$

Label accuracy evaluates the prediction accuracy for each label, from which we can identify which subcellular locations are easier to recognize. The average label accuracy computes the average of L accuracies of labels, and can reflect the total performance.

6) Sensitivity

For a class l,

$$Sensitivity(l) = \frac{TP}{TP + FN} \tag{S10}$$

Sensitivity measures the proportion of actual positives which are correctly identified (TP), and is complementary to the false negative (FN) rate.

7) AUC

AUC is the area under the receiver operating characteristic (ROC) curve. ROC is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the TP rate against the false positive (FP) rate at various threshold settings. AUC can reflect the performance of classification model. The bigger

the AUC, the better the performance.

Supplementary Table 1 List of proteins with high and medium staining

High level stained proteins (348 proteins)			
BAD	SLC25A13	LIG3	CREBBP
AASS	MVP	UFL1	MRE11A
KCNG1	SNX1	SARS	OTC
USP28	CHRDL2	NOP58	ACAA1
CS	HIPK2	CTSA	EYA2
ABCA7	ZC3H15	YBX1	ELAVL1
GOLGA5	ARFGEF1	TP53BP1	IARS2
IDH3G	NUCB2	POLB	PABPC1
DPP8	ACAT1	FOSL2	RAB7A
MLH1	TOP2B	PHF17	ITCH
SDF4	SAR1A	SENP1	PTPRH
ME2	PALB2	OXCT1	HADHA
MAPRE3	SLC25A24	ORC1	FOLH1
PSMC5	CNOT3	MAVS	XRN2
RPLP0	NUDC	FCGBP	DLD
FH	CDC5L	MED15	SMARCB1
POLR2F	CYB5R3	RPL3	TXN2
ST13	ACO2	PRMT5	TIMM9
AHSA1	MTHFD1	APEX1	VAPA
FMR1	AXIN1	USP10	NUBP1
RBL2	TMEM87A	CA2	EIF3E
NDRG1	SH2D4A	ECH1	PDCD5
PLIN3	ETFB	LIG1	GCDH
DFNA5	ZC3HAV1	EIF3A	MAP3K8
XPNPEP1	TFAM	CBX1	ENO3
LRRC59	LPXN	OSBP	CPT1A
CARS	CAMKK2	ATP5B	KRT18
CHD4	SLCO1B3	ALDH5A1	GMNN
SRF	HSPA9	COL4A3BP	RASGRF2
PPWD1	ZNF346	HYAL1	NCL
GALNT3	PECR	HDLBP	DARS
CACYBP	MECR	ACADM	SDHB
CDC20	PRDX1	ABCD3	FOXO3
SET	NDUFA8	SLIRP	AKAP1
CLCC1	GTF3A	HNRNPA2B1	VPS26A
OPTN	PREX1	SNAI1	VAMP7
DEK	GTF2F1	CLPP	RRBP1
PRDX5	TRAP1	ELK1	KTN1
TIMM8A	CANX	ECHS1	CALU
IVD	FXR2	ILF3	GNL3L

PRKCSH	RPL36	PLVAP	ATPIF1
TOP2A	VIMP	H3F3B	RTN3
HSD17B4	IER3IP1	CCNB1	CDK7
LAMTOR5	FBXO18	YARS	TIMM10
GOLM1	MAP7	SP110	COX5B
MRPS9	NMT1	BRIP1	BIN1
TXN	GOLGA1	DERL1	POLR1E
CCDC90B	RMDN3	DBT	PNPT1
LRPPRC	AOX1	COX17	G3BP2
CASP6	PEX5	ESD	ETFA
BBS4	GLYR1	UQCRC2	SERBP1
HDGF	POGZ	GALNT2	GULP1
HSPD1	SLC25A26	SNCA	BHMT
G3BP1	MDH2	TRIM4	NDUFB11
ERLIN2	MTDH	NAPRT1	MKI67
TAF5	PPP4C	HMGA2	DLAT
OXSM	TWF1	PABPC3	TGOLN2
XRCC4	SAR1B	SMARCA5	PTPRR
MSI2	ATP5J	RBM45	PDIA4
PHF6	KAT6B	RER1	NCK1
DUSP23	HLCS	NDUFV3	ADAR
RUSC1	LMNA	SQSTM1	PSMC2
SYVN1	ZFPL1	AK4	RBBP4
IVL	APEH	ABCE1	NDUFS4
COG5	CDCA7L	FOXK1	VDAC2
ZMYND19	DDX21	HSP90B1	CASC4
TERF2IP	MARS	PDIA3	GPRC5B
CDK12	STIM1	TRIM68	CDT1
ATP5H	SDHAF2	SLC3A2	IRF2BP2
BMI1	FEN1	MECP2	CSNK1G1
CLIC4	HINT1	CLIC3	CHD3
HNRNPA3	ELP5	CYCS	CALB2
RELA	KDM2A	GOLGB1	SMARCC1
TOMM20	GOLIM4	MRPL45	MRPS22
SHMT1	AGTRAP	ASB8	TUFM
CYC1	GCC1	PDXDC1	PAK2
NPM1	SHMT2	GLRX5	ACBD3
SLC25A10	DIABLO	EIF4ENIF1	H1FX
P4HB	BCAP31	EVI2B	RXRA
HEXIM1	AKR1C1	H1F0	FHIT
S100A4	IARS	LONP1	HDAC2
GM2A	TLE1	FAM3C	SND1
GTF2E2	BLM	GSTK1	CDC42SE1
PSAP	S100A6	TLK1	ECI2
TOP1	BAG6	SNX2	CLIC1

FIS1	PHB2	IFI30	ADSL
RP11-286N22.8	SLC25A10	MRPL12	IER3IP1
BAD	SLC25A13	LIG3	CREBBP
AASS	MVP	UFL1	MRE11A
KCNG1	SNX1	SARS	OTC
USP28	CHRDL2	NOP58	ACAA1
CS	HIPK2	CTSA	EYA2
ABCA7	ZC3H15	YBX1	ELAVL1
GOLGA5	ARFGEF1	TP53BP1	IARS2
IDH3G	NUCB2	POLB	PABPC1
DPP8	ACAT1	FOSL2	RAB7A
MLH1	TOP2B	PHF17	ITCH
SDF4	SAR1A	SENP1	PTPRH
ME2	PALB2	OXCT1	HADHA
MAPRE3	SLC25A24	ORC1	FOLH1
PSMC5	CNOT3	MAVS	XRN2
RPLP0	NUDC	FCGBP	DLD
FH	CDC5L	MED15	SMARCB1
POLR2F	CYB5R3	RPL3	TXN2
ST13	ACO2	PRMT5	TIMM9
AHSA1	MTHFD1	APEX1	VAPA
Medium level stained proteins (384 proteins)			
BAD	SLC25A13	LIG3	CREBBP
SCIN	PRSS21	AASS	BAZ1B
RNF216	MVP	UFL1	MNAT1
MRE11A	KCNG1	SNX1	SARS
OTC	USP28	CHRDL2	NOP58
ACAA1	CS	HIPK2	CTSA
EYA2	ABCA7	WDR3	MYLK
ZC3H15	YBX1	ELAVL1	GOLGA5
ARFGEF1	TP53BP1	IARS2	IDH3G
NUCB2	POLB	PABPC1	IKBKG
DPP8	ACAT1	RAB7A	MLH1
TOP2B	PHF17	VDAC3	ITCH
SDF4	SAR1A	SENP1	ME2
PALB2	OXCT1	HADHA	MAPRE3
SLC25A24	ORC1	PSMC5	CNOT3
MAVS	XRN2	RPLP0	NUDC
DLD	FH	SUPT16H	CDC5L
MED15	SMARCB1	POLR2F	CYB5R3
RPL3	TXN2	ST13	ACO2
PRMT5	TIMM9	AHSA1	MTHFD1
APEX1	VAPA	FMR1	USP11
PARP4	PRSS54	AXIN1	USP10
NUBP1	RBL2	TMEM87A	CA2

EIF3E	NDRG1	SH2D4A	NEFM
ECH1	PDCD5	PLIN3	ETFB
LIG1	GCDH	DFNA5	ZC3HAV1
SH3GL2	EIF3A	MAP3K8	XPNPEP1
TFAM	KPNB1	CBX1	ENO3
FTSJ3	LRRC59	LPXN	OSBP
CPT1A	CARS	CAMKK2	ATP5B
KRT18	CHD4	ALDH5A1	GMNN
SRF	HSPA9	COL4A3BP	RASGRF2
PPWD1	ZNF346	HYAL1	NCL
GALNT3	PECR	HDLBP	DARS
CACYBP	MRPL37	MECR	ACADM
SDHB	CDC20	PRDX1	ABCD3
FOXO3	SET	NDUFA8	SLIRP
AKAP1	CAT	CLCC1	GTF3A
HNRNPA2B1	VPS26A	OPTN	PREX1
SNAI1	VAMP7	UPF3B	GTF2F1
CLPP	RRBP1	ITPA	PRDX5
TRAP1	ELK1	KTN1	SIX1
TIMM8A	CANX	ECHS1	CALU
IVD	FXR2	ILF3	GNL3L
PRKCSH	RPL36	ATPIF1	MAP1B
TOP2A	VIMP	H3F3B	CA1
HSD17B4	IER3IP1	CCNB1	CDK7
LAMTOR5	YARS	TIMM10	GOLM1
RINT1	MAP7	SP110	COX5B
MRPS9	STAB2	NMT1	TACO1
BRIP1	BIN1	TXN	GOLGA1
DERL1	POLR1E	CCDC90B	RMDN3
DBT	PNPT1	LRPPRC	RPS24
AOX1	COX17	G3BP2	CASP6
PEX5	ESD	ETFA	BBS4
GLYR1	UQCRC2	SAE1	SERBP1
HDGF	ARNT	POGZ	DTL
GALNT2	ACP1	GULP1	HSPD1
SLC25A26	SNCA	G3BP1	MDH2
TRIM4	NDUFB11	ERLIN2	MTDH
NAPRT1	MKI67	TAF5	SSRP1
DSN1	PPP4C	HMGA2	DLAT
OXSM	TWF1	MBIP	PABPC3
TGOLN2	XRCC4	SAR1B	SMARCA5
PTPRR	MSI2	TRIM11	ATP5J
RBM45	PDIA4	PHF6	KAT6B
RER1	NCK1	DUSP23	HLCS
NDUFV3	CBS	ADAR	RUSC1

LMNA	SQSTM1	PSMC2	SYVN1
ZFPL1	AK4	RBBP4	IVL
LMOD1	APEH	ABCE1	NDUFS4
COG5	CDCA7L	FOXK1	VDAC2
ZMYND19	DDX21	HSP90B1	CASC4
TERF2IP	MARS	PDIA3	GOLGA2
GPRC5B	CDK12	STIM1	TRIM68
CDT1	ATP5H	SRP68	SDHAF2
IRF2BP2	BMI1	PXK	FEN1

Supplementary Table 2 Details of the independent cancer protein biomarkers (IDN) dataset

No.	Protein name	Tissues	Normal locations	Cancer locations	Supporting Literature
1	Bax	Lymphoma	Cyto.	Mito.	Nechushtan et al. 1999
2	Cyclin D1	Ovary	Nucl.	Nucl.& Cyto.	Dhar et al. 1999
3	PTEN	Pancreas	Cyto.& Nucl.	Nucl.	Perren et al. 2000
4	BAG-1	Colon	Nucl.	Mito.	Takayama et al. 1998
5	GOLGA5	Thyroid gland	Gol.	Mito.	Klugbauer et al. 1998
6	NQO1	Lung	Cyto.	Nucl.	Winski et al. 2002
7	SOX9	Breast	Nucl.	Cyto.	Bratthauer et al. 2009
8	p53	Breast; Ovary	Nucl.	Nucl.& Cyto.	Moll et al. 1992; Runnebaum et al. 1996
9	TOP2A	Lung	Nucl.	Cyto.	Feldhoff et al. 1994
10	IGFBP	Breast	Nucl.	Cyto.	Akkiprik et al. 2009

Supplementary Table 3 Results of AsemiB^{E} and AsemiBC^{E} tested on IDN

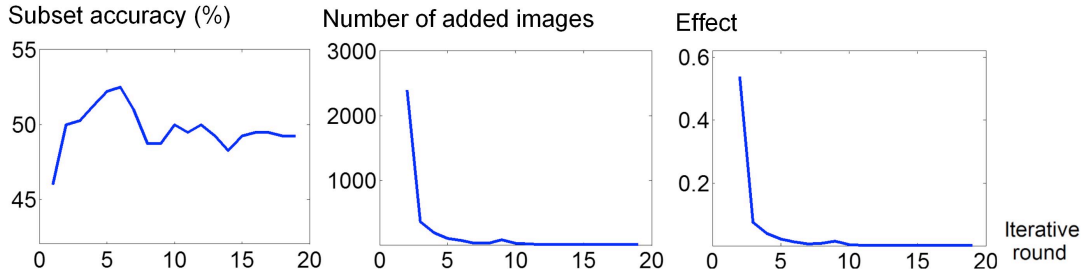
Evaluating metrics	Subset accuracy	Accuracy	Recall	Precision	Average label accuracy
AsemiB^{E}	51.75%	62.54%	40.59%	41.41%	86.75%
AsemiBC^{E}	52.5%	63.1%	41.8%	44.54%	87.04%

Supplementary Table 4 Results of single and ensemble classifiers tested on IDN

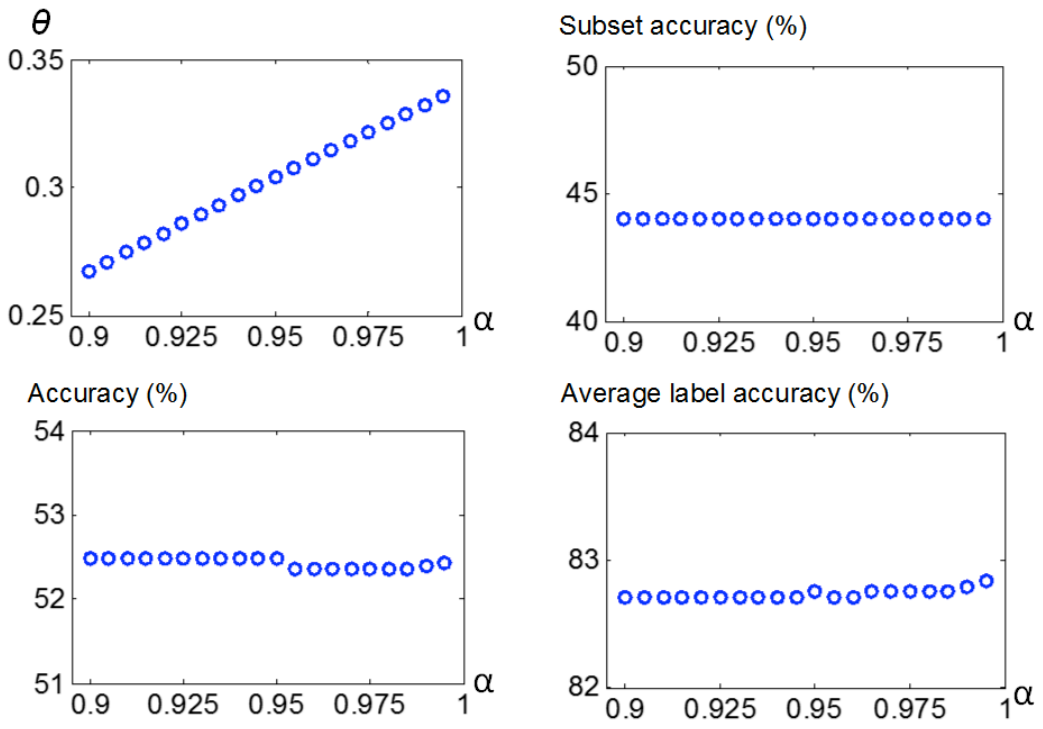
Single classifiers are classifiers using db7 features. Ensemble classifiers are AsemiBCE and AsemiBE. Refer to the text for more details about these classifiers.

Evaluating	Method	Cyto.	ER	Golgi.	Mito.	Nucl.	Vesi.

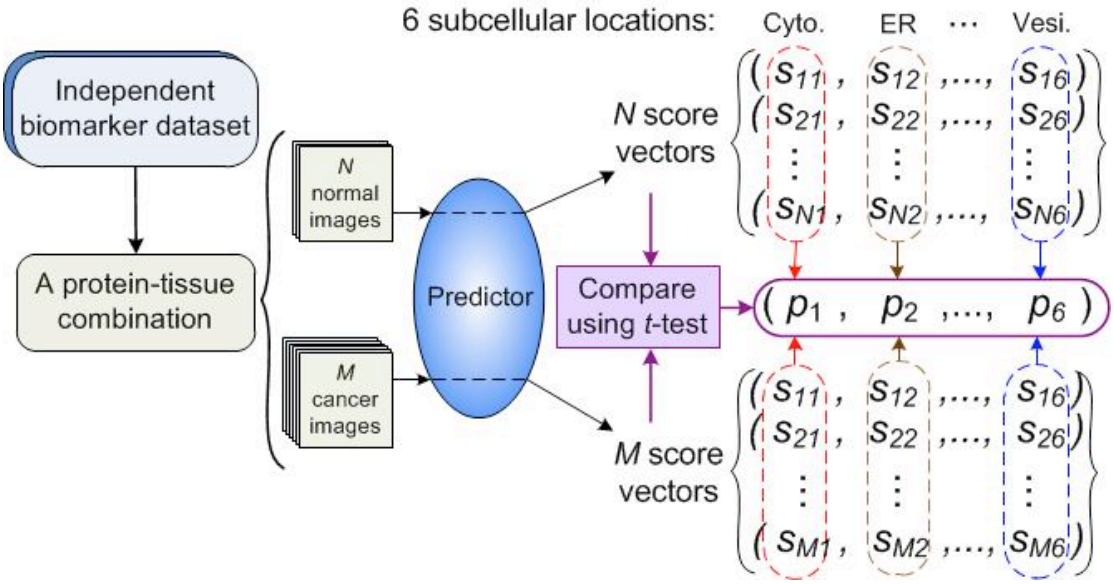
metric							
Sensitivity	single AsemiB	63.59%	66.67%	48.28%	79.25%	81.94%	26.92%
	AsemiB^{E}	64.06%	100%	55.17%	73.58%	85.81%	23.08%
	single AsemiBC	63.59%	66.67%	37.93%	79.25%	81.94%	30.77%
	AsemiBC^{E}	71.43%	100%	51.72%	79.25%	84.52%	34.62%
AUC	single AsemiB	0.6838	0.4492	0.5535	0.7935	0.9115	0.4197
	AsemiB^{E}	0.6921	0.4702	0.5369	0.7982	0.9195	0.3318
	single AsemiBC	0.6925	0.4945	0.5316	0.7919	0.9176	0.4067
	AsemiBC^{E}	0.7176	0.7187	0.5460	0.8085	0.9195	0.4052



Supplementary Fig. 1 Dynamic training process with the stop condition.
If we set $eff_j = 0$, the iteration finally stopped at the 19th iteration, where no more images are added. It is clear that after the 7th iteration, the effects are relatively stable. Besides, the subset accuracy even decreases a little after the 7th iterative round. Considering the stable performance and time saving, we select 0.01 as the threshold of effect to stop further iterations.



Supplementary Fig. 2 Effects of parameter α in dynamic threshold criterion on parameter *θ* and classification performance. We performed a grid search for α from 0.9 to 0.995 on ADN dataset with db1 features. It can be seen that the effects of α on the classification performance are rather slight when α is larger than 0.9.



Supplementary Fig. 3 Flowchart of using *t*-test to measure the significance of translocations. The predictor is either AsemiB^{E} or AsemiBC^{E}.

Reference

Akkiprik,M., et al. (2009) The subcellular localization of IGFBP5 affects its cell growth and migration functions in breast cancer. *BMC Cancer*, **9**, 103.

Bratthauer,G.L. and Vinh,T.N. (2009) Intracellular location of the SOX9 protein in breast disease. *Open Pathol. J.,* **3**, 118-123.

Dhar, K. et al. (1999) Expression and subcellular localization of cyclin D1 protein in epithelial ovarian tumour cells. *Brit. J. Cancer*, **81**, 1174.

Feldhoff,P.W., et al. (1994) Altered subcellular distribution of topoisomerase IIα in a drug-resistant human small cell lung cancer cell line. *Cancer Res.*, **54**, 756-762.

Klugbauer,S., et al. (1998) H.M. Detection of a novel type of RET rearrangement (PTC5) in thyroid carcinomas after Chernobyl and analysis of the involved RET-fused gene RFG5. *Cancer Res.*, **58**, 198-203.

Moll,U.M., et al. (1992) Two distinct mechanisms alter p53 in breast cancer: mutation and nuclear exclusion. *P. Natl. Acad. Sci.*, **89**, 7262-7266.

Nechushtan,A., et al. (1999) Conformation of the Bax C‐terminus regulates subcellular location and cell death. *EMBO J.*, **18**, 2330-2341.

Perren,A. et al. (2000) Mutation and expression analyses reveal differential subcellular compartmentalization of PTEN in endocrine pancreatic tumors compared to normal islet cells. *Am. J. Pathol.*, **157**, 1097-1103.

Runnebaum,I.B., et al. (1996) Subcellular localization of accumulated p53 in ovarian cancer cells. *Gynecol. Oncol.*, **61**, 266-271.

Takayama,S. et al. (1998) Expression and location of Hsp70/Hsc-binding anti-apoptotic protein BAG-1 and its variants in normal tissues and tumor cell lines. *Cancer Res.* **58**, 3116-3131.

Winski,S.L., et al. (2002) Subcellular localization of NAD (P) H: quinone oxidoreductase 1 in human cancer cells. *Cancer Res.*, **62**, 1420-1424.