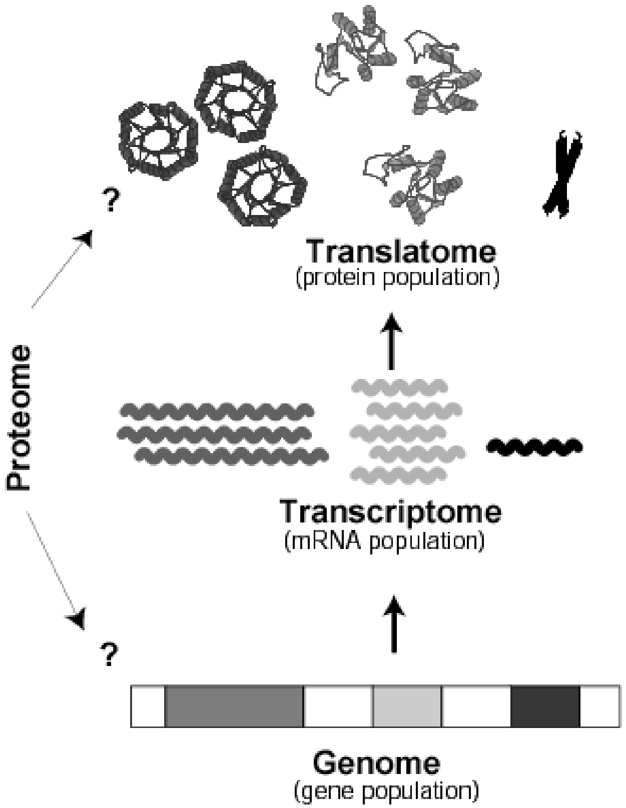
Relating Protein Abundance & mRNA Expression

Mark B Gerstein
Yale (Comp. Bio. & Bioinformatics)

NIDA site visit at Yale
2008.12.03

Why relate amounts of protein & mRNA



[Greenbaum et al. *Bioinformatics* 2002, *18*, 587.]

Gene expression -
major place for **regulation**
(easy to measure)

vs.

Concentration of protein -
major determinant of **activity**

Expectations from simple kinetic models:

$$\frac{dP_i}{dt} = k_{s,i}\,[\mathrm{mRNA}_i] - k_{d,i}\,P_i$$

At steady state: $P_i = \dfrac{k_{s;i}\,[\mathrm{mRNA}_i]}{k_{d,i}}$

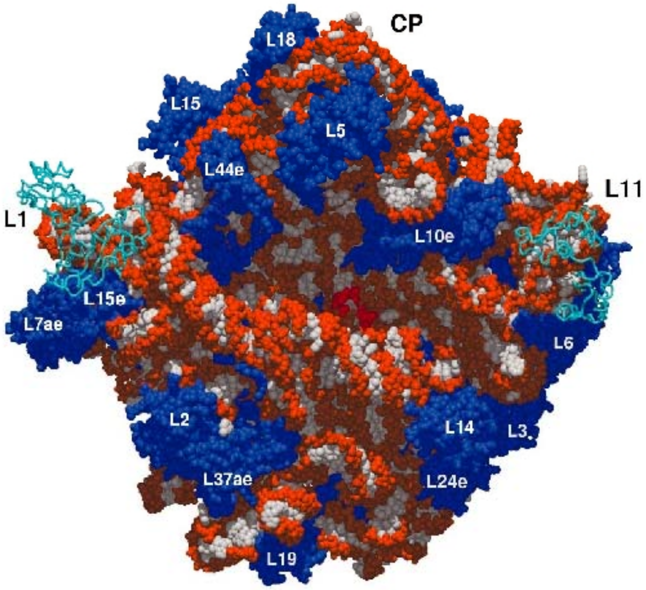
where $k_{s,i}$ and $k_{d,i}$ are the protein synthesis and degradation rate constants

Outliers from trend interesting

Relationship of Protein Abundance to Complexes and Pathways

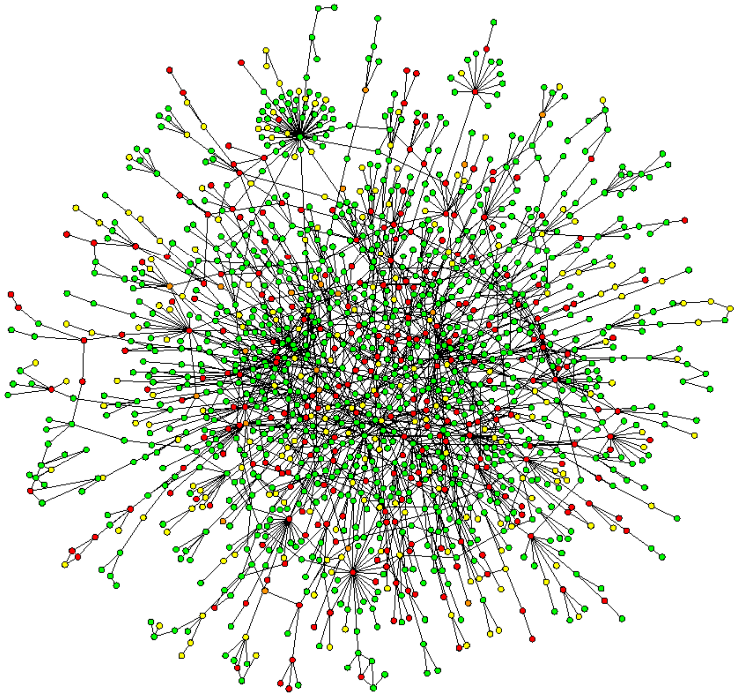
In protein complexes, one expects stoichiometric abundance of component proteins and that mRNA expression levels should be correlated with protein abundance

…Among pathways, this is expected to a lesser degree between interacting proteins



Protein complexes

[Graphic: http://proton.chem.yale.edu]



Protein interaction networks

[Graphic: Jeong et al, Nature, 41:411]

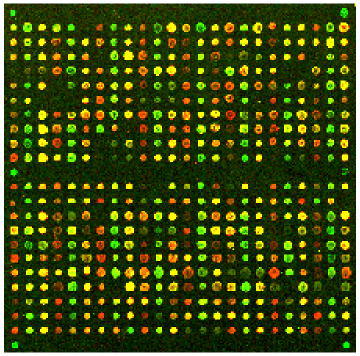
Sources of experimental data

mRNA expression levels

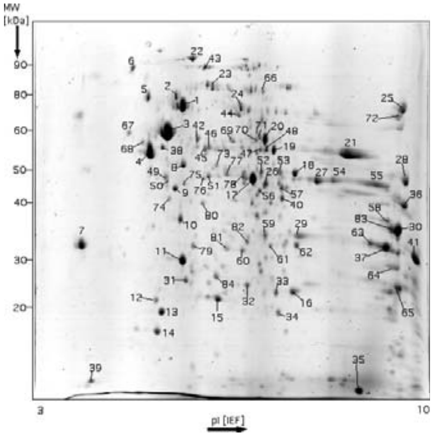
Microarrays

- **Affymetrix**
- **PCR**

SAGE



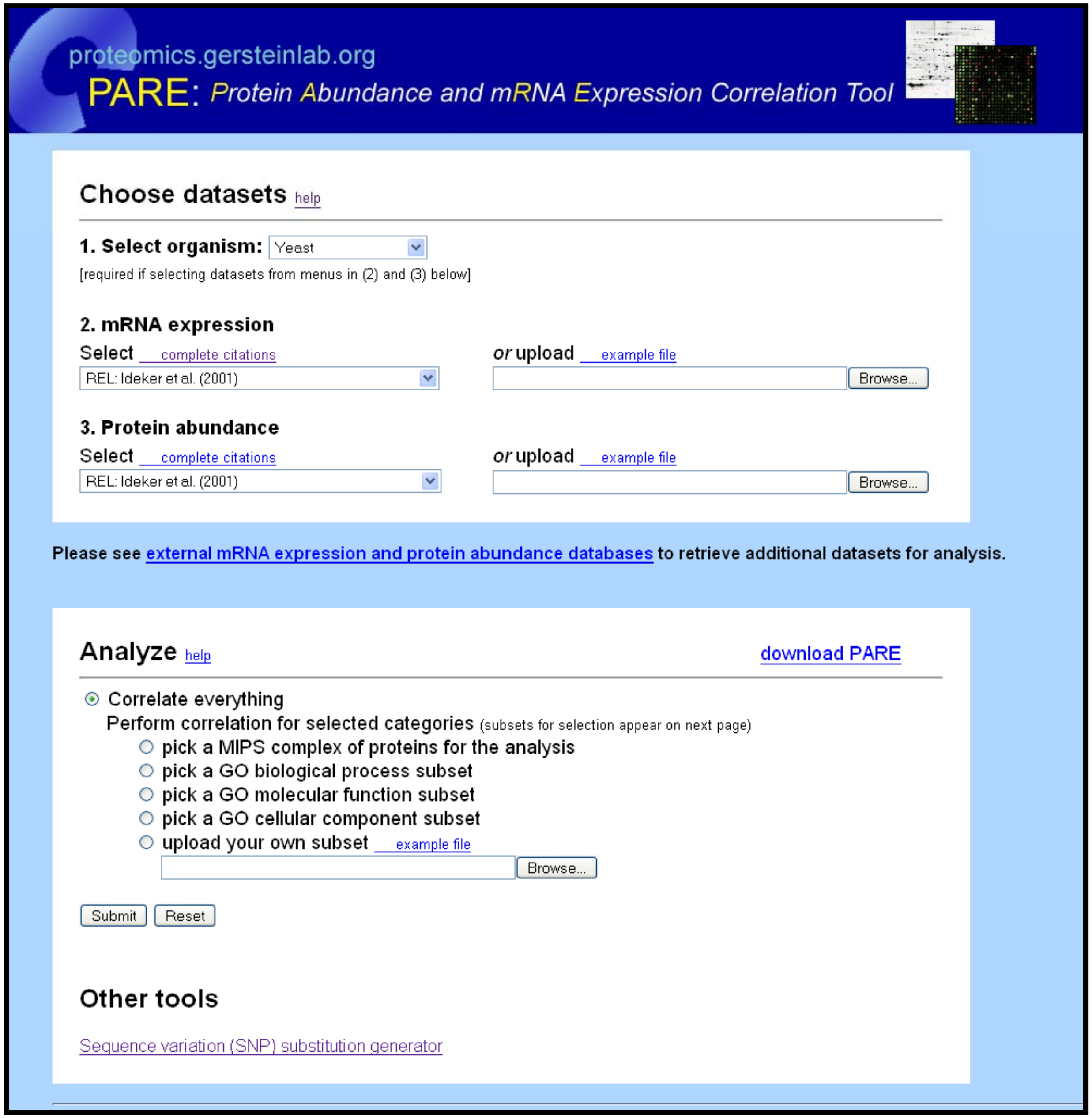
Protein abundance



2D Gel Electrophoresis	Multiple staining options Small dynamic range
DIGE	Cy3 vs. Cy5 labeling Large dynamic range
ICAT, iTRAQ	MS-based Relative abundance (ratio of isotopically labeled species) Large dynamic range
MudPIT	**LC-MS/MS**
SILAC	Stable isotope labeling with amino acids in cell culture - for MS analysis
TAP-Tag	Weissman and O'Shea (Oct. 2003)

[http://www.biology.ucsc.edu/mcd/images/microarray.gif]

PARE: proteomics.gersteinlab.org



proteomics.gersteinlab.org

PARE: *Protein Abundance and mRNA Expression Correlation Tool*

Choose datasets help

1. Select organism: Yeast

[required if selecting datasets from menus in (2) and (3) below]

2. mRNA expression

Select complete citations — REL: Ideker et al. (2001)

or upload example file — Browse...

3. Protein abundance

Select complete citations — REL: Ideker et al. (2001)

or upload example file — Browse...

Please see external mRNA expression and protein abundance databases to retrieve additional datasets for analysis.

Analyze help

download PARE

● Correlate everything

Perform correlation for selected categories (subsets for selection appear on next page)

- ○ pick a MIPS complex of proteins for the analysis
- ○ pick a GO biological process subset
- ○ pick a GO molecular function subset
- ○ pick a GO cellular component subset
- ○ upload your own subset example file — Browse...

Submit Reset

Other tools

Sequence variation (SNP) substitution generator

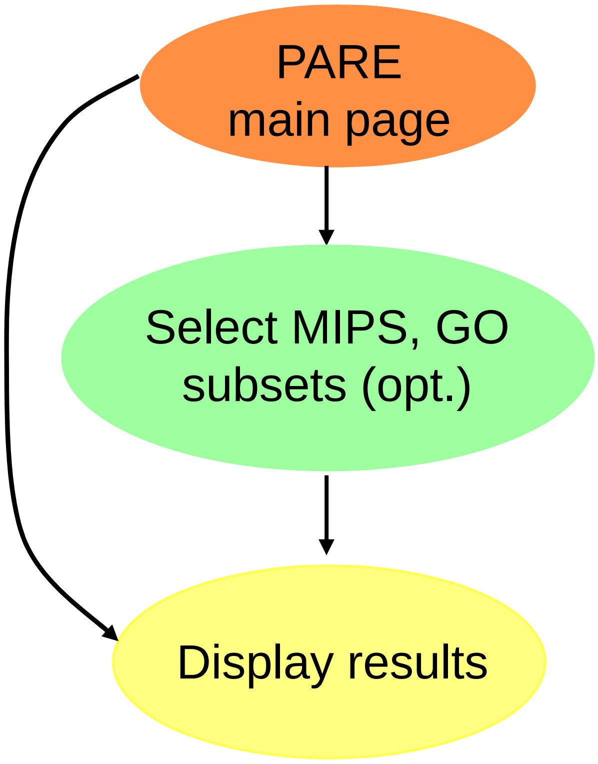
Upload or use pre-loaded mRNA, protein datasets

Open-source code
Downloadable

Analyze all or
analyze MIPS or GO subset

[Yu et al., BMC Bioinfo. '07]

PARE: a web-based tool for correlating mRNA expression and protein abundance



PARE
main page

Select MIPS, GO
subsets (opt.)

Display results

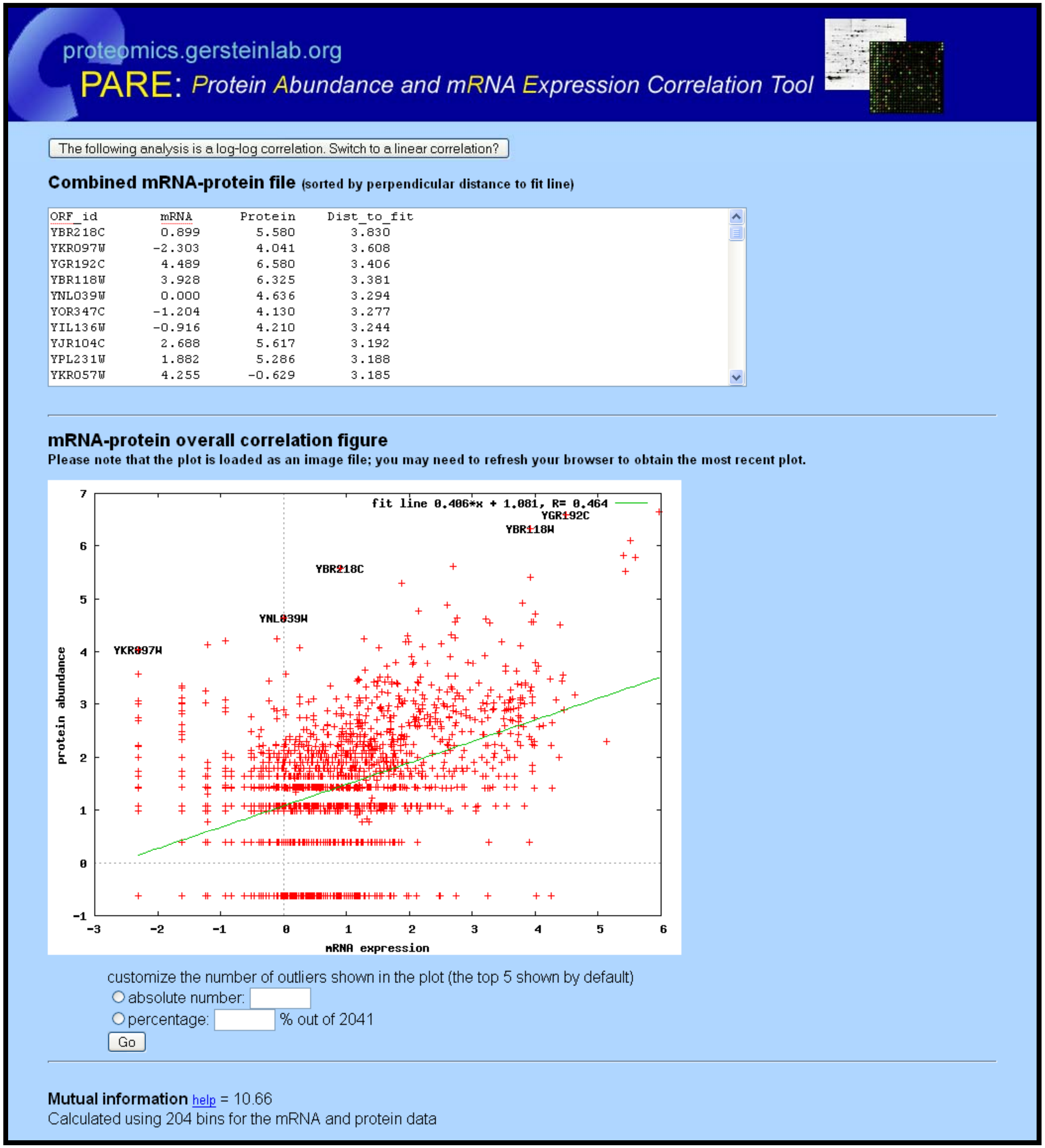
(1) Select mRNA, protein datasets:
- use pre-loaded datasets
- upload datasets

(2) Choose categorization method:
- correlate all
- MIPS complexes
- GO biological processes
- GO molecular function
- GO cellular component

(3) Display
- Linear or log-log correlation for selected subset(s)
- Tabulate data, correlation values for selected subset(
- Label (on plot) and tabulate outlying datapoints

[Yu et al., BMC Bioinfo. '07]

PARE output



PARE: *Protein Abundance and mRNA Expression Correlation Tool*

Combined mRNA-protein file (sorted by perpendicular distance to fit line)

```
ORF_id         mRNA      Protein     Dist_to_fit
YBR218C        0.899       5.580        3.830
YKR097W       -2.303       4.041        3.608
YGR192C        4.489       6.580        3.406
YBR118W        3.928       6.325        3.381
YNL039W        0.000       4.636        3.294
YOR347C       -1.204       4.130        3.277
YIL136W       -0.916       4.210        3.244
YJR104C        2.688       5.617        3.192
YPL231W        1.882       5.286        3.188
YKR057W        4.255      -0.629        3.185
```

mRNA-protein overall correlation figure

Please note that the plot is loaded as an image file; you may need to refresh your browser to obtain the most recent plot.

customize the number of outliers shown in the plot (the top 5 shown by default)

○ absolute number:

○ percentage: % out of 2041

Mutual information help = 10.66

Calculated using 204 bins for the mRNA and protein data

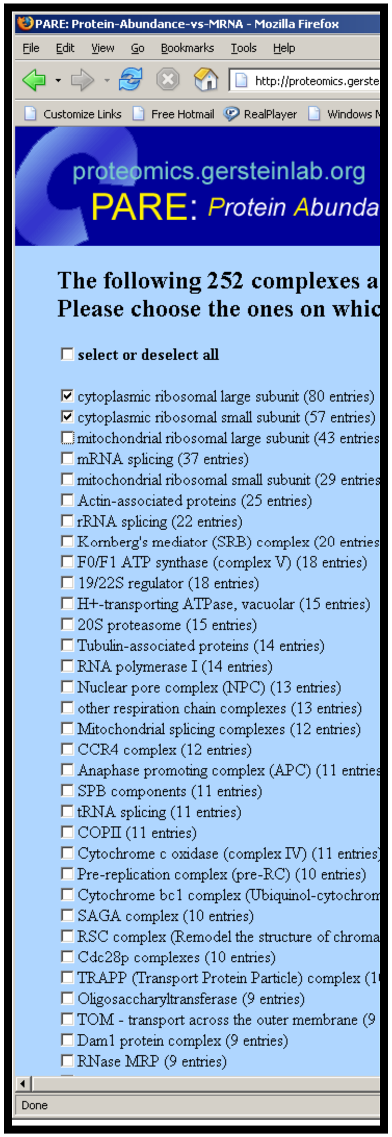
Correlated data

Log-log plot of correlation
- linear fit
- outliers labeled

Calculation of mutual information

[Yu et al., BMC Bioinfo. '07]

Correlation of subsets (GO, MIPS)

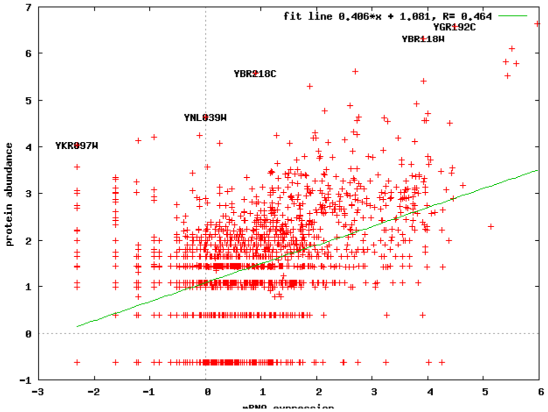


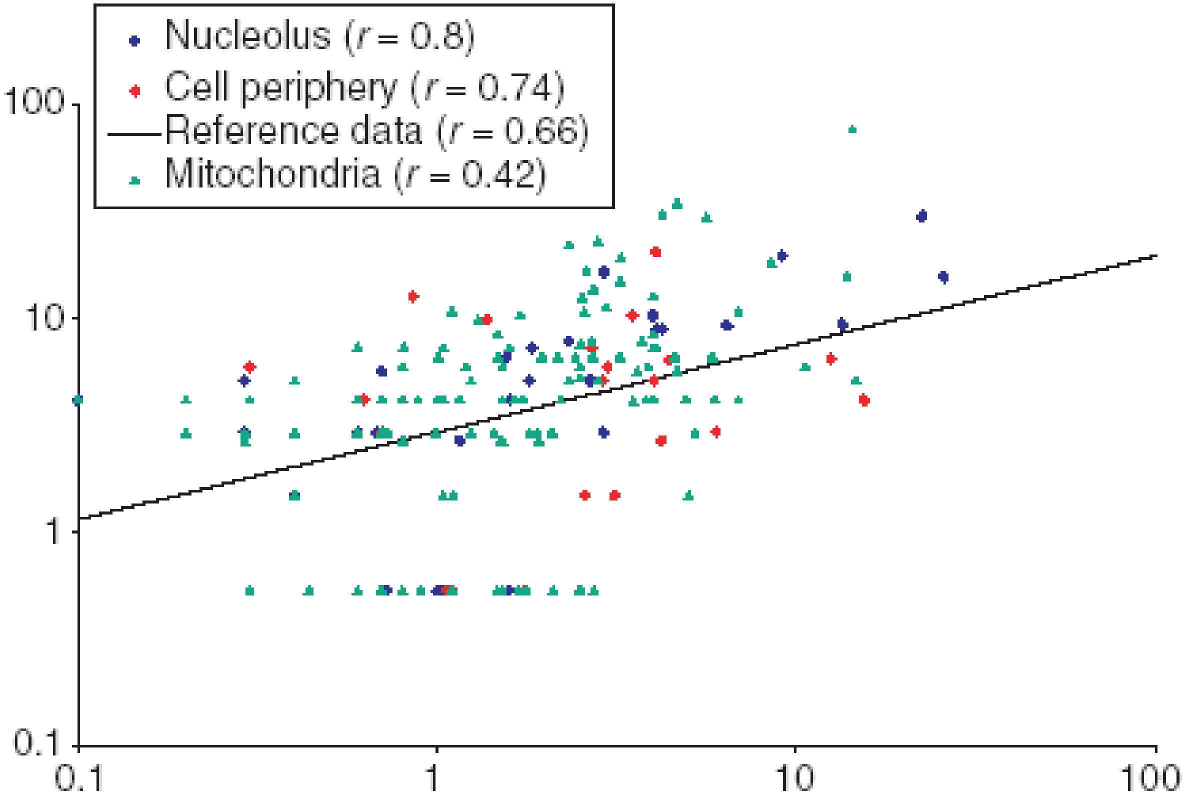
The following 252 complexes a
Please choose the ones on whic

☐ select or deselect all

☑ cytoplasmic ribosomal large subunit (80 entries)
☑ cytoplasmic ribosomal small subunit (57 entries)
☐ mitochondrial ribosomal large subunit (43 entries
☐ mRNA splicing (37 entries)
☐ mitochondrial ribosomal small subunit (29 entries
☐ Actin-associated proteins (25 entries)
☐ rRNA splicing (22 entries)
☐ Kornberg's mediator (SRB) complex (20 entries
☐ F0/F1 ATP synthase (complex V) (18 entries)
☐ 19/22S regulator (18 entries)
☐ H+-transporting ATPase, vacuolar (15 entries)
☐ 20S proteasome (15 entries)
☐ Tubulin-associated proteins (14 entries)
☐ RNA polymerase I (14 entries)
☐ Nuclear pore complex (NPC) (13 entries)
☐ other respiration chain complexes (13 entries)
☐ Mitochondrial splicing complexes (12 entries)
☐ CCR4 complex (12 entries)
☐ Anaphase promoting complex (APC) (11 entries
☐ SPB components (11 entries)
☐ tRNA splicing (11 entries)
☐ COPII (11 entries)
☐ Cytochrome c oxidase (complex IV) (11 entries
☐ Pre-replication complex (pre-RC) (10 entries)
☐ Cytochrome bc1 complex (Ubiquinol-cytochrom
☐ SAGA complex (10 entries)
☐ RSC complex (Remodel the structure of chroma
☐ Cdc28p complexes (10 entries)
☐ TRAPP (Transport Protein Particle) complex (1
☐ Oligosaccharyltransferase (9 entries)
☐ TOM - transport across the outer membrane (9
☐ Dam1 protein complex (9 entries)
☐ RNase MRP (9 entries)

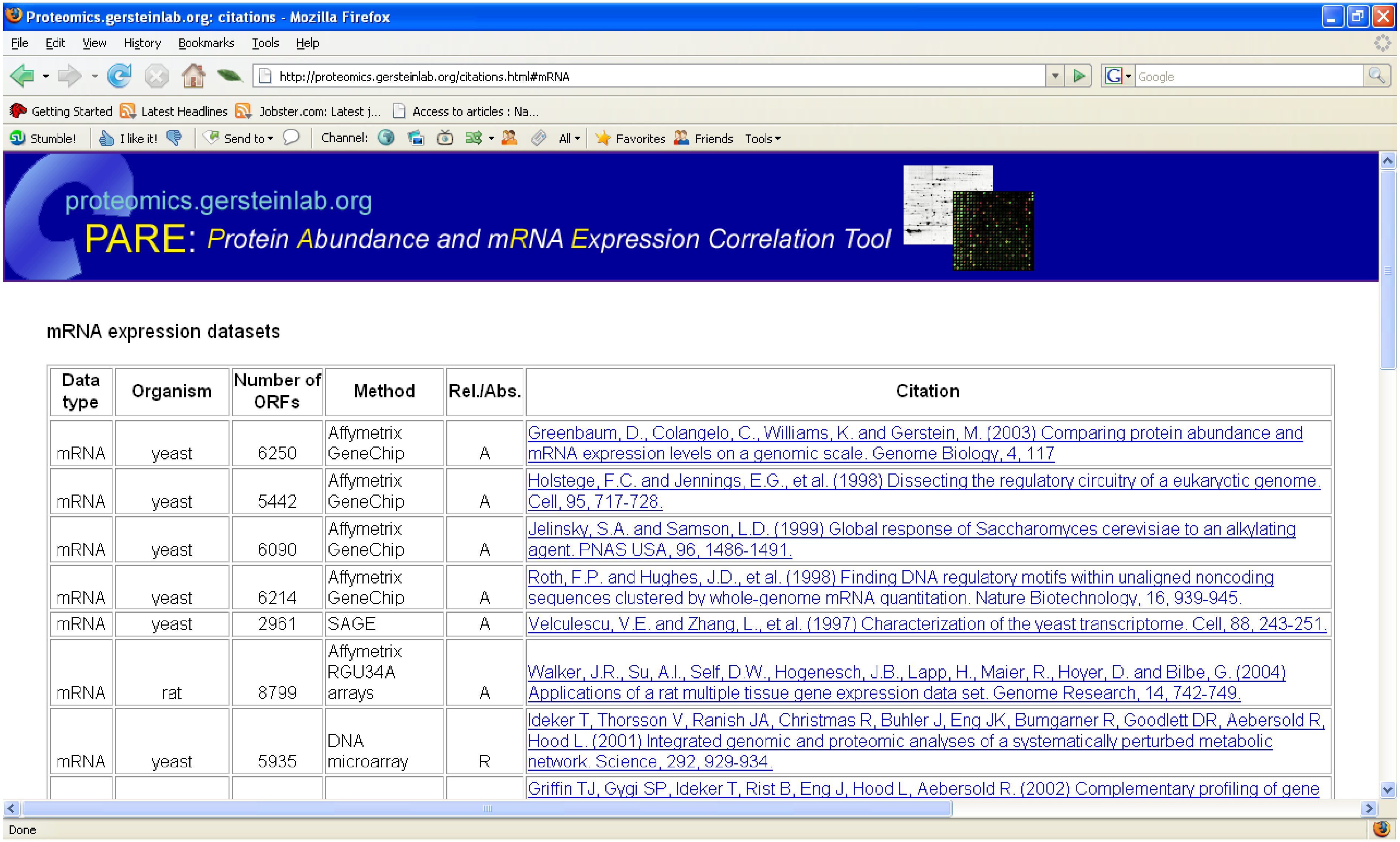
Yeast ref. datasets:
“Correlate all”
vs.
GO cellular component subsets
for particular cellular locations





[Yu et al., BMC Bioinfo. '07]

PARE: pre-loaded datasets



proteomics.gersteinlab.org

PARE: *Protein Abundance and mRNA Expression Correlation Tool*

mRNA expression datasets

Data type	Organism	Number of ORFs	Method	Rel./Abs.	Citation
mRNA	yeast	6250	Affymetrix GeneChip	A	Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biology, 4, 117
mRNA	yeast	5442	Affymetrix GeneChip	A	Holstege, F.C. and Jennings, E.G., et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell, 95, 717-728.
mRNA	yeast	6090	Affymetrix GeneChip	A	Jelinsky, S.A. and Samson, L.D. (1999) Global response of Saccharomyces cerevisiae to an alkylating agent. PNAS USA, 96, 1486-1491.
mRNA	yeast	6214	Affymetrix GeneChip	A	Roth, F.P. and Hughes, J.D., et al. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology, 16, 939-945.
mRNA	yeast	2961	SAGE	A	Velculescu, V.E. and Zhang, L., et al. (1997) Characterization of the yeast transcriptome. Cell, 88, 243-251.
mRNA	rat	8799	Affymetrix RGU34A arrays	A	Walker, J.R., Su, A.I., Self, D.W., Hogenesch, J.B., Lapp, H., Maier, R., Hoyer, D. and Bilbe, G. (2004) Applications of a rat multiple tissue gene expression data set. Genome Research, 14, 742-749.
mRNA	yeast	5935	DNA microarray	R	Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 292, 929-934.
					Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. (2002) Complementary profiling of gene

[Yu et al., BMC Bioinfo. '07]

Connecting PARE with datasets from NIDA investigators

Rat (Nairn lab) - samples from 3 brain regions:
cortex, striatum, hippocampus

Protein abundance - MudPIT datasets
(protein profiling core)
samples for each brain region pooled from 4 rats

mRNA expression - Affymetrix microarray
(microarray core)
microarrays for brain regions for each of
4 rat individuals + replicate measurements (13 datasets)

Mouse (Nairn lab) - 3 brain regions
only iTRAQ data available; no matched mRNA datasets

Rat brain: mRNA expression

mRNA samples

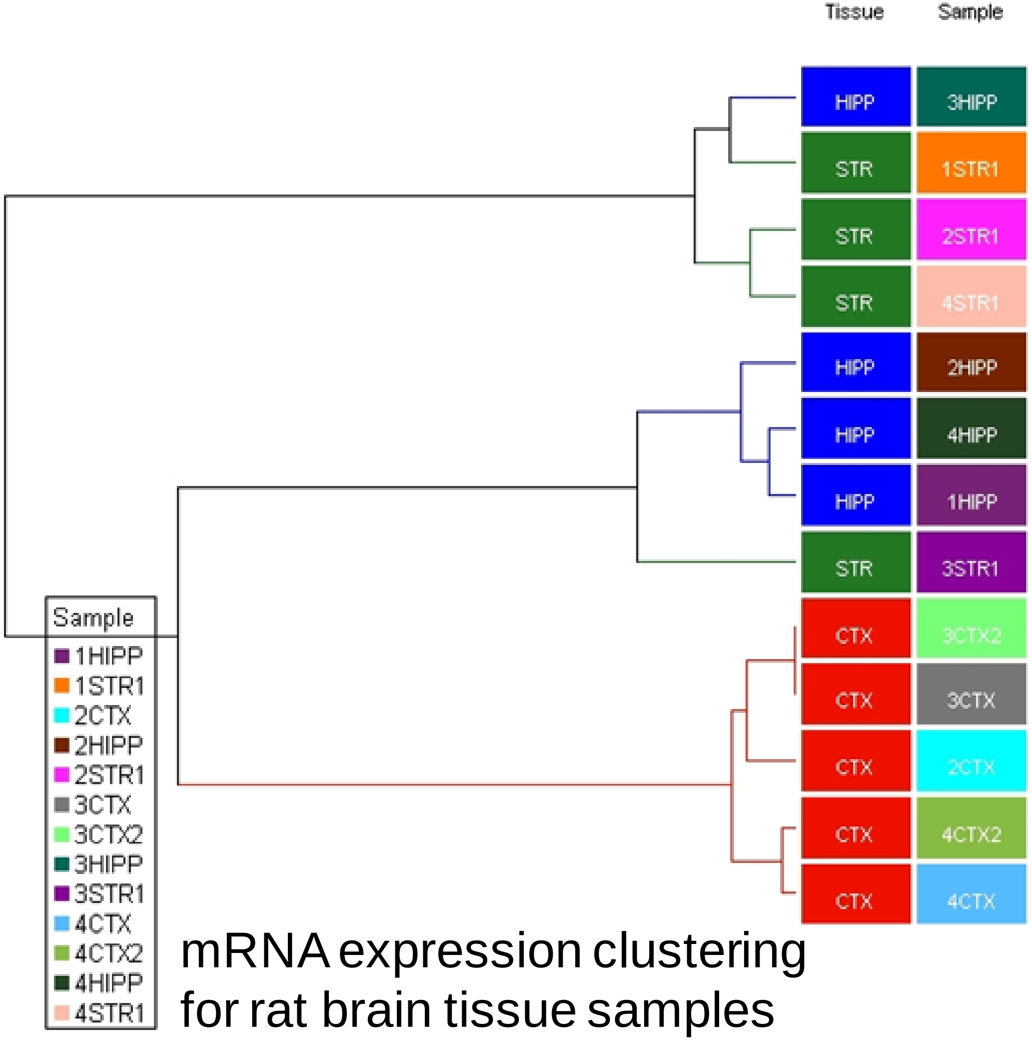
4 individual rats

CTX: indiv. 2, 3, 4 (+2 replicates each for 3, 4)

HIPP: indiv. 1-4

STR: indiv. 1-4

28,894 probes



mRNA expression clustering
for rat brain tissue samples

Rat brain: protein abundance

Rat brain
3 tissues: Cortex (CTX), striatum (STR), hippocampus (HIPP)

Number of matched mRNA/proteins pairs between Affymetrix and MudPIT data:

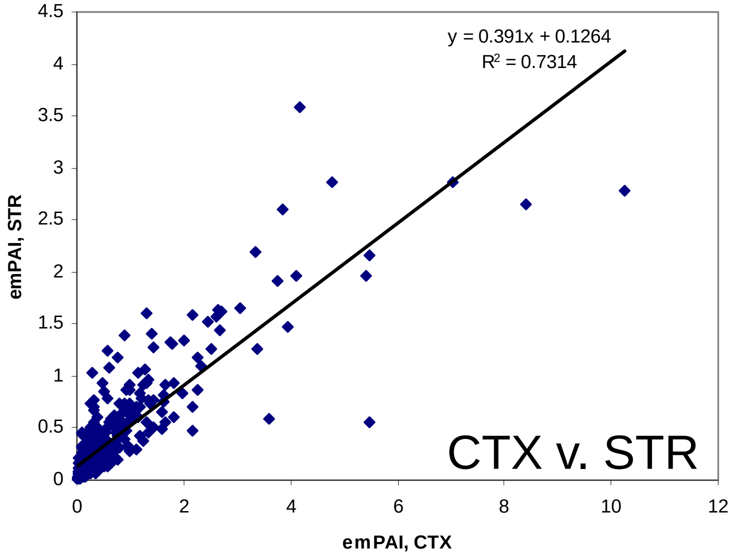
CTX:	830 proteins
STR:	926 proteins
HIPP:	701 proteins

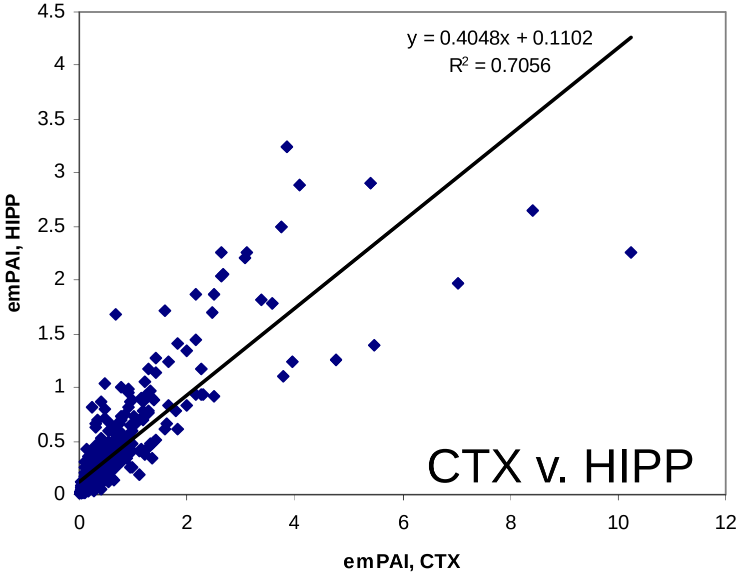
Number of proteins observed in 2 or more tissues (for MudPIT data):

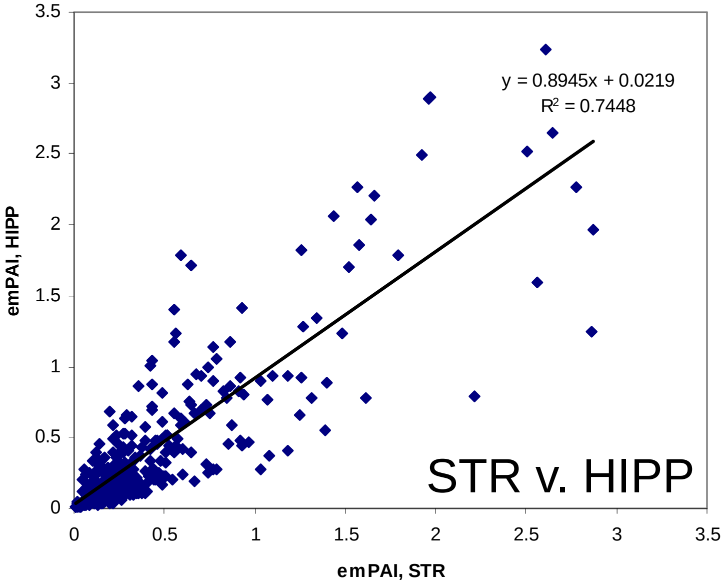
CTX and STR:	517 proteins
CTX and HIPP:	538 proteins
STR and HIPP:	485 proteins
CTX, STR, and HIPP:	421 proteins

Rat brain: protein abundance

emPAI correlations for proteins observed in multiple tissues

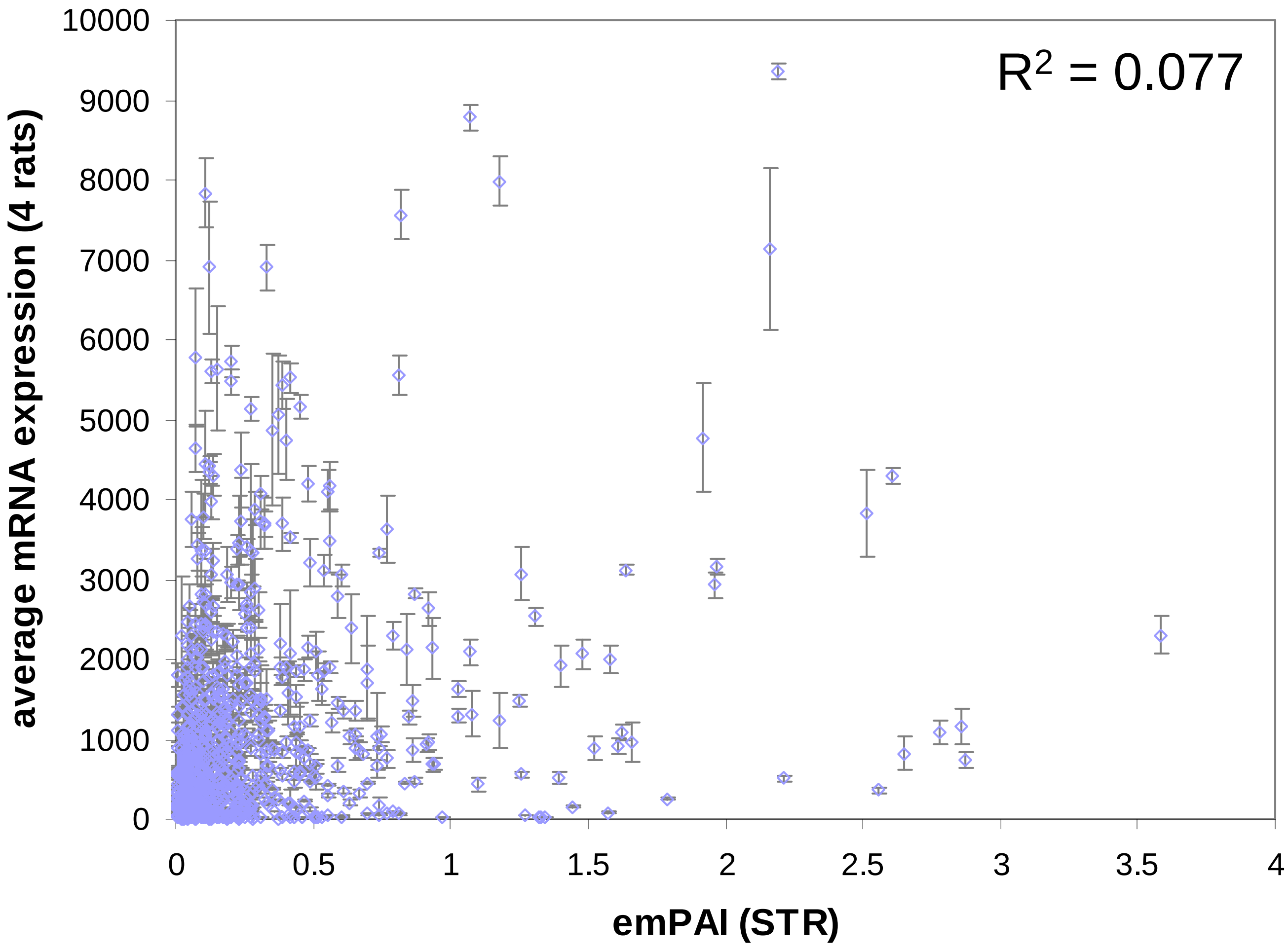






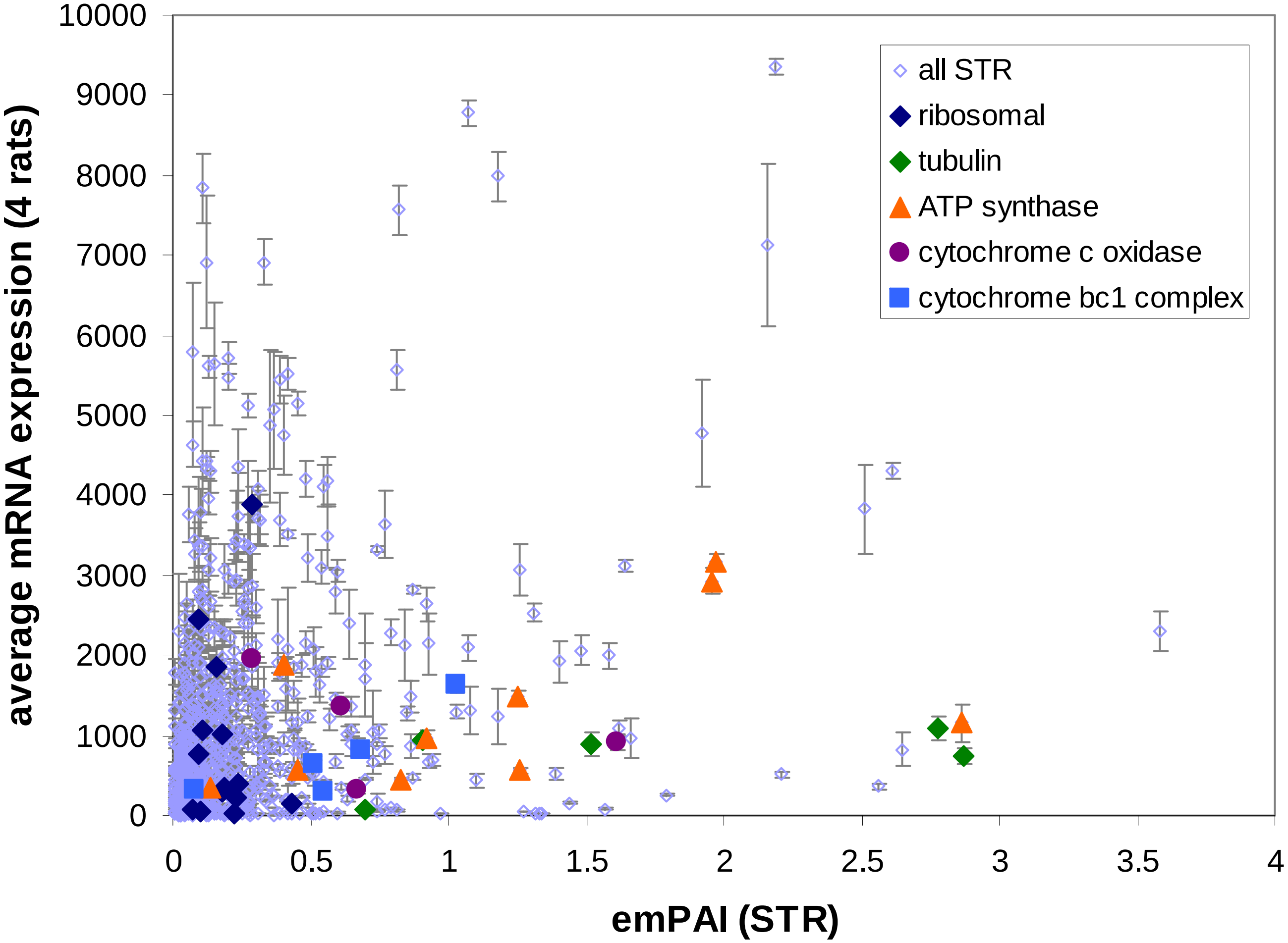
emPAI = exponentially modified protein abundance index,
metric proportional to protein abundance
(Ishihama et al. Mol. Cell. Proteomics (2005) 4, 1265-1272.)

Rat brain: correlation of mRNA expression and protein abundance



Rat brain: emPAI(STR) v. *average* mRNA expression (4 rats: 1STR to 4STR)
926 proteins matched with corresponding mRNA

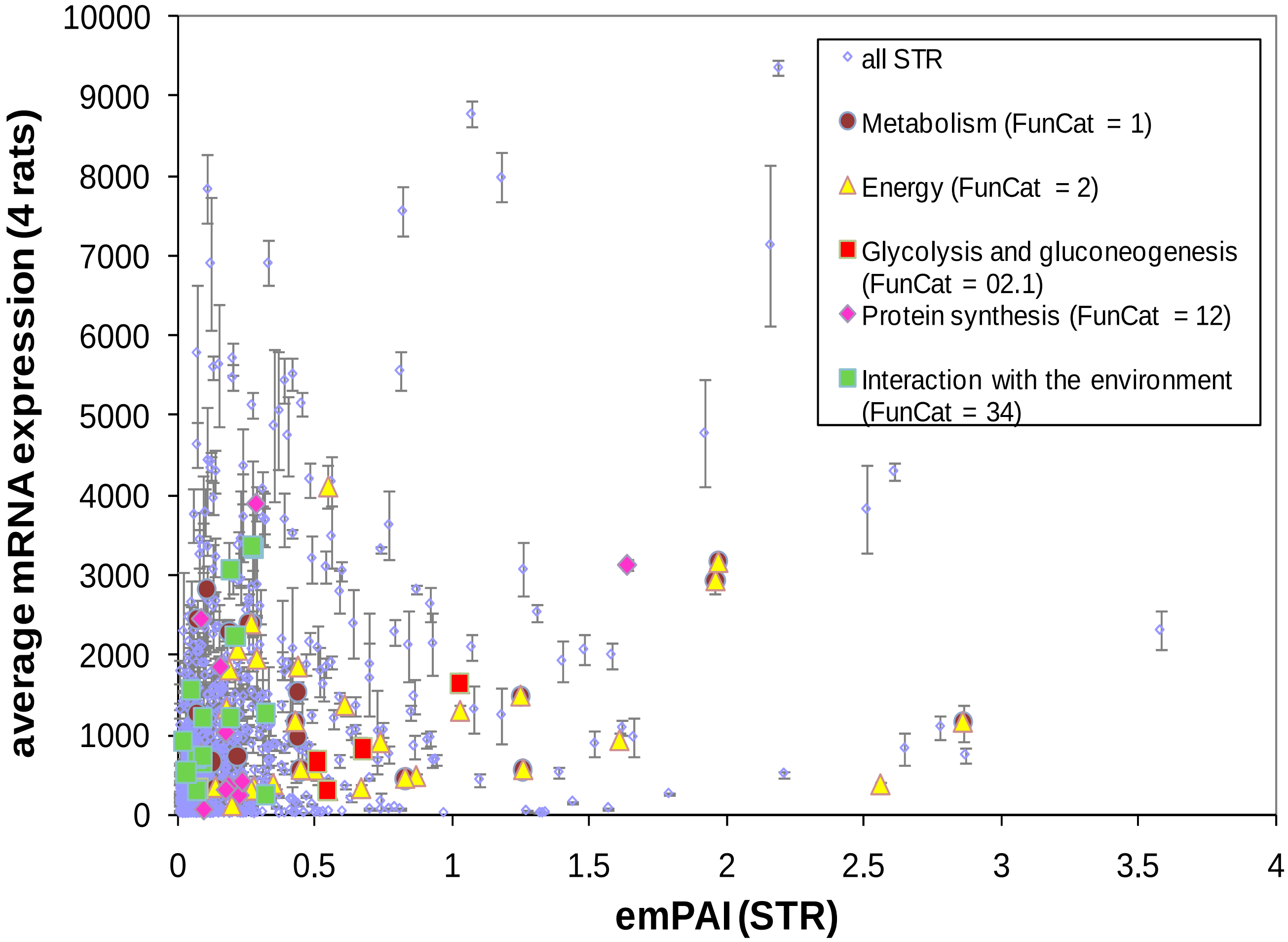
Rat brain: mRNA expression and protein abundance for protein complexes



-Ribosomal proteins observed for STR appear to have low emPAI and a wide range of mRNA expression.

-Proteins that comprise tubulin have a wide range of emPAI (including 2 of the 4 highest values), but a narrow range of mRNA expression.

Rat brain: mRNA expression and protein abundance for functional categories

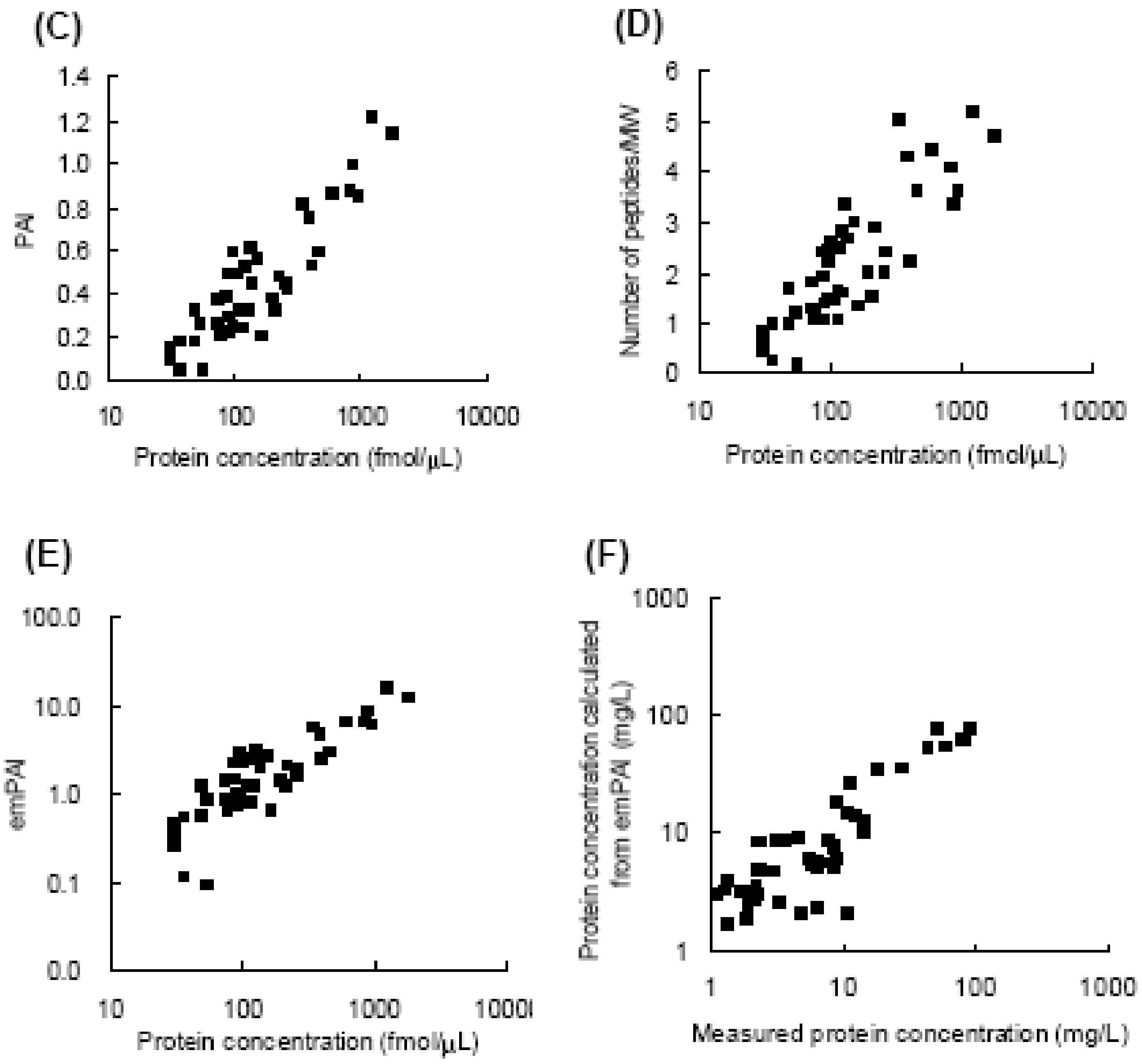


Acknowledgements

MudPIT datasets:
Angus Nairn, Erika Andrade
Kathy Stone, Chris Colangelo, Mark Shifman (YPED)

mRNA datasets:
Shrikant Mane, Aiping (Amy) Lin

PARE:
Anne Burba, Eric Yu



Ishihama et al. Mol. Cell. Proteomics (2005) 4, 1265-1272.