www.sciencemag.org/cgi/content/full/1174301/DC1

Supporting Online Material for

Positive Selection of Tyrosine Loss in Metazoan Evolution

Chris Soon Heng Tan, Adrian Pasculescu, Wendell A. Lim, Tony Pawson,*
Gary D. Bader,* Rune Linding*

*To whom correspondence should be addressed. E-mail: pawson@lunenfeld.ca (T.P.); gary.bader@utoronto.ca (G.D.B.); linding@icr.ac.uk (R.L.)

Published 9 July 2009 on *Science* Express
DOI: 10.1126/science.1174301

This PDF file includes:

Supplementary Information for:
Positive Selection of Tyrosine Loss in Metazoan Evolution

Chris Soon Heng Tan[1,2,3], Adrian Pasculescu[1], Wendell A. Lim[4], Tony Pawson[1,2†], Gary D. Bader[1,2,3†] and Rune Linding[5†]

[1] Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto, Canada

[3] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada

[4] Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, USA

[5] Cellular & Molecular Logic Team, Section of Cell and Molecular Biology, The Institute of Cancer Research (ICR), SW3 6JB, London, UK

[†] Correspondence should be addressed by e-mail to: pawson@lunenfeld.ca, gary.bader@utoronto.ca and linding@icr.ac.uk

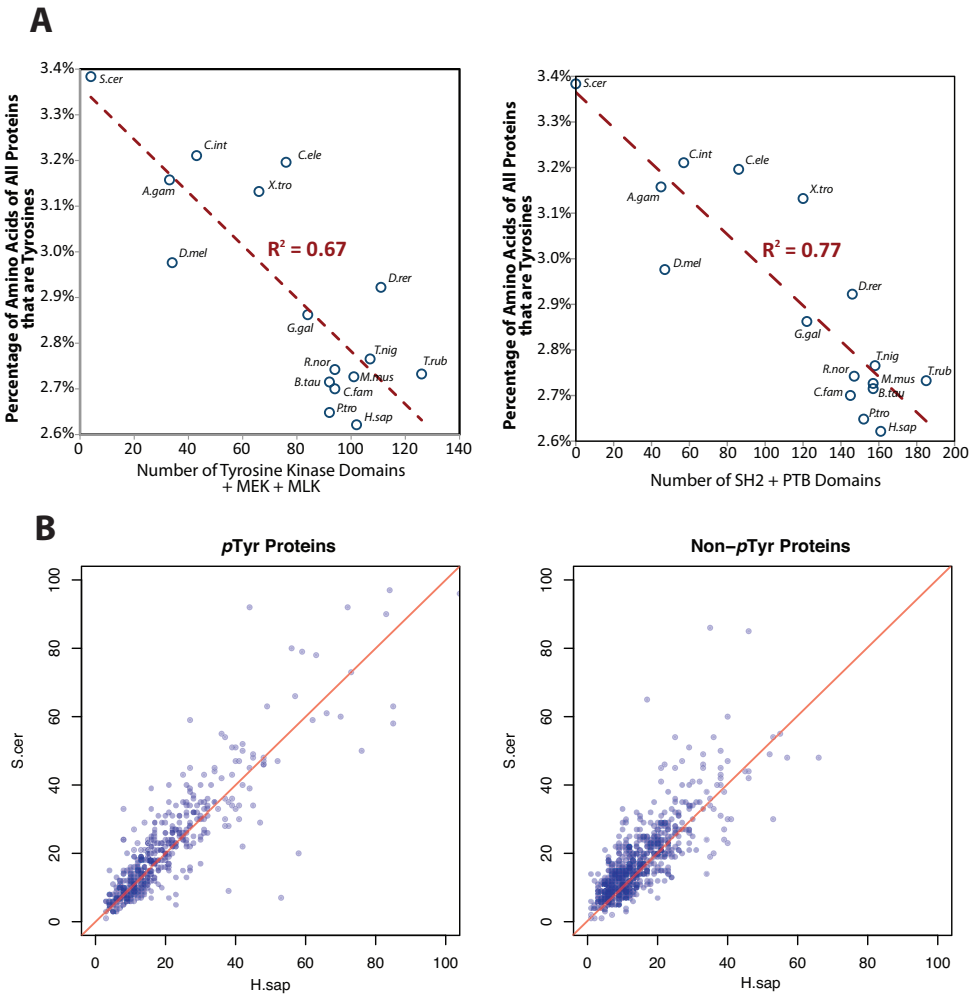

Figure S1: **Correlation of expansion of phospho-tyrosine signaling systems with loss of genome encoded tyrosine residues. A**, MEK and MLK serine/threonine kinases can phosphorylate tyrosines to some degree. Including these kinases in the correlation analysis doesnot significantly change the correlation. We observed a negative correlation of proteome tyrosine content with the number of SH2 and PTB *p*Tyr binding domains. **B**, The portion of amino acids that are tyrosine in orthologous pairs of human and yeast proteins. Every point in the scatter plot represents a human-yeast ortholog pair. Darker spots represent multiple orthologous protein pairs with same tyrosine count. Only proteins with inferred one-to-one human-yeast orthologous relationships were analyzed to avoid biases due to accelerated sequence divergence due to functional redundancy of duplicated genes. Orthologous protein pairs lying above the red diagonal lines ($x = y$) have more tyrosine in yeast than human. The left scatter plot is for 437 human proteins conserved in yeast and known to be tyrosine-phosphorylated and the right plot is for 647 yeast-conserved human proteins not experimentally determined to be tyrosine-phosphorylated. Human Non-*p*Tyr proteins have less tyrosines than human *p*Tyr proteins compared to their orthologous counterparts in yeast (approximate P-value = 7.9×10^{-5}, Mann-Whitney test).

Materials & Methods

Identifying human-yeast orthologous proteins

All known and predicted human and yeast protein sequences were retrieved from Ensembl (release 51) (1) and processed to retain only the longest translation of each gene. Human-yeast orthologous proteins were then inferred using the Inparanoid algorithm (2) and the downloaded sequences, based on stringent bi-directional best BLAST (3) hits with the processed human and yeast protein sequences. To avoid interference from relaxation in evolutionary constraints due to functional redundancy of duplicated genes (paralogs), analysis was restricted to one-to-one human-yeast orthologs.

Collection of experimentally determined human phospho-tyrosines

Experimentally determined phospho-tyrosine sites in human proteins were obtained from the Phospho.ELM (4) and PhosphoSitePlus (5) databases in November 2008 and mapped to the human protein sequences described above. In total, the dataset contains 12659 phospho-tyrosines in 6450 proteins. A human protein is classified as tyrosine-phosphorylated (*p*Tyr) if any of its tyrosines is phosphorylated in our assembled phosphorylation data, or otherwise classified as Non-*p*Tyr protein.

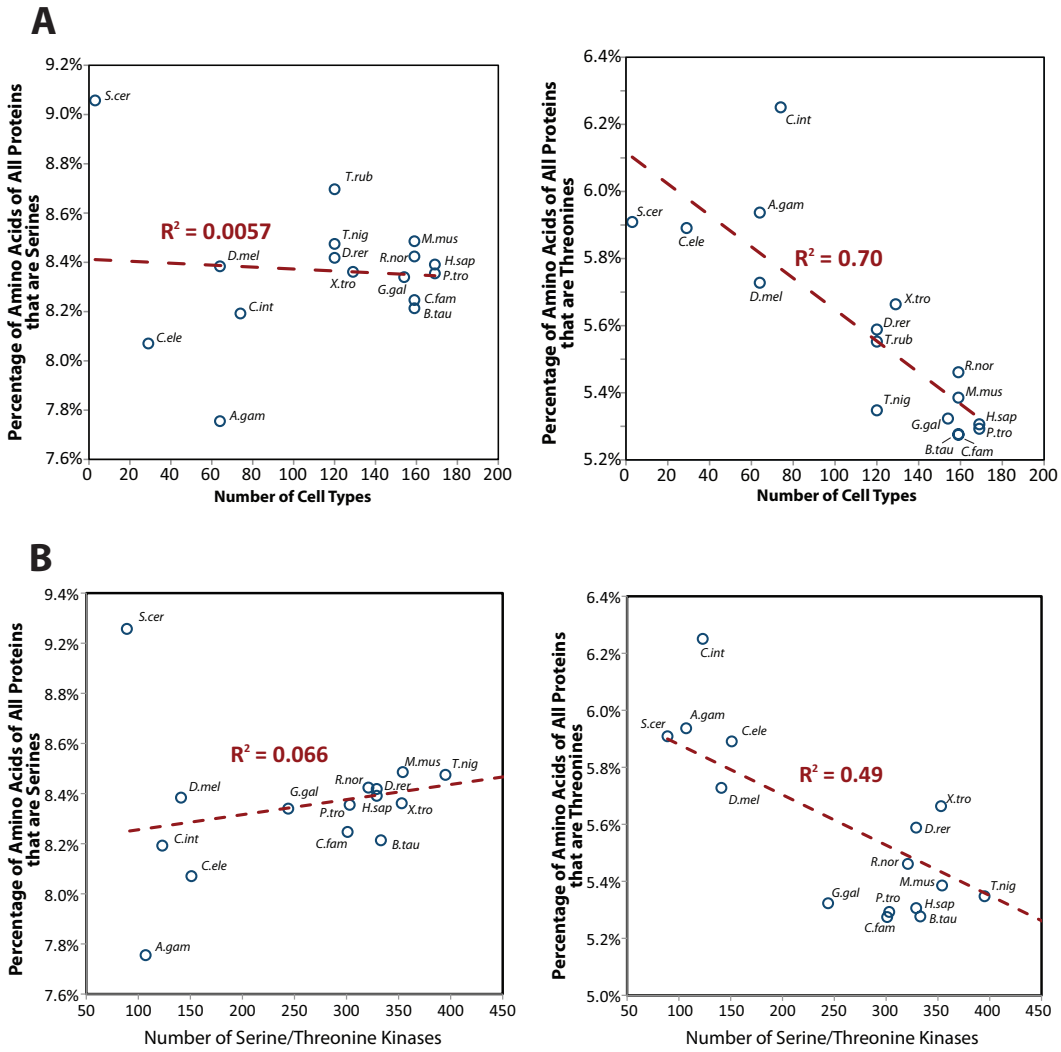

Figure S2: **Correlation of expansion of phospho-serine/threonine signaling systems with loss of genome encoded serine/threonine residues. A**, Number of cell types correlate negatively and significantly with genomic threonine content but not for genomic serine. **B**, Genomically encoded threonine content correlate negatively with number of inferred serine/threonine kinases but this trend was not observed for genomically encoded serine.

Computing the content of tyrosines and other amino-acids in non-redundant protein sequences

For each human gene product, all known and predicted splice variants were aligned using the AMAP multiple sequence alignment software (6), and non-redundant amino acid residue counts were computed from the alignments. We found no substantial difference in tyrosine residue counts using an alternative approach considering only the longest translation of each human gene. Genes coding 200 amino acids or less are excluded from computation to reduce sizable, but non-significant percentage changes in tyrosine content due to small protein size.

Detecting putative tyrosine kinase, serine/threonine kinases, MEK and MLK kinases in different metazoan organisms

Known and inferred protein sequences in the 15 metazoan species and budding yeast were obtained from Ensembl (release 51), processed to retain only the longest translation of each gene. Monosiga brevicolliss protein sequences were obtained from http://genome.jgi-psf.org/Monbr1/Monbr1.home.html (7). Tyrosine kinase domains were detected using HMM models from SMART (8) and Pfam (9) databases using the text-mode pipeline of SMART. Inferred orthologs of human MEK and MLK kinases across the 16 species were retrieved from Ensembl/Compara for the analysis in Fig. S1A (10).

Table S1: **Correlation of all 20 amino acids with distinct cell type. Amino acids are sorted according to increasing Pearsons correlation.**

Amino Acid	Pearson	p-value	Spearman	p-value
D	-0.93	1.65×10^{-7}	-0.93	1.96×10^{-7}
N	-0.92	5.73×10^{-7}	-0.90	3.75×10^{-6}
Y	-0.89	3.99×10^{-6}	-0.89	2.97×10^{-6}
I	-0.87	1.43×10^{-5}	-0.77	4.75×10^{-4}
T	-0.84	5.27×10^{-5}	-0.85	3.74×10^{-5}
F	-0.67	4.58×10^{-3}	-0.36	1.65×10^{-1}
K	-0.58	1.74×10^{-2}	-0.17	5.35×10^{-1}
M	-0.49	5.52×10^{-2}	-0.55	2.75×10^{-2}
S	-0.076	7.81×10^{-1}	0.061	8.22×10^{-1}
V	0.12	6.54×10^{-1}	-0.27	3.21×10^{-1}
A	0.48	$6.20\text{-}\times10^{-2}$	0.36	1.73×10^{-1}
Q	0.62	1.09×10^{-2}	0.61	1.14×10^{-2}
R	0.64	7.63×10^{-3}	0.37	1.61×10^{-1}
H	0.72	1.80×10^{-3}	0.46	7.52×10^{-2}
C	0.78	4.07×10^{-4}	0.40	1.30×10^{-1}
L	0.80	2.20×10^{-4}	0.83	6.66×10^{-5}
G	0.81	1.54×10^{-4}	0.70	2.43×10^{-3}
E	0.83	7.89×10^{-5}	0.83	7.47×10^{-5}
W	0.84	4.81×10^{-5}	0.85	3.45×10^{-5}
P	0.90	3.27×10^{-6}	0.90	3.75×10^{-6}

Statistical Analysis

All statistical tests were performed using the **R** statistical package. Differences in tyrosine content of human-yeast orthologous protein pairs as percentage of all amino acids and absolute tyrosine count were computed and distributions of the computed differences for the pY and Non-*p*Tyr proteins are compared using the Mann-Whitney test.

References and Notes

[1] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. GrŁf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. KŁhŁri, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, S. Searle, Ensembl 2008. *Nucleic Acids Res* **36**, D707–D714 (2008).

[2] M. Remm, C. E. V. Storm, E. L. L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparison. *J. Mol. Biol.* **314**, 1041–1052 (2001).

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

[4] F. Diella, C. M. Gould, C. Chica, A. Via, T. J. Gibson, Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucleic Acids Res.* **36**, D240–D244 (2008).

[5] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, B. Zhang, PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551–1561 (2004).

[6] A. S. Schwartz, L. Pachter, Multiple alignment by sequence annealing. *Bioinformatics* **23**, e24–e29 (2007).

[7] N. King, M. J. Westbrook, S. L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K. J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J. B. Lyons, A. Morris, S. Nichols, D. J. Richter, A. Salamov, J. G. I. Sequencing, P. Bork, W. A. Lim, G. Manning, W. T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I. V. Grigoriev, D. Rokhsar, The genome of the choanoflagellate monosiga brevicollis and the origin of metazoans. *Nature* **451**, 783–788 (2008).

[8] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, P. Bork, SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260 (2006).

[9] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, The pfam protein families database. *Nucleic Acids Res* **36**, D281–D288 (2008).

[10] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney, Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327–335 (2009).

[11] We thank Claus Jørgensen, Jiangzhi Zhang, Karen Colwill, Jing Jin and Kresten Lindorff-Larsen for suggestions and fruitful discussions. This project was in part supported by Genome Canada through the Ontario Genomics Institute and the Canadian Institutes of Health Research (MOP-84324). C.S.H.T. conceived the project. C.S.H.T., W.A.L., G.D.B., T.P. and R.L. designed the experiments. C.S.H.T., R.L. and A.P. performed the experiments. C.S.H.T., G.D.B., W.A.L., T.P. and R.L. wrote the paper. R.L. managed the project.