A Probabilistic Functional Network of Yeast Genes
Insuk Lee, *et al.*
Science **306**, 1555 (2004);
DOI: 10.1126/science.1099511

A Probabilistic Functional Network of Yeast Genes

Insuk Lee,[1] Shailesh V. Date,[1]* Alex T. Adai,[1]†
Edward M. Marcotte[1,2]‡

A conceptual framework for integrating diverse functional genomics data was developed by reinterpreting experiments to provide numerical likelihoods that genes are functionally linked. This allows direct comparison and integration of different classes of data. The resulting probabilistic gene network estimates the functional coupling between genes. Within this framework, we reconstructed an extensive, high-quality functional gene network for *Saccharomyces cerevisiae*, consisting of 4681 (~81%) of the known yeast genes linked by ~34,000 probabilistic linkages comparable in accuracy to small-scale interaction assays. The integrated linkages distinguish true from false-positive interactions in earlier data sets; new interactions emerge from genes' network contexts, as shown for genes in chromatin modification and ribosome biogenesis.

Knowledge of the correct overall structures of gene networks will be invaluable for characterizing the complex roles of individual genes and the interplay between the many systems in a cell. Deriving gene networks from heterogeneous functional genomics data, however, is often difficult, because experiments such as microarray analyses of gene expression (*1*) or systematic protein interaction mapping measure different aspects of gene or protein associations. Affinity purification of proteins analyzed by mass spectrometry (*2*, *3*), for instance, measures the tendency for proteins to be components of the same physical complex, although not necessarily to contact each other directly. By contrast, yeast two-hybrid assays may often indicate direct physical interactions (stable or transient) between proteins (*4*–*6*), whereas synthetic lethal screens (*7*) measure the tendency for genes to compensate for the loss of other genes. Further, these analyses range considerably in accuracy (*8*), and it is not clear a priori which measurements are correct. In spite of these differences, these data sets can, in principle, be computationally integrated, primarily by the reconstruction of network models of the relations between genes (*9*–*12*). Such network reconstructions have largely focused on physical protein interactions and so represent only a subset of biologically important relations.

We sought to construct a more accurate and extensive gene network by considering functional, rather than physical, associations, realizing that each experiment, whether genetic, biochemical, or computational, adds evidence linking pairs of genes, with associated error rates and degree of coverage. In this framework, gene-gene linkages are probabilistic summaries representing functional coupling between genes. Only some of the links represent direct protein-protein interactions; the rest are associations not mediated by physical contact, such as regulatory, genetic, or metabolic coupling, that, nonetheless, represent functional constraints satisfied by the cell during the course of the experiments. Working with probabilistic functional linkages allows many diverse classes of experiments to be integrated into a single, coherent network (Fig. 1), which enables the linkages themselves to be more reliably established.

We first developed a unified scoring scheme for linkages, based on a Bayesian statistics approach. Each experiment is evaluated for its ability to reconstruct known gene pathways and systems by measuring the likelihood that pairs of genes are functionally linked conditioned on the evidence, calculated as a log likelihood score:

$$\mathrm{LLS} = \ln\left(\frac{P(L|E)/{\sim}P(L|E)}{P(L)/{\sim}P(L)}\right)$$

where $P(L|E)$ and ${\sim}P(L|E)$ are the frequencies of linkages (L) observed in the given experiment (E) between annotated genes operating in the same pathway and in different pathways, respectively, whereas $P(L)$ and ${\sim}P(L)$ represent the prior expectations (i.e., the total frequency of linkages between all annotated yeast genes operating in the same pathway and operating in different pathways, respectively). Scores greater than zero indicate that the experiment tends to link genes in the same pathway, with higher scores indicating more confident linkages.

The log likelihood score can be interpreted as being proportional to the accuracy of the experiments and their ability to inform us about cellular pathways. Because each experiment is measured on a common benchmark, different experiments' scores are directly comparable, even when the natures of experiments are distinct (e.g., comparing genetic relations to physical interactions),

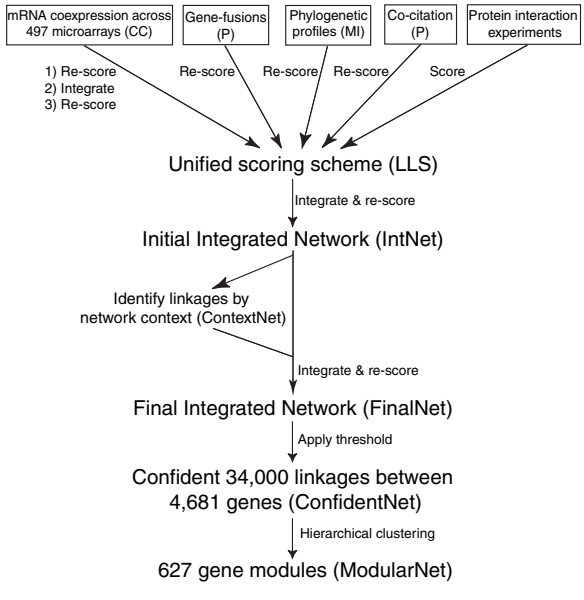


Fig. 1. The method for integrating functional genomics data. Functional genomics data sets are first benchmarked for their relative accuracies; these are used as weights in a probabilistic integration of the data. Several raw data sets already have intrinsic scoring schemes, indicated in parentheses (e.g., CC, correlation coefficients; P, probabilities, and MI, mutual information scores). These data are rescored with LLS, then integrated into an initial network (IntNet). Additional linkages from the genes' network contexts (ContextNet) are then integrated to create the final network (FinalNet), with ~34,000 linkages between 4681 genes (ConfidentNet) scoring higher than the gold standard (small-scale assays of protein interactions). Hierarchical clustering of ConfidentNet defined 627 modules of functionally linked genes spanning 3285 genes ("ModularNet"), approximating the set of cellular systems in yeast.

[1]Center for Systems and Synthetic Biology, and [2]Department of Chemistry and Biochemistry, Institute for Molecular Biology, University of Texas at Austin, Austin, TX 78712–1064, USA.

*Present address: Center for Bioinformatics, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA 19104, USA.
†Present address: Mission Bay Genentech Hall, 600 16th Street, Suite N472D, University of California at San Francisco, San Francisco, CA 94143–2240, USA.
‡To whom correspondence should be addressed. E-mail: marcotte@icmb.utexas.edu

and can be added to indicate confidence of combined evidence.

As scoring "benchmarks," we tested the method against two primary annotation references: the Kyoto-based KEGG pathway database (*13*) and the experimentally observed yeast protein subcellular locations determined by genomewide green fluorescent protein (GFP)–tagging and microscopy (*14*). KEGG scores were used for integrating linkages, with the other benchmark withheld as an independent test of linkage accuracy. Cross-validated benchmarks and benchmarks based on the Gene Ontology (GO) (*15*) and KOG gene annotations (*16*) provided comparable results (*17*).

Seven large-scale yeast protein interaction experiments, including small-scale protein interaction assays collected from the Database of Interacting Proteins (DIP) (*18*), high-throughput mass spectrometry (*2*, *3*), yeast two-hybrid (*4*–*6*), and synthetic lethal assays (*7*), showed similar rankings of accuracy across the four benchmark tests (Fig. 2; fig. S8, A and B). These tests indicate that small-scale experiments (our "gold standard" for high accuracy linkages) have been the most accurate of all, whereas the large-scale experiments vary considerably in quality. Even the least accurate experiments score better than random linkages (for which LLS = 0), highlighting the merit of this method: weak evidence from multiple sources can be combined to provide strong overall evidence for a linkage.

Functional linkages were first inferred on the basis of genes' mRNA coexpression across each of 12 sets of DNA microarray experiments (497 microarray experiments in total), then integrated via a rank-weighted sum of log likelihood scores (*17*) to create the combined set of coexpression-derived linkages. To construct the initial integrated network ("IntNet," Fig. 1), we combined eight categories of data, including the physical and genetic interaction data sets, mRNA coexpression linkages, functional linkages from literature mining (*17*), and computational linkages from two comparative genomics methods, Rosetta stone (gene-fusion) linkages (*19*, *20*) and phylogenetic profiles (*21*). Integrating functional genomics data also allowed discovery of additional relations between genes linked, in turn, to a common set of genes ["ContextNet" (*17*, *22*–*25*)]; these linkages were scored and integrated as above to construct the final gene network ("FinalNet," Fig. 3A). The final network has ~34,000 linkages at an accuracy comparable to the gold standard small-scale interaction assays (Fig. 2), which provides linkages ("ConfidentNet") for more than 4681 yeast genes (~81% of the yeast proteome). The network is reasonably distinct from networks of physical interacting proteins [e.g., sharing only ~16% of linkages with (*11*); see (*17*)].

Adding context-inferred linkages increased clustering of genes (fig. S7, C and D), which produced a highly modular gene network with well-defined subnetworks. We expected these gene clusters to reflect gene systems and modules (*26*–*30*). We could therefore generate a simplified view of the major trends in the network (Fig. 3B) by clustering genes of ConfidentNet according to their connectivities (*17*). Of the 4681 genes, 3285 (~70.2%) were grouped into 627 clusters, reflecting the high degree of modularity. Genes' functions within each cluster are highly coherent (fig. S12), and with 2 to 154 genes per cluster (~5 genes per cluster on average), the clusters effectively capture typical gene pathways and/or systems. A region of the modular network centered on the DNA damage response and repair systems is shown in Fig. 3C. The network is clearly hierarchical: Individual clusters represent distinct systems related to DNA damage response and/or repair; these clusters are in turn connected to modules of cell cycle regulatory genes and chromatin silencing (fig. S13), functionally linked to the DNA damage response and/or repair system. [For cluster descriptions and interactive three-dimensional visualizations, see (*17*).]

One can infer individual genes' functions on the basis of linked neighbors. For example, seven uncharacterized genes are implicated in chromatin remodeling (Fig. 3D). All but 1 of the 18 linkages made by these genes arise from the comparative genomics analysis or from the network context methods, which represent examples of the insights that arise only after data integration. Three of the uncharacterized proteins are predicted by sequence homology to have helicase activity, which is reasonable for a relation to chromatin remodeling; four of these proteins localize to the nucleus, further supporting their association. After this network's construction, one gene, VID21, was implicated in chromatin modification as a component of the NuA4 histone acetyl transferase.

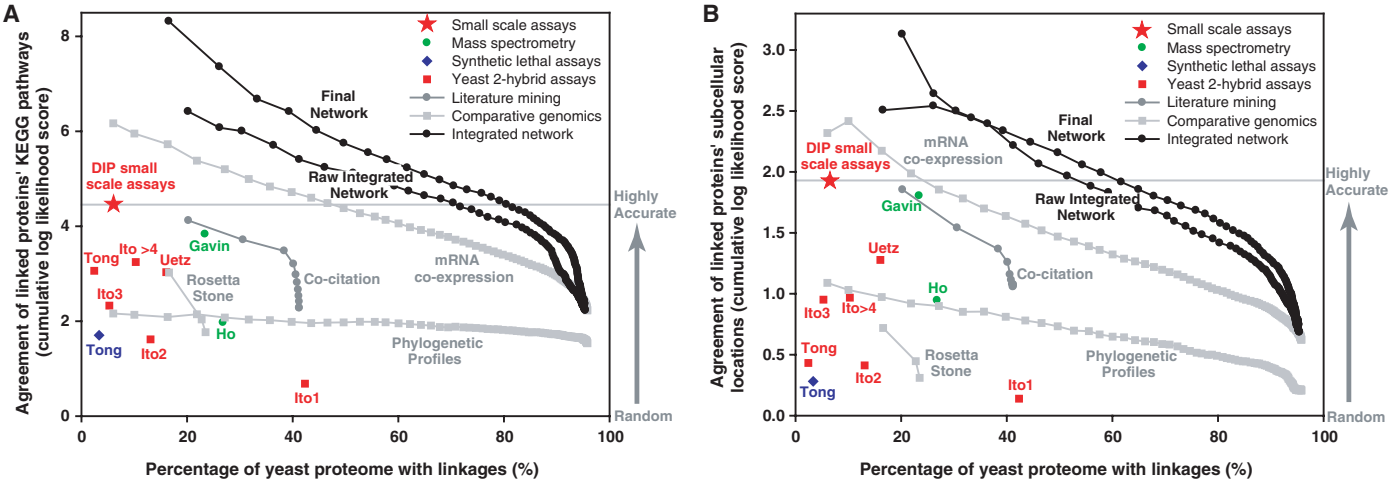
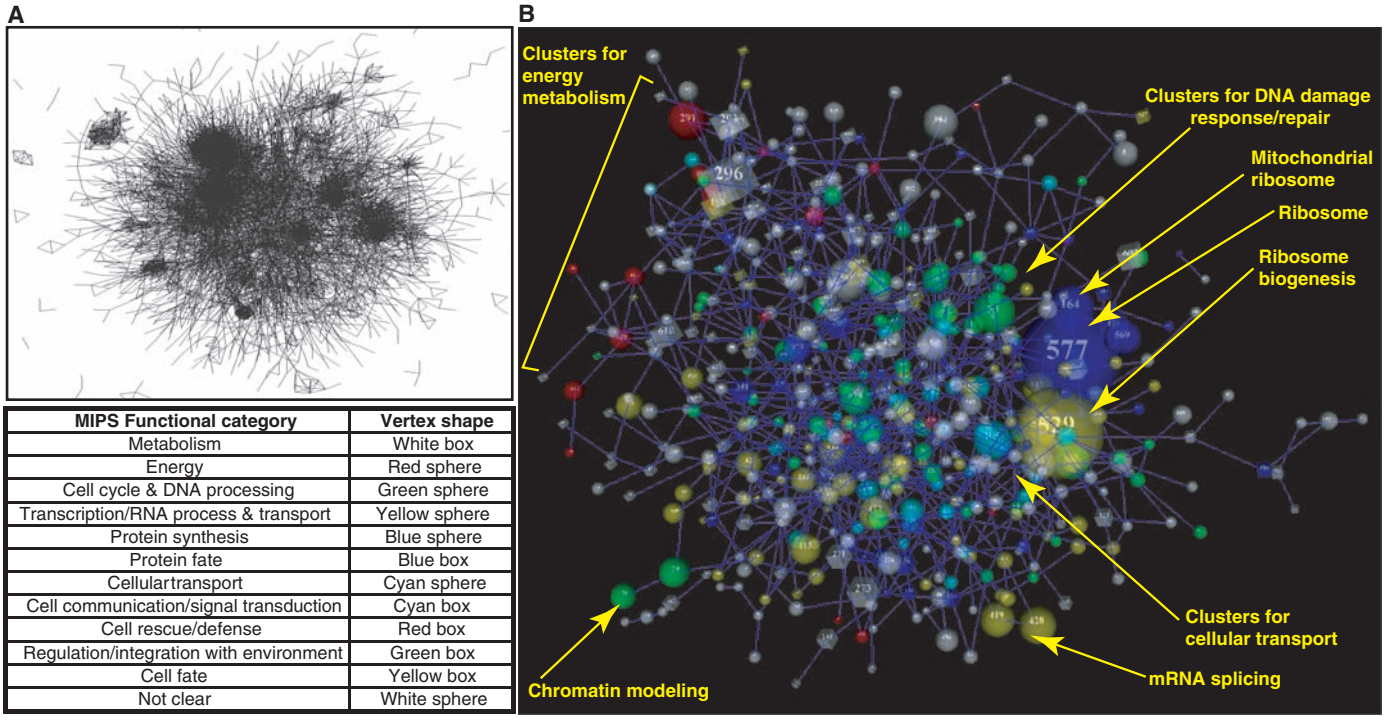


Fig. 2. Benchmarked accuracy and extent of functional genomics data sets and the integrated networks. A critical point is the comparable performance of the networks on distinct benchmarks, which assess the tendencies for linked genes to share (**A**) KEGG pathway annotations (*13*) or (**B**) protein subcellular locations (*14*). Each *x* axis indicates the percentage of protein-encoding yeast genes provided with linkages by the plotted data; each *y* axis indicates relative accuracy, measured as the agreement of the linked genes' annotations on that benchmark. The gold standards of accuracy (red star) for calibrating the benchmarks are small-scale protein-protein interaction data from DIP (*18*). Colored markers indicate experimental linkages; gray markers, computational. The initial integrated network (lower black line), trained using only the KEGG benchmark, has measurably higher accuracy than any individual data set on the subcellular localization benchmark; adding context-inferred linkages in the final network (upper black line) further improves the size and accuracy of the network [see (*17*) for additional benchmarks].

The function of the RNA helicase PRP43, previously thought to be involved only in pre-mRNA splicing and implicated in lariat-intron release from the spliceosome (*31*), is also clarified in the network. PRP43 is linked most strongly to genes of ribosome biogenesis and rRNA processing. The tightest links are to ERB1, RRB1, NUG1, LHP1, and PWP1, the first three of which are confirmed ribosome biogenesis factors. These links derive only from the coexpression and context methods [with a single exception from (*3*)]; data integration is therefore critical. The association of PRP43 with ribosome biogenesis has now been experimentally validated (*32*): the growth defect conferred by a PRP43 conditional lethal mutation corresponds to a rapid and major defect in rRNA processing. These data indicate that rRNA processing is the essential function of PRP43, and it joins a growing group of RNA helicases with two or more distinct functions.

The probabilistic gene network we describe integrates evidence from diverse sources to reconstruct an accurate network, by estimating the functional coupling among yeast genes, and provides a view of the relations between



MIPS Functional category	Vertex shape
Metabolism	White box
Energy	Red sphere
Cell cycle & DNA processing	Green sphere
Transcription/RNA process & transport	Yellow sphere
Protein synthesis	Blue sphere
Protein fate	Blue box
Cellular transport	Cyan sphere
Cell communication/signal transduction	Cyan box
Cell rescue/defense	Red box
Regulation/integration with environment	Green box
Cell fate	Yellow box
Not clear	White sphere

Fig. 3. Features of integrated networks. The final network shows extensive clustering of genes into modules, evident in the "clumping" (**A**). At an intermediate degree of clustering that maximizes cluster size and functional coherence (**B**), 564 (of 627) modules are shown connected by the 950 strongest intermodule linkages. Module colors and shapes indicate associated functions, as defined by Munich Information Center for Protein Sequencing (MIPS) (*34*), with sizes proportional to the number of genes, and connections inversely proportional to the fraction of genes linking the clusters. Portions of the final, confident gene network are shown for (**C**) DNA damage response and/or repair, where modularity gives rise to gene clusters, indicated by similar colors (see also fig. S13), and (**D**) chromatin remodeling, with several uncharacterized genes (red labels). Networks are visualized with Large Graph Layout (LGL) (*35*).

yeast proteins distinct from their physical interactions. The application of this strategy to other organisms, such as to the human genome, is conceptually straightforward: (i) assemble benchmarks for measuring the accuracy of linkages between human genes based on properties shared among genes in the same systems, (ii) assemble gold standard sets of highly accurate interactions for calibrating the benchmarks, and (iii) benchmark functional genomics data for their ability to correctly link human genes, then integrate the data as described. New data can be incorporated in a simple manner [e.g., see (*33*)], serving to reinforce the correct linkages. Thus, the gene network will ultimately converge by successive approximation to the correct structure simply by continued addition of functional genomics data in this framework.

References and Notes

1. J. Gollub *et al.*, *Nucleic Acids Res.* **31**, 94 (2003).
2. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
3. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
4. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
5. T. Ito *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569 (2001).
6. A. H. Tong *et al.*, *Science* **295**, 321 (2002).
7. A. H. Tong *et al.*, *Science* **294**, 2364 (2001).
8. C. von Mering *et al.*, *Nature* **417**, 399 (2002).
9. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, D. Eisenberg, *Nature* **402**, 83 (1999).
10. T. Ideker *et al.*, *Science* **292**, 929 (2001).
11. R. Jansen *et al.*, *Science* **302**, 449 (2003).
12. C. von Mering *et al.*, *Nucleic Acids Res.* **31**, 258 (2003).
13. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, *Nucleic Acids Res.* **30**, 42 (2002).
14. W. K. Huh *et al.*, *Nature* **425**, 686 (2003).
15. S. S. Dwight *et al.*, *Nucleic Acids Res.* **30**, 69 (2002).
16. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
17. Materials and methods, supporting text, figures, tables, and data are available on *Science* Online.
18. I. Xenarios *et al.*, *Nucleic Acids Res.* **30**, 303 (2002).
19. E. M. Marcotte *et al.*, *Science* **285**, 751 (1999).
20. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, C. A. Ouzounis, *Nature* **402**, 86 (1999).
21. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285 (1999).
22. M. Thompson, E. Marcotte, M. Pellegrini, T. Yeates, D. Eisenberg, in *Currents in Computational Molecular Biology*, S. Miyano, R. Shamir, T. Takagi, Eds. (Universal Academy Press, Tokyo, 2000).
23. D. S. Goldberg, F. P. Roth, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4372 (2003).
24. M. P. Samanta, S. Liang, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12579 (2003).
25. T. Schlitt *et al.*, *Genome Res.* **13**, 2568 (2003).
26. L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray, *Nature* **402**, C47 (1999).
27. J. B. Pereira-Leal, A. J. Enright, C. A. Ouzounis, *Proteins* **54**, 49 (2004).
28. V. Spirin, L. A. Mirny, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12123 (2003).
29. A. W. Rives, T. Galitski, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1128 (2003).
30. C. von Mering *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15428 (2003).
31. A. Martin, S. Schneider, B. Schwer, *J. Biol. Chem.* **277**, 17743 (2002).
32. Scott Stevens, personal communication.
33. A. G. Fraser, E. M. Marcotte, *Nat. Genet.* **36**, 559 (2004).
34. H. W. Mewes *et al.*, *Nucleic Acids Res.* **30**, 31 (2002).
35. A. T. Adai, S. V. Date, S. Wieland, E. M. Marcotte, *J. Mol. Biol.* **340**, 179 (2004).

36. We thank S. Stevens for sharing prepublication *PRP43* results and A. Fraser, A. Ramani, Z. Simpson, and A. Johnson for critical comments and discussion. This work is supported by the Welch (F-1515) and Dreyfus Foundations, NSF, a Packard Fellowship (E.M.M.), and NIH (GM067779-01).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5701/1555/DC1
Materials and Methods
SOM Text
Figs. S1 to 14
Tables S1 to S4
Supporting Data S1 to S5

22 April 2004; accepted 8 October 2004

Requirement of JNK2 for Scavenger Receptor A–Mediated Foam Cell Formation in Atherogenesis

Romeo Ricci,[1,2]*† Grzegorz Sumara,[1,2]† Izabela Sumara,[3] Izabela Rozenberg,[1] Michael Kurrer,[4] Alexander Akhmedov,[1] Martin Hersberger,[5] Urs Eriksson,[7] Franz R. Eberli,[1] Burkhard Becher,[6] Jan Borén,[8] Mian Chen,[9] Myron I. Cybulsky,[9] Kathryn J. Moore,[10] Mason W. Freeman,[10] Erwin F. Wagner,[11] Christian M. Matter,[1]‡ Thomas F. Lüscher[1]‡

In vitro studies suggest a role for c-Jun N-terminal kinases (JNKs) in proatherogenic cellular processes. We show that atherosclerosis-prone *ApoE*$^{-/-}$ mice simultaneously lacking JNK2 (*ApoE*$^{-/-}$ *JNK2*$^{-/-}$ mice), but not *ApoE*$^{-/-}$ *JNK1*$^{-/-}$ mice, developed less atherosclerosis than do *ApoE*$^{-/-}$ mice. Pharmacological inhibition of JNK activity efficiently reduced plaque formation. Macrophages lacking JNK2 displayed suppressed foam cell formation caused by defective uptake and degradation of modified lipoproteins and showed increased amounts of the modified lipoprotein-binding and -internalizing scavenger receptor A (SR-A), whose phosphorylation was markedly decreased. Macrophage-restricted deletion of JNK2 was sufficient to decrease atherogenesis. Thus, JNK2-dependent phosphorylation of SR-A promotes uptake of lipids in macrophages, thereby regulating foam cell formation, a critical step in atherogenesis.

Atherosclerosis is the result of complex interactions between modified lipoproteins, monocyte-derived macrophages that become foam cells, T lymphocytes, and cells from the vessel wall (*1*, *2*). The c-Jun N-terminal kinases (JNKs) belong to the mitogen-activated protein kinase (MAPK) family. Ten JNK isoforms have been identified in the human brain, corresponding to alternative spliced isoforms derived from the JNK1, JNK2, and JNK3 genes (*3*). JNK1 and JNK2 are widely expressed. In contrast, JNK3 has a more limited pattern of expression that is largely restricted to brain, heart, and testis. Although mice lacking JNK1 or JNK2 appear morphologically normal, they are immunocompromised because of T-cell defects (*4*). Recent studies in murine disease models defined specific functions for JNK1 and JNK2. JNK1 regulates insulin resistance and obesity (*5*). JNK2 is required for collagen-induced arthritis (*6*). In vitro studies have revealed that JNK proteins act in a variety of pro-atherogenic cellular processes involving endothelial cell activation, T-effector cell differentiation and proliferation, and migration of vascular smooth muscle cells (VSMCs) (*7*).

To investigate the role of JNK in atherosclerotic plaque formation in vivo, we used atherosclerosis-prone apolipoprotein E

[1]Cardiovascular Research, Institute of Physiology, and Division of Cardiology, University Hospital Zurich, CH-8057 Zurich, Switzerland. [2]Institute of Cell Biology, [3]Institute of Biochemistry, Eidgenössische Technische Hochschule, Hönggerberg, CH-8093 Zurich, Switzerland. [4]Department of Pathology, [5]Institute of Clinical Chemistry, [6]Department of Neurology/Neuroimmunology Unit, University Hospital Zurich, CH-8091 Zurich, Switzerland. [7]Experimental Critical Care Medicine, Department of Research and Medicine A, Basel University Hospital, CH-4031 Basel, Switzerland. [8]Wallenberg Laboratory for Cardiovascular Research, Goteborg University, Goteborg S-4345, Sweden. [9]Toronto General Research Institute and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada. [10]Lipid Metabolism Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. [11]Institute of Molecular Pathology, A-1030 Vienna, Austria.

*To whom correspondence should be addressed. E-mail: romeo.ricci@cell.biol.ethz.ca
†These authors contributed equally to this work.
‡These authors contributed equally to this work.