Positive Selection of Tyrosine Loss in Metazoan Evolution

Chris Soon Heng Tan, *et al.*

Science **325**, 1686 (2009);
DOI: 10.1126/science.1174301

tion by simply disrupting or generating protein-protein interactions (fig. S2B).

An important implication of flexibility in phosphorylation site positioning is that combinatorial control by multiple kinases is readily evolved. Indeed, the protein kinase Ime2, a distant relative of Cdk1 that is expressed solely in meiotic cells, phosphorylates a large number of Cdk1 substrates at distinct sites but can still have the same effect as Cdk1 on substrate function (*17*).

The evolution of Cdk1 signaling appears to share features with the evolution of transcriptional regulation (fig. S7). Transcriptional regulators and Cdks both maintain their biochemical specificities (the DNA consensus motif and peptide consensus motif, respectively) over long evolutionary time scales. However, in both cases there is rapid evolution of the intergenic and disordered regions, respectively, that contain these motifs. In transcriptional regulation, DNA sequence motifs can function from many positions relative to the gene being controlled and, because of their short length and sequence degeneracy, can evolve rapidly (*18*–*20*). Similarly, many Cdk1 phosphorylation sites are not tightly constrained within the protein target sequence, and the signals for phosphorylation are short and easily evolved. These features allow cell-cycle control mechanisms to adapt rapidly to developmental challenges and opportunities that arise over time.

References and Notes

1. D. O. Morgan, *The Cell Cycle: Principles of Control* (New Science Press, London, 2007).
2. Materials and methods are available as supporting material on *Science* online.
3. A. C. Bishop *et al.*, *Nature* **407**, 395 (2000).
4. J. A. Ubersax *et al.*, *Nature* **425**, 859 (2003).
5. S. C. Strickfaden *et al.*, *Cell* **128**, 519 (2007).
6. Z. Serber, J. E. Ferrell Jr., *Cell* **128**, 441 (2007).
7. M. B. Yaffe, A. E. Elia, *Curr. Opin. Cell Biol.* **13**, 131 (2001).
8. T. Pawson, G. D. Gish, P. Nash, *Trends Cell Biol.* **11**, 504 (2001).
9. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037 (2004).
10. B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, A. D. Johnson, *PLoS Biol.* **6**, e38 (2008).
11. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).
12. K. Kakiuchi *et al.*, *Biochemistry* **46**, 7781 (2007).
13. D. Bridges, G. B. Moorhead, *Sci. STKE* **2005**, re10 (2005).
14. A. M. Moses, J. K. Heriche, R. Durbin, *Genome Biol.* **8**, R23 (2007).
15. A. M. Moses, M. E. Liku, J. J. Li, R. Durbin, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17713 (2007).
16. L. S. Drury, J. F. Diffley, *Curr. Biol.* **19**, 530 (2009).
17. L. J. Holt, J. E. Hutti, L. C. Cantley, D. O. Morgan, *Mol. Cell* **25**, 689 (2007).
18. G. A. Wray *et al.*, *Mol. Biol. Evol.* **20**, 1377 (2003).
19. S. B. Carroll, J. K. Grenier, S. D. Weatherbee, *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell, Malden, MA, 2005).
20. B. B. Tuch, H. Li, A. D. Johnson, *Science* **319**, 1797 (2008).
21. S. Ghaemmaghami *et al.*, *Nature* **425**, 737 (2003).

22. We thank J. Feldman, R. Fletterick, M. Jacobson, H. Li, M. Matyskiela, P. O'Farrell, M. Sullivan, and S. Naylor for helpful comments; A. K. Dunker, E. Garner, C. Oldfield, K. Shimizu, and T. Ishida for disorder prediction algorithms; the Broad Institute, Sanger Center, Génolevures, and the Joint Genome Institute for genome sequence data; and O. Jensen, C. Zhang, and K. Shokat for reagents. This work was supported by grants from NIH (GM50684 to D.O.M., HG3456 to S.P.G., and GM037049 to A.D.J.) and fellowships from NSF (to L.J.H. and B.B.T.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/325/5948/1682/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S6
References
Databases S1 and S2

27 February 2009; accepted 15 July 2009
10.1126/science.1172867

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]* Rune Linding[5]*

John Nash showed that within a complex system, individuals are best off if they make the best decision that they can, taking into account the decisions of the other individuals. Here, we investigate whether similar principles influence the evolution of signaling networks in multicellular animals. Specifically, by analyzing a set of metazoan species we observed a striking negative correlation of genomically encoded tyrosine content with biological complexity (as measured by the number of cell types in each organism). We discuss how this observed tyrosine loss correlates with the expansion of tyrosine kinases in the evolution of the metazoan lineage and how it may relate to the optimization of signaling systems in multicellular animals. We propose that this phenomenon illustrates genome-wide adaptive evolution to accommodate beneficial genetic perturbation.

It is a biological paradox that organism complexity shows limited correlation with gene repertoire size (*1*). However, some protein families (*2*) have expanded with organism complexity as measured by number of cell types (*3*), especially those involved in regulation, such as tyrosine kinases in signaling, cell-cell communication, and tissue boundary formation (*4*, *5*). We observed a striking negative correlation of genomically encoded tyrosine content with the number of distinct cell types in metazoan species (Spearman's $\rho = -0.89$, approximate $P = 3.0 \times 10^{-6}$; Pearson's $\rho = -0.89$, approximate $P = 4.0 \times 10^{-6}$) (Fig. 1A). Thus, metazoans with more cell types have proportionally less potential tyrosine phosphosites. Similarly, we observed that the number of tyrosine kinase domains correlates negatively with genomic tyrosine content (Spearman's $\rho = -0.68$, approximate $P = 3.7 \times 10^{-3}$; Pearson's $\rho = -0.81$, approximate $P = 1.3 \times 10^{-4}$) (Fig. 1B). Including dual-specificity mixed-lineage kinases (MLKs) and mitogen-activated protein kinase kinases (MEKs) revealed a similar pattern (fig. S1A).

These observations suggest an evolutionary model in which the acquisition of a tyrosine kinase results in systems-level adaptation to remove deleterious phosphorylation events that cause aberrant cellular behavior and diseases (*4*). Assuming that a cell begins with a single tyrosine kinase, which is subsequently duplicated, it follows that the kinases may functionally diverge, as a result of relaxation in evolutionary constraints, to phosphorylate new substrates. Emerging kinase specificities could be retained if new substrates confer selection advantage. However, it is unlikely that every new phosphorylation event is beneficial. We hypothesize that optimization of newly emerged signaling networks would follow (*6*) through the elimination of detrimental phosphorylation events by tyrosine-removing mutations. Even if many new phosphorylation sites are not deleterious, an organism with minimized noisy signaling systems is likely to have a fitness advantage. This scenario is repeated with the subsequent duplication of tyrosine kinases leading to more tyrosine residues lost (*7*).

Despite several recent systematic phosphoproteomic studies (*8*), many human proteins have no observed phosphotyrosines. Our model suggests that tyrosine loss had occurred predominantly in these proteins in order to minimize tyrosine phosphorylation. To test this hypothesis, we investigated differences in tyrosine loss between these proteins (Non-pTyr) and those that are tyrosine phosphorylated (pTyr). Comparing members of these two groups to their orthologous proteins in *S. cerevisiae* (*7*), which lack conventional tyrosine kinases, enabled us to assess the degree of tyrosine loss that may be triggered by the onset of phosphotyrosine signaling in metazoans.

[1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto M5G 1X5, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto M5S 1A8, Canada. [3]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto M5S 3E1, Canada. [4]Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA. [5]Cellular and Molecular Logic Team, Section of Cell and Molecular Biology, The Institute of Cancer Research (ICR), London, SW3 6JB, UK.

*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.)

A significantly smaller fraction of amino acids are tyrosines in human proteins than in their yeast orthologs (approximate $P = 3.5 \times 10^{-4}$, paired Wilcoxon signed rank test) (Fig. 1C). However, this phenomenon was statistically more pronounced in non-pTyr proteins than in pTyr proteins (approximate $P = 5.1 \times 10^{-9}$, Mann-Whitney test) (Fig. 1C). A similar trend was observed on the basis of absolute tyrosine residue counts (approximate $P = 2.0 \times 10^{-7}$, Mann-Whitney test) (fig. S1B) and on a higher confidence subset of pTyr proteins that either have multiple phosphotyrosines or have sites observed in multiple studies (approximate $P = 1.3 \times 10^{-7}$, Mann-Whitney test).

Thus, tyrosine loss was strongly favored in human protein evolution, most notably in protein subsets that are not known to be tyrosine-phosphorylated. Genetic drift (*9*) is unlikely to account for these differences observed in a large number of evolutionarily distant human-yeast protein orthologs. Because tyrosine is an essential and the most expensive amino acid to biosynthesize (*10*) after tryptophan and phenylalanine, essentiality and biosynthetic cost could be major factors in the observed loss. However, this is unlikely because we observed a strong positive correlation of number of cell types with tryptophan and a weaker negative correlation for phenylalanine (table S1). Instead, we propose that positive selection of tyrosine-removing mutations occurred in the metazoan lineage to reduce adventitious tyrosine phosphorylation, at least in part. This optimization process probably shaped signaling networks crucial for the development of multicellular animals. Additionally, this could provide a mechanism to prevent unspecific phosphorylation events that operates with the evolution of domains, consensus motifs (*11*), and contextual factors to colocalize kinases with their substrates (*11*–*13*). We observed a slightly stronger negative correlation of genomically encoded tyrosine content with the number of inferred phosphotyrosine-binding domains than tyrosine kinase domain count (Spearman's $\rho = -0.81$, Pearson's $\rho = -0.88$) (fig. S1A), which is in agreement with the notion that tyrosine phosphorylation exerts parts of its functional effects through creating binding sites for phosphobinding domains like Src homology 2 (SH2) and phosphotyrosine binding domain (PTB) (*14*).

The choanoflagellate *Monosiga brevicollis*, which is a member of the only known unicellular lineage with canonical tyrosine kinases (*15*), is an outlier in the cell-type correlation studied above. This observation is consistent with the emerging picture that choanoflagellates represent a distinct evolutionary branch from metazoans in which phosphotyrosine-signaling systems have been used for divergent functions (*16*, *17*). Nevertheless, the *Monosiga* analysis is still consistent with optimization of phosphotyrosine signaling in this lineage; compared with the metazoans analyzed here, *Monosiga* has higher numbers of tyrosine kinases

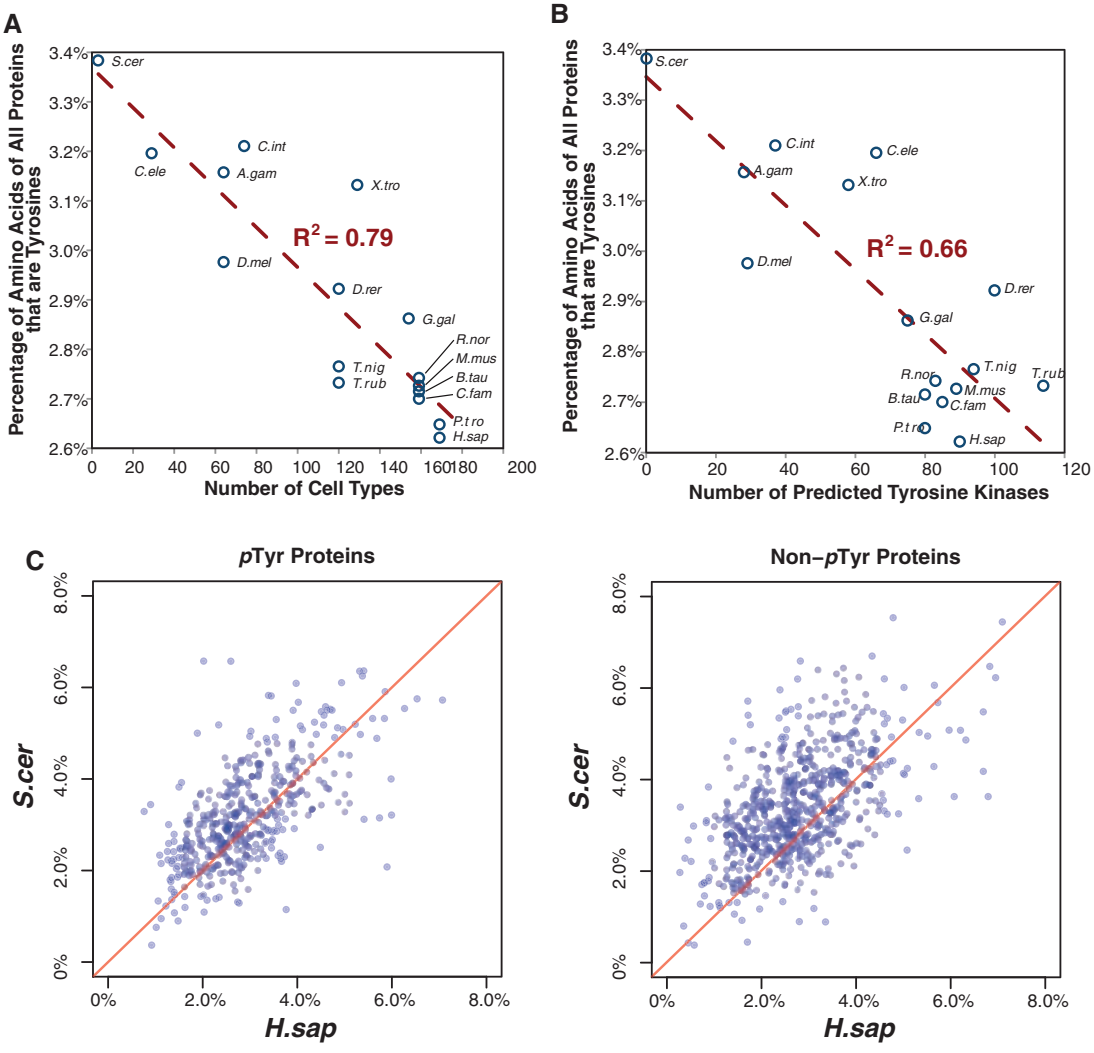

Fig. 1. Correlation of expansion of phosphotyrosine-signaling systems with loss of genome-encoded tyrosine residues. (**A**) The genomically encoded tyrosine content in metazoan organisms and yeast correlate negatively and significantly with organism complexity as measured by distinct cell types (*2*). Bakers' yeast (*S. cerevisiae*) is included as a unicellular eukaryote for comparison. The species analyzed are yeast (*S. cerevisiae*), worm, (*C. elegans*), sea squirt (*C. intestinalis*), fly (*D. melanogaster*), mosquito (*A. gambiae*), zebrafish (*D. rerio*), tetraodon pufferfish (*T. nigroviridis*), Japanese pufferfish (*T. rubripes*), frog (*X. tropicalis*), chicken (*G. gallus*), dog (*C. familiaris*), cow (*B. taurus*), mouse (*M. musculus*), rat (*R. norvegicus*), chimpanzee (*P. troglodytes*), and human (*H. sapiens*). (**B**) The number of tyrosine kinase domains in metazoans and yeast correlates negatively and significantly with the number of distinct cell types. (**C**) The fraction of tyrosines in human-yeast ortholog protein pairs. Every point in the scatter plot represents a human-yeast ortholog protein pair where the (x, y) values denote the tyrosine content in human and yeast proteins, respectively. For simplicity, only proteins with an inferred one-to-one orthologous relationship between human and yeast are analyzed (for example, to avoid accelerated sequence divergence due to functional redundancy of paralogs). Orthologous protein pairs lying above the red diagonal ($x = y$) lines have higher tyrosine composition in yeast than in 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 human proteins conserved in yeast not known to be tyrosine phosphorylated.

[127 (7)] and lower genomically encoded tyrosine content (2.3%).

Other factors, such as tyrosine sulfation, could have contributed to the observed tyrosine loss, which raises the question of whether other post-translational modifications and regulatory mechanisms are under similar evolutionary selection that could also explain genomic GC conversion. We observed a strong negative correlation of number of cell types with amino acids that can be methylated or glycosylated (table S1). The numbers of genomically encoded threonine showed strong negative correlations with serine/threonine kinase and cell-type numbers, although these trends were not observed with serine (fig. S2), which suggests possible coarse-grained functional differences between serine- and threonine phosphorylation in metazoans.

Our findings suggest that the implementation of tyrosine kinase signaling, as a biological innovation that probably assisted in the development of multicellular organisms, required system-level adaptive mutations. Analogous to the arguments of John Nash in his dissertation (*18*), this phenomenon highlights a general principle of adaptive evolution pertaining to the introduction of new components into a complex system and parallels the evolution of some human societies in which the local populations have to adjust and adapt to the influx of immigrants contributing to the societies' economic development. This principle may serve as an important framework when considering the evolution and fidelity of complex biological systems. Finally, this work raises the possibility that complex regulatory diseases, such as cancer, might result from systems-wide adaptive changes in human genomes and signaling systems.

References and Notes

1. E. Szathmáry, F. Jordán, C. Pál, *Science* **292**, 1315 (2001).
2. C. Vogel, C. Chothia, *PLOS Comput. Biol.* **2**, e48 (2006).
3. J. T. Bonner, *Integ. Biol* **1**, 28 (1998).
4. T. Hunter, *Curr. Opin. Cell Biol.* http://dx.doi.org/10.1016/j.ceb.2009.01.028 (2009).
5. Z. Songyang, L. C. Cantley, *Trends Biochem. Sci.* **20**, 470 (1995).
6. A. Zarrinpar, S.-H. Park, W. A. Lim, *Nature* **426**, 676 (2003).
7. Materials and methods are available as supporting material on *Science* Online.
8. C. Jørgensen, R. Linding, *Brief. Funct. Genomics Proteomics* **7**, 17 (2008).
9. S. Wright, *Genetics* **16**, 97 (1931).
10. D. W. Raiford *et al.*, *J. Mol. Evol.* **67**, 621 (2008).
11. M. B. Yaffe *et al.*, *Nat. Biotechnol.* **19**, 348 (2001).
12. R. Linding *et al.*, *Cell* **129**, 1415 (2007).
13. M. L. Miller *et al.*, *Sci. Signal.* **1**, ra2 (2008).
14. B. T. Seet, I. Dikic, M.-M. Zhou, T. Pawson, *Nat. Rev. Mol. Cell Biol.* **7**, 473 (2006).
15. N. King *et al.*, *Nature* **451**, 783 (2008).
16. D. Pincus, I. Letunic, P. Bork, W. A. Lim, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9680 (2008).
17. G. Manning, S. L. Young, W. T. Miller, Y. Zhai, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9674 (2008).
18. J. Nash, thesis, Princeton University, Princeton, NJ (1950).

19. We thank C. Jørgensen, J. Zhang, K. Colwill, T. J. Gibson, J. Jin, and K. 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.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1174301/DC1
Materials and Methods
Figs. S1 and S2
Table S1
References and Notes

31 March 2009; accepted 25 June 2009
Published online 9 July 2009;
10.1126/science.1174301
Include this information when citing this paper.

Evolution of a Novel Phenolic Pathway for Pollen Development

Michiyo Matsuno,[1]* Vincent Compagnon,[1]* Guillaume A. Schoch,[2] Martine Schmitt,[3] Delphine Debayle,[1] Jean-Etienne Bassard,[1] Brigitte Pollet,[4] Alain Hehn,[5] Dimitri Heintz,[1] Pascaline Ullmann,[1] Catherine Lapierre,[4] François Bernier,[1] Jürgen Ehlting,[6] Danièle Werck-Reichhart[1]†

Metabolic plasticity, which largely relies on the creation of new genes, is an essential feature of plant adaptation and speciation and has led to the evolution of large gene families. A typical example is provided by the diversification of the cytochrome P450 enzymes in plants. We describe here a retroposition, neofunctionalization, and duplication sequence that, via selective and local amino acid replacement, led to the evolution of a novel phenolic pathway in Brassicaceae. This pathway involves a cascade of six successive hydroxylations by two partially redundant cytochromes P450, leading to the formation of N^1,N^5-di(hydroxyferuloyl)-N^{10}-sinapoylspermidine, a major pollen constituent and so-far-overlooked player in phenylpropanoid metabolism. This example shows how positive Darwinian selection can favor structured clusters of nonsynonymous substitutions that are needed for the transition of enzymes to new functions.

Plant adaptation relies on a remarkable metabolic plasticity that is reflected by the evolution of large gene families. New family members may emerge by segment duplication (sometimes followed by exon shuffling), gene fusion, or retroposition; that is, reverse transcription of mRNAs followed by insertion of the intronless cDNA into the parent genome (*1*). A striking example of duplication for gene cluster evolution in the triterpenoid metabolism was recently described (*2*). A survey of the *Arabidopsis* and rice genomes identified several retroposons (*3*, *4*), but retrogene evolution for the acquisition of novel functions was well described only in mammals or insects (*1*).

We recently performed a global coexpression analysis to predict the function of orphan cytochrome P450 genes in *Arabidopsis thaliana* (*5*). A candidate emerged with a predicted function in a new branch of phenolic metabolism. *CYP98A8* (AT1G74540) and its paralog *CYP98A9* (AT1G74550) are two chromosome 1–clustered duplications of an ancestor of *CYP98A3*, the latter previously shown to *meta*-hydroxylate the *p*-coumaric esters of shikimic/quinic acids to form lignin monomers (*6*, *7*). CYP98A8 and CYP98A9 share only about 50% protein identity with CYP98A3, but they do belong to the monophyletic CYP98A clade that includes all confirmed 4-coumaroyl shikimate/quinate meta-hydroxylases; but in protein phylogenies, CYP98A8 and CYP98A9 appear separated from vascular plant sequences by CYP98s from conifers and moss (fig. S1B). This means either that *CYP98A8* and *CYP98A9* diverged before angiosperm evolution or that they appeared recently and evolved fast, and were thus placed at the base of the clade because of long-branch attraction. In favor of the latter hypothesis, cDNA phylogenies show that *CYP98A8* and *CYP98A9* form a sister group with *CYP98A3* (Fig. 1A and fig. S1D), and no potential orthologs of *CYP98A8* exist in the TIGR Plant Transcript Assemblies and PlantGDB databases, except for a single transcript assembly (*CYP98A53*) from *Brassica napus*. *CYP98A8* and *CYP98A9* are intronless, whereas all other known higher plant *CYP98A*s contain two introns at conserved positions, including one that is considered a signature of the expanded *CYP71* clan to which the *CYP98* family belongs (*8*). This indicates a gene birth event via mRNA-mediated transposition.

Ratios of nonsynonymous and synonymous substitution rates (*dN*/*dS* or ω) were first es-

[1]Institut de Biologie Moléculaire des Plantes, CNRS-UPR2357, Université de Strasbourg, 28 Rue Goethe, 67083 Strasbourg Cedex, France. [2]Pharma Research, Discovery Technologies, F. Hoffmann-La Roche, CH-4070 Basel, Switzerland. [3]Institut Gilbert Laustriat LC1, UMR CNRS/ULP 7175, Faculté de Pharmacie de Strasbourg, 74 Route du Rhin, 67401 Illkirch, France. [4]UMR 206 Chimie Biologique AgroParisTech/INRA, AgroParisTech, 78850 Thiverval-Grignon, France. [5]UMR Nancy-Université, INRA Agronomie et Environnement, 2 Avenue Forêt de Haye, Boite Postale 172, 54505 Vandoeuvre-lès-Nancy, France. [6]Centre for Forest Biology, University of Victoria, Post Office Box 3020 STN CSC, Victoria, British Columbia V8W 3N5, Canada.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: daniele.werck@ibmp-ulp.u-strasbg.fr