$\beta_L$ are assigned values to give a basic reproduction number, $R_0$, for local transmission in a patch of 5, $R_0$ for mass-action transmission in a patch of 0.4, and an $R_0$ of 0.02 between any two patches. Seasonal variation in contact rates is characterized by $\varepsilon_p$, and is assumed to be opposite in phase for the Northern and Southern hemispheres (represented by setting $\varepsilon_p = +0.25$ for $1 < p \leq M/2$, and $\varepsilon_p = -0.25$ for $M/2 < p \leq M$).

The probability that a host is infected upon exposure depends on the strain $s$, and the host's immune history. Each strain is characterized by $A$ epitopes, each consisting of $C$ codons (three nucleotide bases). Immunity is assumed to be specific to the set of amino acids to which the host has been exposed at each codon, and for simplicity, no functional constraints are imposed on the amino acid sequences. For the default values of $A = 4$ and $C = 3$ used here, a total of $4 \times 10^{15}$ strains are possible. Antigenic distance, $d(s,H)$ between a strain $s$ and the immune history of a host, $H$, is then simply defined as the number of codons in strain $s$ for which the amino acid has not been previously encountered by the host. The level of cross-protection from infection provided at a certain antigenic distance is given by the function $f(d)$, where we assume $f(d) = \theta_1 + (\theta_0 - \theta_1)(d - n_t)/(AC - n_t)$ for $d \geq n_t$, $f(d) = \theta_1$ for $0 < d < n_t$, and $f(d) = 1$ for $d = 0$. $n_t$ (=2) is the threshold level of change necessary for cross-protection to drop below the maximal level set by $\theta_1$(=0.99), and $\theta_0$ (=0.25) is the minimal level of cross-protection mounted against a strain with no similarity to a previously encountered strain at the codons modelled (=0 for no cross-protection). These assumptions reflect empirical studies suggesting that two or more substitutions at key antigenic sites are required to escape pre-existing immunity[26,27].

The probability that a host will be infected by a strain following exposure is given by

$$\varphi_{p,i}(t,s) = \{1 - \omega \exp[-(t - T_{p,i})/\tau]\}\{1 - f[d(s,H_{p,i})]\}$$

where $\tau$ (=270 days) is the decay timescale (half-life = $\tau_{1/2}$ = 187 days ≈ 6 months) and $\omega$ (=1) is the peak level ($0 \leq \omega \leq 1$) of a short-lived strain-transcending immune response that protects against reinfection in the weeks after an infection (see main text). If a host is infected following exposure, its immune history is updated to include the new strain. If exposure does not result in infection, we assume that no immune response is raised to the new strain but that pre-existing immune responses are boosted (akin to the 'original antigenic sin' response[28]), by resetting $T_{p,i}$ to $t - 6$ immediately after exposure.

There is a probability $\delta$ (=$10^{-5}$) per base per day that a nucleotide substitution will occur in the virus in an infected host and the resulting mutant strain will replace the pre-existing strain in that individual. The individual is then infectious with the new strain. All strains are assumed to have the same intrinsic transmissibility (but see Supplementary Information), and model runs were started from near the single-strain equilibrium.

Received 18 November 2002; accepted 21 February 2003; doi:10.1038/nature01509.


1. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56,** 152–179 (1992).
2. Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16,** 1457–1465 (1999).
3. Cox, N. *et al.* in *Options for the Control of Influenza II* (eds Hannoun, C., Kendal, A. P., Klenk, H. D. & Ruben, F. L.) 223–230 (Elsevier Science, Amsterdam, 1993).
4. Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94,** 7712–7718 (1997).
5. Lindstrom, S. E. *et al.* Comparative analysis of evolutionary mechanisms of the hemagglutinin and three internal protein genes of influenza B virus: multiple cocirculating lineages and frequent reassortment of the NP, M, and NS genes. *J. Virol.* **73,** 4413–4426 (1999).
6. Daly, J., Wood, J. & Robertson, J. in *Textbook of Influenza* (eds Nicholson, K., Webster, R. & Hay, A.) 168–177 (Blackwell Science, Oxford, 1998).
7. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286,** 1921–1925 (1999).
8. Cox, N. & Regnery, H. in *Options for the Control of Influenza III* (eds Brown, L. E., Hampson, Q. W. & Webster, R. G.) 591–598 (Elsevier Science, Amsterdam, 1996).
9. Glezen, W. P. & Couch, R. B. in *Viral Infections of Humans* (eds Evans, A. S. & Kaslow, R. A.) 473–505 (Plenum Medical, New York, 1997).
10. Gill, P. W. & Murphy, A. M. Naturally acquired immunity to influenza type A. A further prospective study. *Med. J. Aust.* **2,** 761–765 (1977).
11. Frank, A. L., Taber, L. H. & Porter, C. M. Influenza B virus reinfection. *Am. J. Epidemiol.* **125,** 576–586 (1987).
12. Frank, A. L. & Taber, L. H. Variation in frequency of natural reinfection with influenza A viruses. *J. Med. Virol.* **12,** 17–23 (1983).
13. Yetter, R. A., Lehrer, S., Ramphal, R. & Small, P. A. Outcome of influenza infection: effect of site of initial infection and heterotypic immunity. *Infect. Immun.* **29,** 654–662 (1980).
14. Sonoguchi, T., Naito, H., Hara, M., Takeuchi, Y. & Fukumi, H. Cross-subtype protection in humans during sequential, overlapping, and/or concurrent epidemics caused by H3N2 and H1N1 influenza viruses. *J. Infect. Dis.* **151,** 81–88 (1985).
15. Yewdell, J. W., Bennink, J., Smith, G. L. & Moss, B. Influenza A virus nucleoprotein is a major target antigen for crossreactive anti-influenza A virus cytotoxic T lymphocytes. *Proc. Natl Acad. Sci. USA* **82,** 1785–1789 (1985).
16. Skoner, D. P. *et al.* Effect of influenza A virus infection on natural and adaptive cellular immunity. *Clin. Immunol. Immunopathol.* **79,** 294–302 (1996).
17. Seo, S. H., Peiris, M. & Webster, R. G. Protective cross-reactive cellular immunity to lethal A/Goose/Guangdong/1/96-like H5N1 influenza virus is correlated with the proportion of pulmonary CD8(+) T cells expressing gamma interferon. *J. Virol.* **76,** 4886–4890 (2002).
18. Sambhara, S. *et al.* Heterosubtypic immunity against human influenza A viruses, including recently emerged avian H5 and H9, induced by FLU-ISCOM vaccine in mice requires both cytotoxic T-lymphocyte and macrophage function. *Cell Immunol.* **211,** 143–153 (2001).
19. McElhaney, J. E., Meneilly, G. S., Pinkoski, M. J., Lechelt, K. E. & Bleackley, R. C. Vaccine-related determinants of the interleukin-2 response to influenza vaccination in healthy young and elderly adults. *Vaccine* **13,** 6–10 (1995).
20. Nicholson, K. G., Webster, R. G. & Hay, A. J. *Textbook of Influenza* (Blackwell Science, Oxford, 1998).
21. Earn, D. J., Dushoff, J. & Levin, S. A. Ecology and evolution of the flu. *Trends Ecol. Evol.* **17,** 334–340 (2002).
22. Layne, S. P. *et al.* A global lab against influenza. *Science* **293,** 1729 (2001).
23. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286,** 1921–1925 (1999).
24. Moser, M. R. *et al.* An outbreak of influenza aboard a commercial airliner. *Am. J. Epidemiol.* **110,** 1–6 (1979).
25. Benenson, A. S. *Control of Communicable Diseases in Man* (American Public Health Association, Washington DC, 1975).
26. Wilson, I. A. & Cox, N. J. Structural basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.* **8,** 737–771 (1990).
27. Smith, C. B., Cox, N. J., Subbarao, K., Taber, L. H. & Glezen, W. P. Molecular epidemiology of influenza A(H3N2) virus reinfections. *J. Infect. Dis.* **185,** 980–985 (2002).
28. Hoskins, T. W., Davies, J. R., Smith, A. J., Miller, C. L. & Allchin, A. Assessment of inactivated influenza-A vaccine after three outbreaks of influenza A at Christ's Hospital. *Lancet* **1,** 33–35 (1979).
29. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony* (Sinauer, Sunderland, 1998).
30. Morbidity and Mortality Weekly Reports (Centers for Disease Control and Prevention, annual reports on influenza frequency, 1986–2002).


**Supplementary Information** accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

**Acknowledgements** We thank N. J. Cox for discussions. N.M.F. thanks the Royal Society, Howard Hughes Medical Institute and Medical Research Council, A.P.G. thanks the Miller Institute, and R.M.B. thanks the NIH for research funding.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to N.M.F. (e-mail: neil.ferguson@ic.ac.uk).

# Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events

**John E. Bowers*, Brad A. Chapman*, Junkang Rong & Andrew H. Paterson**

*Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA*

** These authors contributed equally to this work*

**Conservation of gene order in vertebrates is evident after hundreds of millions of years of divergence[1,2], but comparisons of the *Arabidopsis thaliana* sequence[3] to partial gene orders of other angiosperms (flowering plants) sharing common ancestry ~170–235 million years ago[4] yield conflicting results[5–11]. This difference may be largely due to the propensity of angiosperms to undergo chromosomal duplication ('polyploidization') and subsequent gene loss[12] ('diploidization'); these evolutionary mechanisms have profound consequences for comparative biology. Here we integrate a phylogenetic approach (relating chromosomal duplications to the tree of life) with a genomic approach (mitigating information lost to diploidization) to show that a genome-wide duplication[3,13–17] post-dates the divergence of *Arabidopsis* from most dicots. We also show that an inferred ancestral gene order for *Arabidopsis* reveals more synteny with other dicots (exemplified by cotton), and that additional, more ancient duplication events affect more distant taxonomic comparisons. By using partial sequence data for many diverse taxa to better relate the evolutionary history of completely sequenced genomes to the tree of life, we foster comparative approaches to the study of genome organization, consequences of polyploidy, and the molecular basis of quantitative traits.**

Angiosperms sustain humanity by providing oxygen, medicines, food, feed, fibre, fuel, erosion and flooding control, soil regener-

ation and other benefits[18], warranting increased exploration of their genomic diversity. The ~1,000-fold variation in angiosperm genome size, from 125 million base pairs (Mbp) for *Arabidopsis thaliana*[3] to 124,852 Mbp for *Fritillaria assyriaca*[19], motivates comparative approaches, using data from smaller genomes to accelerate the study of larger genomes. It was recognized recently that even small angiosperm genomes contain much duplication[3,13–17], but robust application of this finding to comparative biology has awaited a means to directly relate chromosomal events to the tree of life in a manner that is not subject to the differing rates of various molecular clocks[20,21].

After revising our earlier analysis of *Arabidopsis thaliana* chromosomal duplication[13] to reflect matches between inferred protein (instead of DNA) sequences and to include 26,028 (instead of 15,199) genes (obtained from NCBI), we circumscribed 34 non-overlapping chromosomal segment pairs that include 23,177 *Arabidopsis* genes (representing 89% of the total) (Fig. 1a). Circumscription of these segment pairs (here called the α group) is conservative, as $\chi^2$ tests comparing the observed number of gene duplications comprising each pair to the number expected in a chromosomal region of equal size if duplications are randomly distributed over the genome, show a maximal random likelihood of $2.8 \times 10^{-84}$ (for α17). Of 2,851 (11%) genes in putatively non-duplicated regions, 1,570 (55%) were near centromeres, and the rest in gaps between the duplications.

After using interpolation to infer an ancestral gene order (see Supplementary Information) that accounts for the composition of the 26 'large' (Fig. 1 legend) α segment pairs that collectively comprise 83% of the transcriptome, a second iteration of analysis revealed additional, more ancient duplications. Nested within the 26 α segment pairs, we circumscribed 29 additional duplications (Fig. 1b). All 29 segments are mosaics of genes from each of their four 'descendant' modern chromosome segments, and most are only evident by analysis of the inferred ancestral gene order. The 29 segment pairs comprise two subpopulations (called here β and γ). β01–β22 include 13,449 genes (51.6% of the transcriptome) and are non-overlapping, with a maximal probability of random occurrence of $3 \times 10^{-3}$ (for β19). β18, 20 and 22 each comprise <1% of the transcriptome, and like the 'small' α pairs were grouped for further analyses. γ01–γ07 include 5,287 genes (20.3% of the transcriptome) with a maximal probability of random occurrence of $<2 \times 10^{-3}$ (for γ04), and overlap with many β segments. Distinct identities of the β and γ groups are reinforced by the finding that among 78 cases in which protein matches were found in both groups, 59 (75%) showed greater distance between the members of the γ pairs.

To relate the events that produced the α, β and γ segment pairs to the angiosperm family tree, we compared all syntenic *Arabidopsis* gene pairs from each duplication event (respectively), to individual genes from representatives of the gymnosperms (Pinaceae), monocotyledonous angiosperms (*Oryza*), asterids (Solanaceae), distantly related rosids (*Glycine*, *Medicago*), closely related rosids (Malvaceae) and a confamilial Brassicaceae genus (*Brassica*), respectively (see Fig. 2 for GenBank taxon IDs). Although the ideal comparison would involve analysis of individual gene trees that included full-length sequences for all taxa simultaneously, the fragmentary nature of expressed sequence tags (ESTs) that comprise most plant gene databases necessitated a modified approach. First, using genes for which adequate sequence information existed (based on criteria described in Fig. 2 and Methods), we made phylogenetically rooted trees using *Physcomitrella* (a moss) as an outgroup, then evaluated the frequency at which heterologous proteins clustered internally or externally to the *Arabidopsis* duplicates. For many additional genes, it was possible to determine whether inferred protein sequences from the syntenic *Arabidopsis* genes were more, or less, similar to one another than to the heterologous protein by evaluating PAM (per cent accepted mutation)-based pairwise distances (Fig. 2 and Methods). Both analyses considered only matches of ≥35 amino acids (105 nucleotides). (Lists of genes and their arrangements in the α, β and γ duplications, inferred ancestral gene orders, frequencies of internal versus external rooted trees and PAM-based pairwise distances for each α, β and γ segment pair, and the accession numbers plus chromosomal and map locations of the cotton DNA sequences used for synteny analysis are available as Supplementary Information.)

The *Arabidopsis* α duplication event pre-dated its divergence from *Brassica* about 14.5–20.4 million years (Myr) ago[4], but post-dated its divergence from the Malvaceae 83–86 Myr ago[22]. Rooted trees and PAM-based pairwise distances (Fig. 2b) showed that 49% and 64% (respectively) of *Brassica* sequences were more similar to one duplicated *Arabidopsis* sequence than was the other *Arabidopsis* sequence. By contrast, only 6–19% of Malvaceae and more distant sequences clustered internally to the *Arabidopsis* syntenic duplicates.

The β event pre-dated *Arabidopsis* divergence from the other dicots studied, but post-dated divergence from the monocots about 170–235 Myr ago[4]. Rooted trees and pairwise distances revealed internal clustering rates of 43–79% across the dicots, but only 14–33% with monocot or gymnosperm ESTs. The frequencies of internal clustering with *Oryza* rooted trees overlapped both the dicot and gymnosperm values (which did not overlap one another), but the larger number of pairwise distances differentiated *Oryza* from the dicots.

The γ event appears to pre-date monocot–dicot divergence. Predominantly internal clustering (47–87%) was found for sequences from all angiosperms studied. For gymnosperms, internal clustering was less frequent (31–47%) but not significantly different from that for angiosperms. Relating the γ event to the angiosperm–gymnosperm divergence ~300 Myr ago[23] awaits more data.

Different models for *Arabidopsis* karyotypic evolution can be explored by considering the antiquity of duplication events, and the size of inferred duplicate chromosomal segments. The 'footprints' of the α event of ≤86 Myr ago include 57 adjacent syntenic regions with opposite orientation and order explicable by localized inversions (Fig. 1 legend) that cover 89% of the genome, with an average length of about 9 centimorgans (cM). Using an estimated rate of structural mutations per chromosome pair per million years of divergence[24] ($0.14 \pm 0.06$), and assuming that the present $n = 5$ chromosomes has typified the *Arabidopsis* lineage in the past, about 60 rearrangements would be predicted, yielding modern chromosomal segments that average 8.3 cM in length. The possibility that $n = 8$ better represents the taxon's history[25] would suggest greater divergence, with modern segments averaging 5.2 cM. If the lineage has varied between five and eight chromosomes, the fit of observed to predicted[24] values improves.

Establishing the provenance of ancient genome-wide duplication events revises and extends understanding of angiosperm evolutionary history, showing that much if not all of the flowering plant lineage is palaeo-polyploid. Although our α event falls squarely in the range (65–100 Myr ago) during which much evidence supports a genome-wide duplication[3,13–17], only two studies infer additional duplications. (1) Vision *et al.*[15] inferred four 'age classes' of duplications that bound 48, 39, 11 and 3% of the *Arabidopsis* genes, respectively, dismissing two additional classes as probable artefacts. However, <25% of all genes fall in two or more blocks, thus only one age class (C, 48%) could be inferred to involve most of the genome. Our α event involves 89% of *Arabidopsis* genes, close to the total of all four age classes[15]. By inferring pre-duplication gene orders, then searching for more ancient duplication, we show that the β event involved ≥51.6% of the genes, more than the largest age class[15] and virtually all of which were also involved in the α event. The γ event, which may be 100 Myr older than the most ancient age class[15], covers more of the transcriptome (20.3%) than two of the

four age classes. (2) A recent study[17] using synonymous (third-nucleotide, $K_s$) rather than overall[15] nucleotide substitution rates also contraindicates the four age classes, instead supporting our inference of two to three genome-wide duplications based on $K_s$ values similar to those found among our syntenic gene pairs (1.18 ± 0.69, 2.33 ± 1.03 and 2.82 ± 1.16 for α, β and γ, respectively, calculated as described[26]).

Using consensus among many gene trees to relate *Arabidopsis*

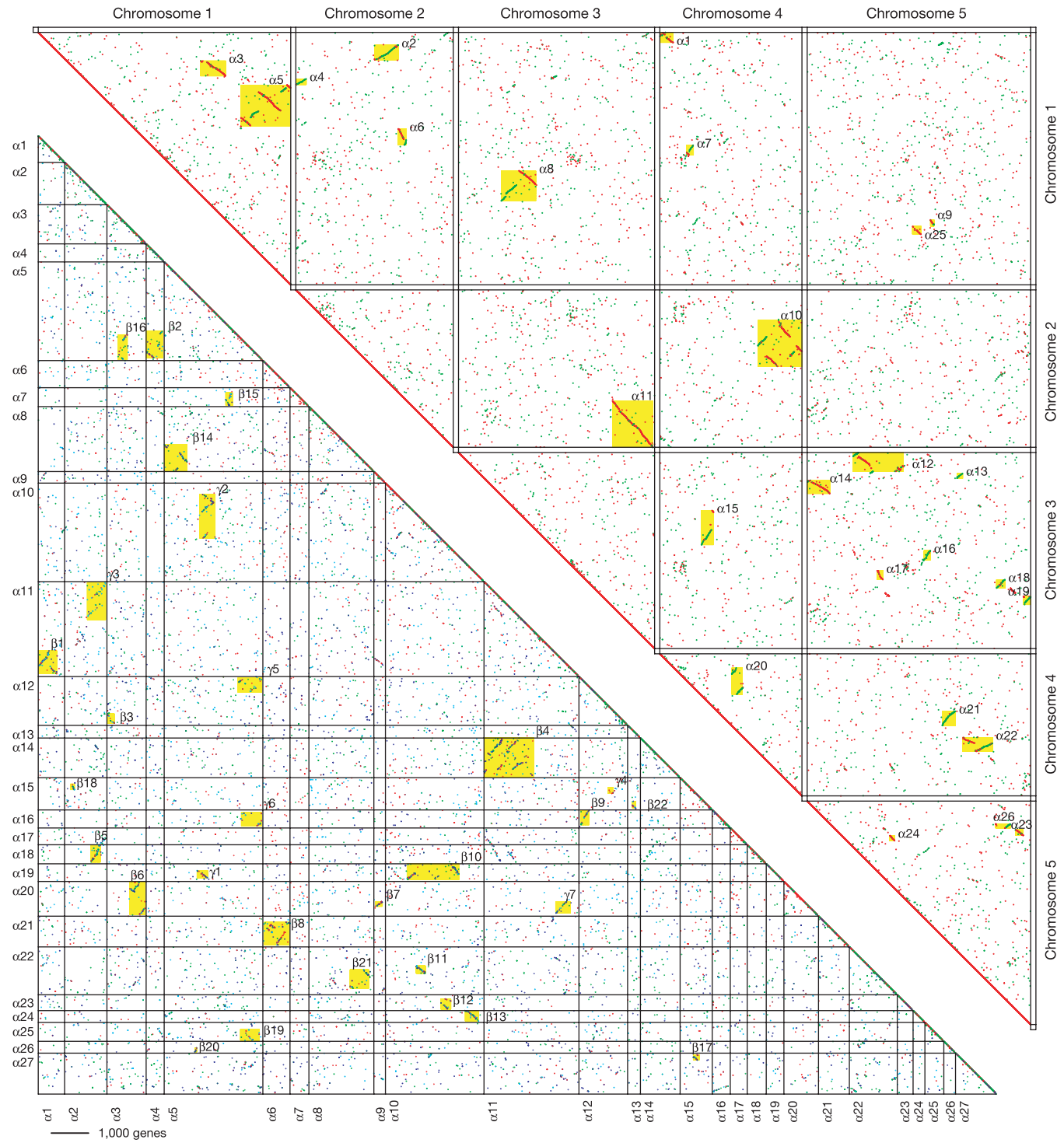
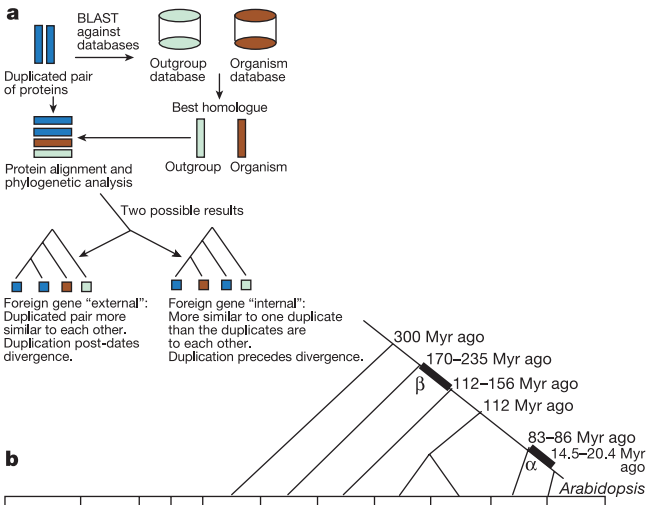



**Figure 1** Arrangement of duplicated protein-encoding genes in *Arabidopsis thaliana*. Top right, α duplications. Both *x* and *y* axes represent 26,028 genes in their chromosomal order. The best-matching gene pairs are plotted, colour-coded to indicate same (red) or opposite (green) transcriptional orientations. For further analysis, 57 adjacent duplicated regions with opposite orientation and order explicable by localized inversions were combined into 26 'large' duplications (α01–α26) that each included ≥1% (260) of the genes. Eight shorter duplications were pooled (α27). Lower left, β and γ duplications. Both *x* and *y* axes represent 21,749 genes, in an inferred ancestral order that accounts for the composition of the 26 large α duplications (at left and bottom). Twenty-nine β or γ duplications (see text) are highlighted. Colours show how the four modern *Arabidopsis* chromosome segments contribute to β or γ duplications, distinguishing contributions to the segments at left and bottom respectively from the: (1) lower-numbered chromosomes (red); (2) higher- and lower-numbered chromosomes (light blue); (3) lower- and higher-numbered chromosomes (dark blue); (4) higher-numbered chromosomes (green). Higher-resolution versions of the figure and lists of gene orders are available (see Supplementary Information).




| Duplicated segment | Analysis method | No. of genes | % syntenic duplicates | Pinaceae (TaxID: 3318) | Oryza (TaxID: 4527) | Solanaceae (TaxID: 4070) | Medicago (TaxID: 3877) | Glycine (TaxID: 3846) | Malvaceae (TaxID: 3629) | Brassica (TaxID: 3705) |
|---|---|---|---|---|---|---|---|---|---|---|
| Event 1 (α01-27) | Rooted trees | 23177 | 29.7 | 0.06$^{C}$ (1671) | 0.12$^{BC}$ (1280) | 0.18$^{B}$ (2095) | 0.19$^{B}$ (2116) | 0.18$^{B}$ (1950) | 0.17$^{B}$ (1832) | 0.49$^{A}$ (1836) |
| Event 1 (α01-27) | Distance | | | 0.11$^{C}$ (2587) | 0.12$^{C}$ (2011) | 0.19$^{B}$ (3322) | 0.20$^{B}$ (3390) | 0.19$^{B}$ (3130) | 0.19$^{B}$ (2854) | 0.65$^{A}$ (3259) |
| Event 2 (β01-22) | Rooted trees | 13449 | 16.3 | 0.14$^{B}$ (666) | 0.24$^{AB}$ (500) | 0.44$^{A}$ (829) | 0.45$^{A}$ (844) | 0.44$^{A}$ (809) | 0.43$^{A}$ (746) | 0.58$^{A}$ (718) |
| Event 2 (β01-22) | Distance | | | 0.28$^{C}$ (1031) | 0.33$^{C}$ (744) | 0.59$^{B}$ (1281) | 0.66$^{B}$ (1289) | 0.64$^{AB}$ (1255) | 0.57$^{B}$ (1121) | 0.79$^{A}$ (1259) |
| Event 3 (γ01-07) | Rooted trees | 5287 | 13.1 | 0.31 (257) | 0.47 (182) | 0.70 (297) | 0.66 (305) | 0.66 (288) | 0.59 ( 277) | 0.67 (284) |
| Event 3 (γ01-07) | Distance | | | 0.47 (377) | 0.59 (263) | 0.87 (470) | 0.86 (485) | 0.87 (441) | 0.74 (421) | 0.83 (451) |

**Figure 2** Phylogenetic methodology for dating *Arabidopsis* duplications. **a**, Schematic of data flow for rooted gene tree and/or PAM-based pairwise protein distance analysis described in Methods. **b**, For each taxon (as listed) and analysis method (rooted trees, pairwise distances), the fraction of internal trees with superscripted letters indicating differences significant at $P \geq 0.05$, and number of trees that could be analysed (parentheses) are shown, and aligned with phylogenetic relationships among the taxa (above). Taxonomic divergence dates are as cited in the text. Detailed data for each individual α, β and γ segment pair are available (see Supplementary Information).

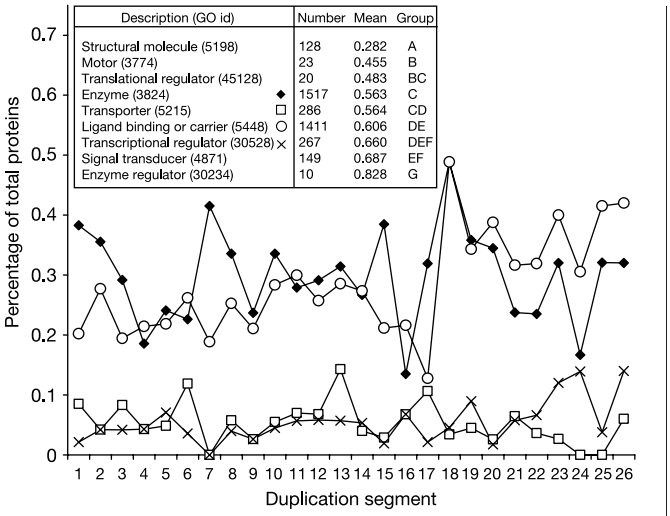



**Figure 3** Gene family composition contributes to different levels of divergence between duplicated chromosomal segments. Inset, Gene- Ontology-based (http://www.geneontology.org) molecular function annotations of syntenic duplicated gene pairs, including the number in each category, and mean protein similarity values (calculated by protdist). Tukey's analysis of square-root-transformed protein distances shows significant differences (A–G) in the degree of similarity for different gene families. Main panel; for α01–α26 duplications, percentages of the total proteins (genes) are plotted for the five most abundant protein families. A duplicate pair can be in more than one family if different domains perform different functions, so the sum can exceed 100%.

duplication events to its divergence from other angiosperms addresses significant pitfalls of 'dating' approaches based on molecular clocks[15–17]. Within each duplication event, much variation exists among syntenic gene pairs in the frequencies of internal/external clustering, but we question the interpretation[15–17] that such variation largely reflects the antiquity of duplication events. Individual chromosome segment pairs vary widely in the abundance of different classes of genes, which in turn have different degrees of conservation (Fig. 3). For example, structural proteins are well conserved, and enzyme regulators are especially diverse. Substantial differences in the extent of divergence between segment pairs (for example, average $K_s$ values) can be explained simply by this sampling variation. A recent report[21] of $K_s$ variation among plant genes of up to 14-fold, generally higher than previously reported, urges further caution in equating $K_s$ with time.

A consensus of data from many genes is needed to 'date' duplication events, in that occasional false trees and/or distance estimates may occur due to rapid evolution of one duplicated copy, ancient proximal duplication followed by deletion of the true orthologue, failure to sample true orthologues in EST data, interlocus convergence such as gene conversion, or other factors. For example, the Pinaceae–*Arabidopsis* divergence[23] far pre-dates the α event, yet pine sequences show 6–11% internal trees.

By merging a phylogenetic approach to relating chromosomal duplications to the tree of life, with a genomic approach to mitigating information lost to diploidization, increased levels of angiosperm synteny (and associated opportunities for comparative biology) may be revealed. Virtually all previous comparative studies, including ours[13], may have underestimated synteny between taxa by using 'one-to-one' comparisons to *Arabidopsis*[5–11], which are appropriate only for closely related taxa that diverged from *Arabidopsis* after the α event (such as other Brassicaceae). Phylogenetic data provide a framework for making more appropriate intergenomic comparisons, by determining whether chromosomal duplications within taxa pre-date or post-date divergence among taxa. Although abundant genomic duplication makes this approach especially important in the angiosperms, it may contribute to better genomic comparisons in many lineages.

In large angiosperm genomes that are not likely to be sequenced soon (such as many major crops), much benefit might be gained from the sequences of botanical models by alignment of detailed genetic maps based upon conserved sequence-tagged sites. Comparison to a detailed cotton map (Fig. 4) showed that the two different members of most *Arabidopsis* α segment pairs provide complementary synteny information. For each member of the 26 large α segment pairs, we determined the number of 'best matches' (using BLASTN with an expectation value of $E < 10^{-9}$) with genes on each of the 13 cotton chromosomes. To normalize over the different gene numbers on each cotton chromosome and *Arabidopsis* segment, we subtracted the random expectation (calculated by standard contingency methods). The paired *Arabidopsis* segments showed highly significant correlation ($r = 0.197$, 336 d.f., $P < 0.01$) of the residual frequencies of 'best matches' to the 13 cotton chromosomes. Comparisons of non-paired segments to one another showed no correlation ($r = -0.02$). Only after assembling the two members of α segment pairs into inferred ancestral orders did patterns of *Arabidopsis*–cotton synteny become clear (to be fully described elsewhere but see example, Fig. 4). In addition to its applied value for cotton improvement, this finding further supports the view that the α event post-dated *Arabidopsis*–Malvaceae divergence. Using entire cotton chromosomes to assess the predictive value associated with pairing of duplicated *Arabidopsis* segments is very conservative, but is at present necessary. Sufficient data to analyse synteny in the 8.3-cM-average bins predicted[24] to persist after ≤86 Myr of divergence awaits mapping of more cotton genes.

Of special interest is synteny between monocots and eudicots, the two main branches of the angiosperms. One-to-one comparisons[27] show discernible parallels of the partial *Oryza* (rice) sequence to

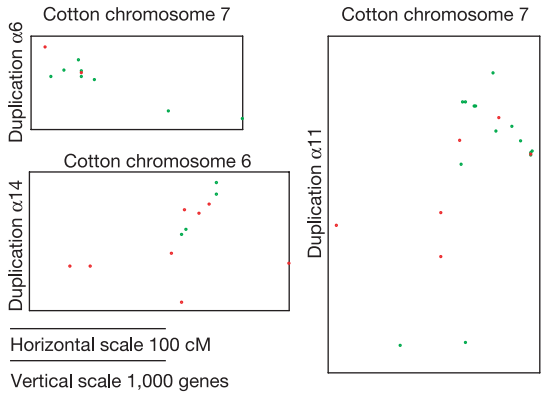



**Figure 4** Examples of *Arabidopsis*–cotton synteny. Axes reflect relative gene orders (*Arabidopsis*) and map distances in cM (cotton). Cotton homoeologous series 7 corresponds in non-overlapping regions to *Arabidopsis* α06 (derived from parts of chromosomes 1 (red) and 2 (green)) and α11 (from chromosomes 2 (red) and 3 (green)) based on 10 and 20 matches respectively, numbers of matches expected to occur by chance at likelihoods of $2.6 \times 10^{-5}$ and $1.9 \times 10^{-3}$ based on $\chi^2$ analysis. A sub-central region of cotton chromosome 6 corresponds to *Arabidopsis* α14 (from chromosomes 3, 5) based on 12 matches, and a chance likelihood of $2.5 \times 10^{-5}$.

*Arabidopsis*, although "… conservation is less extensive than previously predicted…" (in ref. 24). Factoring in two *Arabidopsis* duplications (α, β) plus at least one *Oryza* duplication[27] that post-date divergence of these taxa is likely to increase the level of conservation found.

By this synergistic approach, analysis and interpretation of whole-genome sequences benefits from partial sequence data for larger-genome taxa, in turn building contextual information important to using comparative approaches to accelerate analysis of larger genomes. Improved delineation of comparative gene arrangements promises new insights into fundamental questions, and invigorated progress towards practical applications of genomics. For example, the notion that particular gene arrangements may confer increased fitness has long awaited sufficient data to discern conserved organization of chromosomal segments that are larger, or contain more genes, than would be explicable by chance. Detailed study of the levels and patterns of diploidization and divergence in different gene families, and divergent taxa, may begin to shed light on the long-debated adaptive significance, if any, of polyploidy. Much knowledge exists about genomic changes that follow polyploidization[12], but little is known about which specific events underlie the co-evolution of divergent genomes in a common nucleus to yield new phenotypes. Only 30% of *Arabidopsis* genes have retained syntenic copies in the ≤86 Myr since the α duplication—comparison to the 70% synteny of human and mouse proteins after 100 Myr (ref. 1) highlights the potential impact of gene loss on angiosperm evolution.

Better delineation of comparative gene arrangements may also aid in annotation of genomic sequences for other angiosperms, and also promises increased use of rapidly expanding knowledge about the action(s) of individual genes in facile models to identify convergent or parallel evolution of allelic variants important to agriculture or development. Such a comparative approach may especially accelerate progress towards identifying those genes that underlie complex physiological, morphological or behavioural phenotypes, and are discernible only as quantitative trait loci[28]. □

## Methods

### Duplication analysis

A total of 26,028 *Arabidopsis* gene sequences were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), encoded by their chromosomal order and transcriptional orientation, and compared to each other using BLASTP[29]. Only the top five, or two, non-self protein matches that met a threshold of $E < 10^{-10}$ were considered in α-duplication, or β/γ-duplication analysis, respectively. Circumscription of individual duplicated segments was largely as described[13], using additional data to extend the limits of a duplication if a BLAST match revealed a pair of genes in consistent orientation within 20 or fewer genes (counted by combining both members of the pair) from the prior terminus. Adjacent syntenic regions with opposite orientation and order explicable by localized inversions were combined for further analyses, arbitrarily designating the linear order of the lower-numbered chromosomal region as ancestral, and assigning the inversion to the higher-numbered chromosome. In six cases, the proposed inversions involved two duplication regions.

### Gene tree analysis

Each duplicated syntenic gene pair was compared to each taxon-specific sequence (using nucleotide databases created by batch NCBI download of taxon IDs 3318, 3629, 3705, 3846, 3877, 4070 and 4527) using TBLASTN, results parsed, and the best match exceeding $E < 10^{-5}$ translated to protein in the frame of the TBLASTN match, then (using BLAST) checked against all *Arabidopsis* proteins to confirm that the original genes were recovered. Rooted trees were made by including *Physcomitrella* (taxon ID 3217) sequences. The most similar regions between a duplicated pair, the best-matching homologue and the moss outgroup (where relevant) were determined using pairwise alignment as implemented by the 'water' program in EMBOSS (http://www.emboss.org), then aligned using CLUSTALW version 1.82 and the default parameters (PAM matrix; gap opening penalty = 10.0; gap extension penalty = 0.2). For rooted analyses, 100 bootstrap replicates were created using the EMBOSS interface to the 'seqboot' program in PHYLIP version 3.6 (http://evolution.genetics.washington.edu/phylip.html), and protein parsimony trees created using the 'protpars' program in PHYLIP (default parameters). A consensus tree rooted with moss was created using the EMBOSS interfaces to the 'consense' program in PHYLIP, and the resulting tree file was parsed to determine the position of the taxon-specific sequence relative to the *Arabidopsis* duplicates. For non-rooted analyses, protein distances for *Arabidopsis* duplicates and the other taxon were calculated using 'protdist' (PAM matrix, default parameters) in PHYLIP, and compared to determine the most similar pair. NCBI downloads, BLAST parsing, sequence translation and phylogenetic analyses were automated using Python scripts available from http://www.plantgenome.uga.edu/project-bioinformatics.htm.

The fractions of 'internal trees' associated with each duplication cycle were compared using one-way analysis of variance (ANOVA) for correlated samples and Tukey's HSD analysis for post-ANOVA comparisons between organisms (R. Lowry, chapter 15 in http://faculty.vassar.edu/lowry/webtext.html). Individual duplicated segment pairs were considered treatments, and the indicated taxa were conditions, accounting for correlations that may result from comparing identical genes in different taxa. This is conservative, because in many cases different regions of an *Arabidopsis* gene matched ESTs from different taxa, reducing the correlation problem.

### *Arabidopsis*–cotton synteny

The order of 2,102 sequence-tagged sites (GenBank accession numbers provided in Supplementary Information) along the chromosomes of a hypothetical cotton progenitor pre-dating A/D-subgenome divergence was inferred by using anchor probes mapped in both diploid and tetraploid cottons[30] plus new loci to interpolate the locations of sequence-tagged sites mapped in only a subset of these populations. This map was compared to the inferred ancestral gene order pre-dating the *Arabidopsis* α duplication by finding the most similar BLASTN match of $E < 10^{-9}$ to any *Arabidopsis* gene.

Received 25 October 2002; accepted 5 February 2003; doi:10.1038/nature01521.

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
2. Smith, S. S. *et al.* Analyses of the extent of shared synteny and conserved gene orders between the genome of *Fugu rubripes* and Human 20q. *Genome Res.* **12,** 776–784 (2002).
3. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
4. Yang, Y.-W., Lai, K.-N., Tai, P.-Y. & Li, W.-H. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J. Mol. Evol.* **48,** 597–604 (1999).
5. Salse, J., Benoit, P., Cooke, R. & Delseny, M. Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing Rice Genome Sequencing Project. *Nucleic Acids Res.* **30,** 2317–2328 (2002).
6. Mayer, K. *et al.* Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. *Genome Res.* **11,** 1167–1174 (2001).
7. Ku, H., Vision, T., Liu, J. & Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA* **97,** 9121–9126 (2000).
8. Grant, D., Cregan, P. & Shoemaker, R. C. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **97,** 4168–4173 (2000).
9. Lee, J. M., Grant, D., Vallejos, C. E. & Shoemaker, R. C. Genome organization in dicots. II. *Arabidopsis* as a 'bridging species' to resolve genome evolution events among legumes. *Theor. Appl. Genet.* **103,** 765–773 (2001).
10. Rossberg, M. *et al.* Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and Capsella genomes. *Plant Cell* **13,** 979–988 (2001).
11. Liu, H., Sachidanandam, R. & Stein, L. Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. *Genome Res.* **11,** 2020–2026 (2001).
12. Eckhardt, N. A sense of self: The role of DNA sequence elimination in allopolyploidization. *Plant Cell* **13,** 1699–1704 (2001).
13. Paterson, A. H. *et al.* Comparative genomics of plant chromosomes. *Plant Cell* **12,** 1523–1539 (2000).

14. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12,** 1093–1101 (2000).
15. Vision, T. D., Brown, D. B. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290,** 2114–2117 (2000).
16. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
17. Simillion, C., Vandepoele, K., Van Montagu, M. C. E., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99,** 13627–13632 (2002).
18. Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* **418,** 671–677 (2002).
19. Bennett, M. D. & Smith, J. B. Nuclear DNA amounts in angiosperms. *Proc. R. Soc. Lond. B* **274,** 227–274 (1976).
20. Strauss, E. Can mitochondrial clocks keep time? *Science* **283,** 1435–1438 (1999).
21. Zhang, L., Vision, T. J. & Gaut, B. S. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19,** 1464–1473 (2002).
22. Benton, M. J. *The Fossil Record 2* (Chapman and Hall, New York, 1993).
23. Bowe, L. M., Coat, G. & dePamphilis, C. W. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA* **97,** 4092–4097 (2000).
24. Paterson, A. H. *et al.* Toward a unified genetic map of higher plants, transcending the monocot–dicot divergence. *Nature Genet.* **14,** 380–382 (1996).
25. Koch, M., Bishop, J. & Mitchell-Olds, T. Molecular systematics and evolution of *Arabidopsis* and *Arabis*. *Plant Biol.* **1,** 529–537 (1999).
26. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32–43 (2000).
27. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
28. Paterson, A. H. *et al.* Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335,** 721–726 (1988).
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
30. Brubaker, C. L., Paterson, A. H. & Wendel, J. F. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **42,** 184–203 (1999).

**Supplementary Information** accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

**Acknowledgements** We thank A. Feltus, J. C. Kissinger and S. Schulze for comments on the manuscript, and the Paterson Lab for technical support. This work was supported by the US Department of Agriculture National Research Initiative and Initiative for Future Agriculture and Food Safety, the US National Science Foundation Plant Genome Research Program, the Howard Hughes Medical Institute Graduate Fellowship Program, the International Consortium for Sugarcane Biotechnology, the Georgia Cotton Commission/Cotton Inc., and the Georgia Agricultural Experiment Station.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to A.H.P. (e-mail: paterson@uga.edu).

# The role of presenilin cofactors in the γ-secretase complex

**Nobumasa Takasugi*, Taisuke Tomita*, Ikuo Hayashi*, Makiko Tsuruoka*, Manabu Niimura*, Yasuko Takahashi*, Gopal Thinakaran† & Takeshi Iwatsubo***

** Department of Neuropathology and Neuroscience, Graduate School of Pharmaceutical Sciences, University of Tokyo, 7-3-1 Hongo, Bunkyoku, Tokyo 113-0033, Japan*
*† Department of Neurobiology, Pharmacology and Physiology, The University of Chicago, 924 East 57th Street, Chicago, Illinois 60637, USA*

**Mutations in presenilin genes account for the majority of the cases of the familial form of Alzheimer's disease (FAD). Presenilin is essential for γ-secretase activity, a proteolytic activity involved in intramembrane cleavage of Notch and β-amyloid precursor protein (βAPP)[1,2]. Cleavage of βAPP by FAD mutant presenilin results in the overproduction of highly amyloidogenic amyloid β42 peptides[3–6]. γ-Secretase activity requires the formation of a stable, high-molecular-mass protein complex[7–11] that, in addition to the endoproteolysed fragmented form of presenilin, contains essential cofactors including nicastrin[12–14], APH-1 (refs 15–18) and PEN-2 (refs 16, 19). However, the role of each protein in complex formation and the generation of enzymatic activity is unclear. Here we show that *Drosophila* APH-1 (Aph-1) increases the stability of *Drosophila* presenilin (Psn) holoprotein in the complex. Depletion of PEN-2 by RNA interference prevents endoproteolysis of presenilin and promotes stabilization of the holoprotein in both *Drosophila* and mammalian cells, including primary neurons. Co-expression of *Drosophila* Pen-2 with Aph-1 and nicastrin increases the formation of Psn fragments as well as γ-secretase activity. Thus, APH-1 stabilizes the presenilin holoprotein in the complex, whereas PEN-2 is required for endoproteolytic processing of presenilin and conferring γ-secretase activity to the complex.**

Presenilin is essential for γ-secretase cleavage, which releases amyloid β-peptide (Aβ) and the intracellular domain of Notch by intramembranous proteolysis of β-amyloid precursor protein (βAPP) and Notch, respectively[1,2]. Presenilin mediates γ-secretase function by forming a highly stable protein complex of high relative molecular mass (high-$M_r$) together with a set of cofactor proteins[7–11]. In addition to nicastrin (NCT), a type I single-pass membrane glycoprotein[12], two additional putative presenilin cofactors have been identified: APH-1, a multi-transmembrane protein coded by a gene whose deletion leads to hypoplasia of the anterior pharynx in *Caenorhabditis elegans*, was found to be a Notch pathway member possibly involved in presenilin function[15]; *aph-1* was also identified as one of the presenilin enhancer genes (*pen-1*) together with *pen-2*, which codes for a double-membrane-spanning protein[16]. NCT, APH-1 and PEN-2 are required for γ-secretase function and accumulation of presenilin fragments[12–14,16–19], although the differential roles of each cofactor in the formation of the high-$M_r$ presenilin protein complex, and whether NCT, APH-1 and PEN-2 represent the principal presenilin cofactors to confer γ-secretase activity, remain elusive.

We stably transfected *Drosophila* S2 cells with Aph-1, and found a significant increase in the levels of endogenous Psn holoprotein. This was markedly enhanced by expression of Aph-1 and *Drosophila* NCT (Nct), but was not observed in cells transfected with Nct alone. The levels of Psn fragments involved in the active form of γ-secretase were not altered in either case (Fig. 1a). We next treated stably co-transfected S2 cells overexpressing Aph-1 and Nct with cycloheximide (CHX) to block total cellular protein synthesis, and examined the stability of Psn and other proteins. In mock-transfected S2 cells (transfected with an empty vector alone) only small amounts of Psn holoprotein were detectable, which rapidly degraded within about 4 h of CHX treatment; however, fragments of Psn were relatively more abundant and highly stable, in a manner similar to that of mammalian presenilin[7,11] (Fig. 1b, left panel). In contrast, Psn holoprotein levels were significantly increased by overexpression of Aph-1 or of Aph-1 and Nct, and remained highly stabilized (Fig. 1b, middle and right panels, respectively). Levels of Aph-1 and Nct in co-transfected S2 cells were also highly stable during the period of CHX treatment, suggesting that Psn (including holoprotein) forms a highly stabilized protein complex together with Aph-1 and Nct under these conditions (Fig. 1b, middle and right panels). To gain support for this hypothesis, we then separated CHAPSO-solubilized membrane fractions of S2 cells stably transfected with Aph-1 and Nct by glycerol velocity gradient centrifugation[10,11]. Psn holoprotein in cells overexpressing Aph-1 and Nct was fractionated totally in high-$M_r$ ranges of 232–440K together with Psn fragments (Fig. 1c, lower panel, FL). This was in contrast to the exclusive low-$M_r$ distribution of short-lived Psn holoproteins in S2 cells in the absence of Aph-1 overexpression (Fig. 1c, upper panel, FL). Furthermore, most of the Aph-1 and Nct proteins were found in the high-$M_r$ fractions (Fig. 1c, lower panel). Taken together, our data support the hypothesis that Aph-1 represents