Inferring complex phylogenies

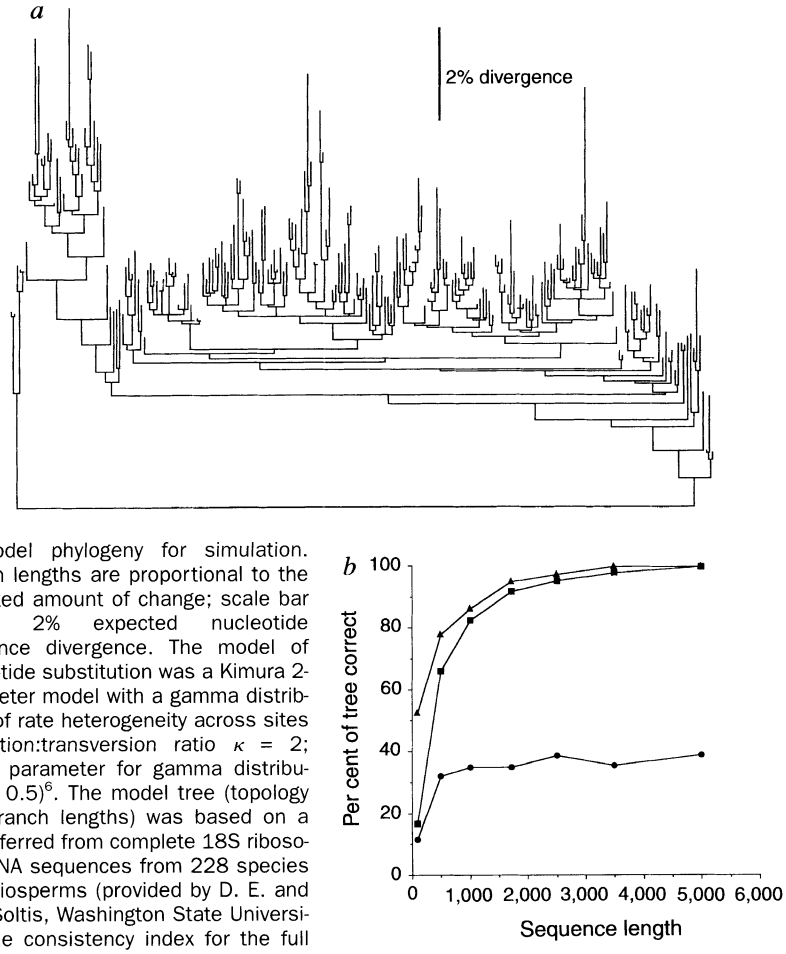
SIR — Previous studies have demonstrated that immense molecular data sets are often needed for accurate phylogenetic estimation. For instance, under some conditions of extreme evolutionary rate variation, correct recovery of the phylogeny of just four taxa requires both an accurate DNA substitution model and information on tens of thousands to millions of nucleotides (sometimes exceeding the size of whole genomes)[1]. These facts, together with the large number of possible solutions, have led many systematists to question the practicality of obtaining accurate reconstructions for phylogenies involving hundreds or thousands of species[2,3]. Despite this scepticism, however, there is considerable interest in phylogenetic analyses of many sequences, such as studies of the radiation of flowering plants[4] or global diversity of human mitochondrial genomes[5].

To test the feasibility of such studies and to predict the length of sequence necessary to infer complex phylogenies accurately, I simulated the phylogeny shown in part *a* of the figure. The model tree is based on an estimated phylogeny of 228 species of angiosperms inferred from complete 18S ribosomal RNA genes, and contains considerable rate heterogeneity among lineages and many short internal branches. The model of DNA evolution for the simulation included a transition:transversion bias and rate heterogeneity across sites (see figure legend). Expected levels of change along branches were based on the observed divergences in the 18S ribosomal DNA data.

I tested three simple algorithms for approximating phylogenetic solutions: stepwise addition under the parsimony criterion; neighbour joining under the minimum evolution criterion; and the unweighted pair-group method of arithmetic averages (UPGMA) clustering algorithm[6]. The results (*b* in the figure) demonstrate that the model phylogeny can be accurately reconstructed with sequences only 5,000 nucleotides long. Considerably more nucleotides are needed to reconstruct some four-taxon trees[1].

This result is surprising because there are only three possible solutions for the four-taxon problem, but approximately 1.2×10^{502} solutions for the 228-taxon problem. Even simple methods that incorporate little of the evolutionary complexity of the model (such as parsimony) perform remarkably well (and actually outperform neighbour joining, even when pairwise distances match the model of substitution exactly). However, only the UPGMA method, which is highly sensitive to unequal rates of change[6], failed to reconstruct more than 99% of the branches in the model tree with 5,000 nucleotides. Furthermore, branch-swapping on the stepwise-addition parsimony tree (based on the 5,000-nucleotide data set) revealed four equally good solutions, one of which matches the model tree exactly — including two unresolved polytomies (the other three solutions are various resolutions of these two polytomies). Branch-swapping on initial trees estimated from smaller data sets revealed hundreds to tens-of-thousands of equally good solutions (all of which were within 4% accuracy of the initial solution).

The explanation for the unexpected ease of reconstruction appears to lie in the dispersal of noise in the data set (homoplasy, or convergences and reversals of character states) across the many branches in the tree, so that covarying patterns of homoplasy between any two taxa are relatively rare. This allows the phylogenetic signal, even where weak, to be detected



a, Model phylogeny for simulation. Branch lengths are proportional to the expected amount of change; scale bar shows 2% expected nucleotide sequence divergence. The model of nucleotide substitution was a Kimura 2-parameter model with a gamma distribution of rate heterogeneity across sites (transition:transversion ratio $\kappa = 2$; shape parameter for gamma distribution = 0.5)[6]. The model tree (topology and branch lengths) was based on a tree inferred from complete 18S ribosomal DNA sequences from 228 species of angiosperms (provided by D. E. and P. S. Soltis, Washington State University). The consistency index for the full simulated data set was 0.43. The C source code used for the simulations can be obtained from the World Wide Web at http://mw511.biol.berkeley.edu/software.html (siminator.c, by J. P. Huelsenbeck); details of the model tree are available from the author (hillis@bull.zo.utexas.edu). *b*, Efficiency of three algorithms for estimating the model tree from the simulated data sets. All analyses were conducted using PAUP* 4.0d42 (provided by D. L. Swofford, Smithsonian Institution). In parsimony analyses (stepwise addition, triangles), all changes and sites were weighted equally. Neighbour joining (squares) and UPGMA analyses (circles) were based on pairwise distances that matched the model of evolution (maximum likelihood distances, with the appropriate parameters for modelled transition:transversion ratio and gamma distribution of rate heterogeneity)[6]. Results based on other pairwise distances (including uncorrected *p*-distances and Kimura 2-parameter distances with gamma correction) were nearly indistinguishable from those shown, and the corresponding trees were never more than 2% different in accuracy from the trees based on maximum likelihood distances.

1. Hillis, D. M., Huelsenbeck, J. P. & Swofford, D. L. *Nature* **369**, 363–364 (1994).
2. Patterson, C., Williams, D. M. & Humphries, C. J. *Annu. Rev. Ecol. Syst.* **24**, 153–188 (1993).
3. Hillis, D. M. *Syst. Biol.* **44**, 3–16 (1995).
4. Chase, M. W. *et al. Ann. Missouri Bot. Gard.* **80**, 528–580 (1993).
5. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. *Science* **253**, 1503–1507 (1991).
6. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* 2nd edn (eds Hillis, D. M, Moritz, C. & Mable, B. K.) 407–514 (Sinauer, Sunderland, Massachusetts, 1996).
7. Lake, J. A. *Mol. Biol. Evol.* **4**, 167–191 (1987).
8. Graur, D., Duret, L. & Guoy, M. *Nature* **379**, 333–335 (1996).
9. Huelsenbeck, J. P. & Hillis, D. M. in *Molecular Zoology: Advances, Strategies, and Protocols* (eds Ferraris, J. D. & Palumbi, S. R.) 19–45 (Wiley-Liss, New York, 1996).

above the background noise. Although the amount of data needed for accurate reconstruction of such complex trees is approximately two to three times greater than has been collected to date for angiosperms, this study demonstrates that at least some problems of this order of complexity are tractable with existing methods. This simulation also demonstrates that adding large numbers of additional taxa to phylogenetic analyses may increase the accuracy of the estimated trees and at the same time reduce the need for computationally complex methods of analysis.

These results run counter to a recent trend to break phylogenetic problems into a series of four-taxon questions[7,8]; instead, including large numbers of taxa in an analysis may be the best way to ensure phylogenetic accuracy (although few studies to date have examined sequences long enough to provide complete resolution of complex trees). The optimal density of taxa on the tree and the necessary sequence length for complete resolution of the branching order will vary from problem to problem, so parametric bootstrapping techniques[9] should be useful for designing full-scale studies using preliminary data sets.

David M. Hillis
Department of Zoology,
University of Texas,
Austin, Texas 78712, USA
e-mail: hillis@bull.zo.utexas.edu

Mesoderm induction in zebrafish

SIR — Mesoderm formation and its dorsoventral specification are essential processes in vertebrate early development. The mesoderm produces many types of tissues such as notochord, muscle, bone and blood cells, arranged in the correct positions along the dorsoventral axis. The mechanisms underlying these processes are best understood in the amphibian *Xenopus*, whereas little is known about this process in other vertebrates. In amphibia, the mesoderm differentiates in the equatorial region of the blastula-stage embryo through inductive interactions with vegetal blastomeres; the dorsoventral polarity of the mesoderm also depends on signals from dorsovegetal blastomeres[1–3]. Are similar mechanisms involved in the development of teleost fish? Which tissues are responsible for mesoderm induction in the teleosts? We have addressed these questions using zebrafish embryos.

We focused on the yolk cell because the presumptive mesoderm region in the blastula-stage embryo is formed at the vegetal marginal zone of the blastoderm, close to the yolk cell. To examine the ability of the yolk cell to induce mesoderm, we have developed a transplantation method (Fig. 1). We directly transplanted biotin-labelled yolk cells, from which the blastoderm had been removed, onto the animal-pole region of unlabelled host embryos with their yolk syncytial layer facing the host animal pole. The yolk syncytial layer, which is unique to teleosts, covers the animal region of the yolk cell under the blastoderm. During incubation, the host blastomeres covered the surface of the grafted syncytial layer. We examined the recombinants at early gastrula stages with the mesodermal markers *no tail* (*ntl*, a pan-mesodermal marker)[4] and *goosecoid* (*gsc*, a dorsal mesodermal marker)[5].

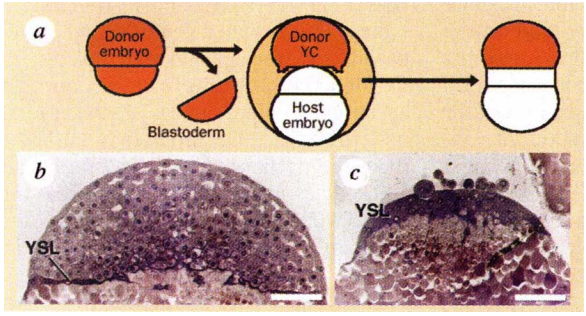


FIG. 1 Transplantation of the yolk cell (YC). *a*, Schematic representation of cell transplantation. Biotin-labelled donor cell, from which the blastoderm has been removed, was transplanted onto the animal-pole region of an unlabelled host embryo of the same developmental stage (mid-blastula stage). Recombinants were examined at 50%-epiboly with mesodermal markers by *in situ* hybridization. *b*, Light micrograph of a host embryo at the mid-blastula stage. In the marginal (MZ), yolk syncytial layer (YSL) and blastoderm cells are tightly attached to each other. *c*, Donor yolk cell. Under our experimental conditions, we could not remove the MZ cells completely with the YC intact, probably due to the tight adhesion of MZ cells with the YSL. However, those MZ cells, when transplanted into the host embryo, do not induce ectopic expression of *ntl* (data not shown). Therefore, we conclude that the few MZ cells attached to donor YCs do not affect the results obtained from our YC transplantation experiments. Scale bars, 0.1 mm.

We transplanted yolk cells prepared from mid-blastula into the host embryos of the equivalent developmental stages. As shown in Fig. 2*a*, ectopic expression of *ntl* is induced as a circumferential ring around the grafted yolk cell (all positive out of the 72 operations), similar to the normal expression pattern on the host vegetal side (Fig. 2*d*). Histological examination reveals that ectopic expression is induced in the unlabelled host cells close to the yolk

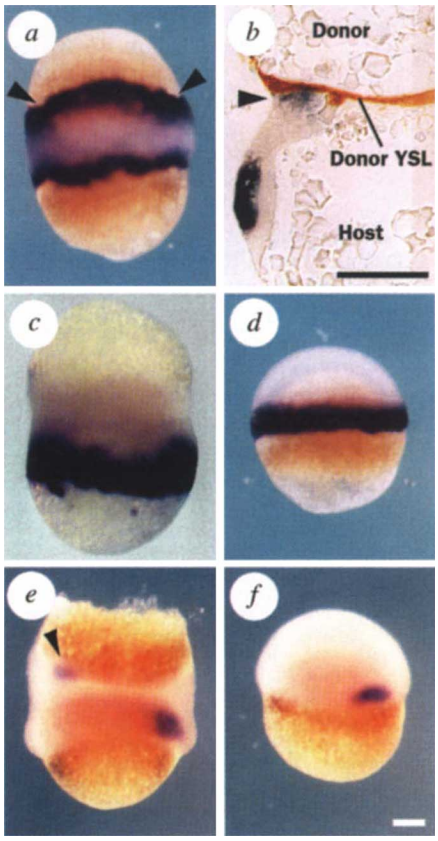


FIG. 2 Induction of *ntl* and *gsc* expression by the transplanted yolk cell. *a*, Expression of *ntl* in an embryo transplanted with a YC at the 512-cell-to-sphere stage and fixed at 50%-epiboly. Ectopic expression (indicated by arrows) is evident in a circumferential ring around the grafted YC. *b*, Histological section of a YC-transplanted embryo. Ectopic *ntl* expression (arrowhead) is seen in the host cells (unlabelled) near the donor YSL (stained brown after biotin–avidin peroxidase staining). *c*, No ectopic expression of *ntl* is observed when an unfertilized egg is transplanted instead of the YC. *d*, Expression of *ntl* in normal embryo. *e*, Expression of *gsc* in an embryo transplanted with a YC. Localized expression of *gsc* is induced near the transplanted YC (arrowhead). *f*, Normal expression pattern of *gsc* at 50%-epiboly. Scale bars, 0.1 mm.

syncytial layer of the transplanted yolk cell (Fig. 2*b*), suggesting that the syncytial layer is important in the induction of *ntl* expression. At all stages examined (512-cell embryo to 'sphere'), the yolk cell shows the same inducing activity. We also transplanted unfertilized eggs but found no induction of *ntl* (Fig. 2*c*). These results clearly show that the mesoderm-inducing signals are derived from the yolk cell and passed through the syncytial layer to the blastoderm cells.

We then examined the expression pattern of *gsc* in the recombinants. In addition to endogenous *gsc* expression, which is localized in the presumptive dorsal region, ectopic *gsc* expression is frequently induced near the transplanted yolk cell (60 out of 66 operations) (Fig. 2*e*, *f*). In most recombinants, ectopic expression is