# Life with 25,000 Genes

R. Scott Poethig

*Genome Res.* 2001 11: 313-316
Access the most recent version at doi:10.1101/gr



Commentary

# Life with 25,000 Genes

R. Scott Poethig

*Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA*

Plants make the earth a good place for humans to live. They produce the oxygen we breathe, the food we eat, fuel for our cars and factories, fiber for the clothes we wear, wood for the houses we live in, and chemicals that keep us healthy or help cure us when we get sick. Plus, they are pleasant to look at, fun to grow, and intellectually interesting. Given all of this, it is surprising that they remain so poorly understood.

The completion of the *Arabidopsis* genome sequence will do much to change this (*Arabidopsis* Genome Initiative 2000; Lin et al. 1999; Mayer et al. 1999; Salanoubat et al. 2000; Tabata et al. 2000; Theologis et al. 2000). As the first plant genome to be sequenced, this is rightly heralded as a landmark event. With the array of molecular genetic tools available for *Arabidopsis* and the impetus provided by the 2010 project, it will not be long before we know the physiological and developmental function of every gene in this species (Chory et al. 2000; Somerville and Dangl 2000). It is easy to take these events for granted in the era of genomic sequencing and reverse genetics. However, none of this would have happened without the major change in plant biology that has taken place in the last 15 years (Fink 1998). For example, in the late 1970's and early 1980's, students in Ian Sussex' plant biology laboratory completed Ph.D. theses on no less than 11 species, most of which were crop plants (my contribution was tobacco). At the time of Sussex's retirement in 1997, all but one of the students in his lab were studying *Arabidopsis* (I. Sussex, pers. comm.). The recent explosion of interest in *Arabidopsis*—a weed of absolutely no economic importance—is unprecedented in the history of plant biology and provided this field with its first widely adopted model system. *Arabidopsis* is a small plant with a rapid life cycle and has been used in genetic studies for decades, but it was the discovery that *Arabidopsis* has a small genome with very little repeated DNA (Leutwiler et al. 1984; Pruitt and Meyerowitz 1986) that is primarily responsible for its recent popularity. The widespread adoption of *Arabidopsis* as an experimental system has resulted in rapid progress in many areas and has produced the international effort that led to the sequencing of the *Arabidopsis* genome.

The *Arabidopsis* genome is 125 Mb and encodes ~25,500 genes (*Arabidopsis* Genome Initiative 2000). *Arabidopsis* therefore has significantly more genes than yeast (Goffeau et al. 1996), *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium 1998) or *Drosophila* (Adams et al. 2000). This is primarily because *Arabidopsis* genes often occur in more than one copy. Seventeen percent of the genes in *Arabidopsis* occur as tandem arrays of two or more closely related genes and ~60% of the genome is segmentally duplicated, albeit in a highly rearranged fashion; the number of unique types of genes in *Arabidopsis* (12,000) is actually about equal to the number of gene types in worms (14,000) and flies (11, 000). The types of genes present in *Arabidopsis* reinforce what has been learned from previous sequencing projects about the evolution of eukaryotes. Genes required for eukaryotic cell function such as components of the cytoskeleton, or essential processes such as DNA replication, repair and recombination, cell division, protein synthesis, and vesicle trafficking, are largely conserved between *Arabidopsis* and other eukaryotes. In contrast, genes involved in regulatory processes such as signal transduction and transcriptional regulation are quite different in *Arabidopsis*, yeast, *C. elegans*, and *Drosophila*. For example, *Arabidopsis* has no genes similar to the components of major signaling pathways in animals, such as the *Wingless/Wnt, Hedgehog, Notch/lin12, JAK/STAT, TGF-b*β*/SMADS*, and receptor tyrosine kinase/Ras pathways. Instead, signaling in *Arabidopsis* depends largely on receptor Ser/Thr kinases, of which there are 340, and a novel family of receptors related to bacterial two-component histidine kinases. Plants have also evolved a diverse array of transcription factors not found in animals. *Arabidopsis* has ~1500 transcription factors, 1.3 times as many as in Drosophila and 1.7× as many as *C. elegans* or yeast (Riechmann et al. 2000). Forty-five percent of the families of transcription factors found in *Arabidopsis* are unique to plants. These differences should not be surprising, given that multicellularity originated independently in plants and animals, and plants have cell walls and animals do not (Meyerowitz 1999). The important lesson is that plants are as different from other organisms as they are the same, and are interesting for both reasons.

The fact that *Arabidopsis* has more genes than either *C. elegans* or *Drosophila* begs the question: Why does a structurally and behaviorally simple organism have more genes than organisms with a nervous system and cells that move? Although the answer to this question is unknown, there are some obvious possibili-

**E-MAIL spoethig@sas.upenn.edu; FAX (215)898-8780.**
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.

ties. In contrast to animals, plants are autotrophs and can synthesize what they need to survive from air, light, water and a few mineral nutrients. In addition, they produce a huge variety of secondary compounds used for defense, disease resistance and a variety of other purposes; in fact, the number of secondary compounds produced by all plants is estimated to be as high as 100,000 (*Arabidopsis* Genome Initiative 2000). Many of the enzymes that participate in these metabolic pathways are present in multiple copies in the genome. The extent to which this genetic diversity leads to functional diversity is unclear, but the potential for enormous biochemical diversity is certainly present in the *Arabidopsis* genome. In short, animals are structurally more complex than plants, but plants probably do a lot more biochemistry. Another possible explanation for the difference in gene number between plants and animals is the remarkable ability of plants to deal with genetic imbalance. Variation in chromosome number resulting from either polyploidy or aneuploidy, as well as variation in gene dose due to segmental duplications and deficiencies, is tolerated much better by plants than by animals. Duplication and diversification of gene function may therefore be a more important source of novelty in plants than in animals. Of course, it is also possible that much of the genetic redundancy in *Arabidopsis* is functionally irrelevant. *Arabidopsis* may have more genes than either worms or flies because it can, not because it needs to.

While it is interesting to compare *Arabidopsis* to species from which it diverged 1.6 billion years ago (Wang et al. 1999), the value of the *Arabidopsis* genome sequence lies primarily in what it reveals about plants. One surprise is the extent of segmental duplication (*Arabidopsis* Genome Initiative 2000; Blanc et al. 2000; Vision et al. 2000). This observation confirms the results of interspecific comparisons of genome organization (Kowaleski et al. 1994; Paterson, 1996; Grant et al. 2000; Ku et al. 2000;) and suggests that the lineage leading to *Arabidopsis* underwent at least one genome duplication event, followed by extensive gene loss and rearrangement (Vision et al. 2000). Polyploidy is common in plants but was not predicted for a plant with five chromosomes and an unusually small genome. Estimates based on assumptions about the rate of amino acid substitution indicate that one large-scale duplication occurred ~100 Myr ago, and suggest that other duplications may have occurred 140, 170, and 200 Myr ago (Vision et al. 2000). The latter duplications are old enough to have occurred prior to the divergence of major angiosperm lineages (including the split between monocots and dicots), and should therefore be shared by many different flowering plants (Figure 1). Extensive genome rearrangements can occur within a few generations in newly polyploid plants (Song et al. 1995), so it is unclear how useful information about

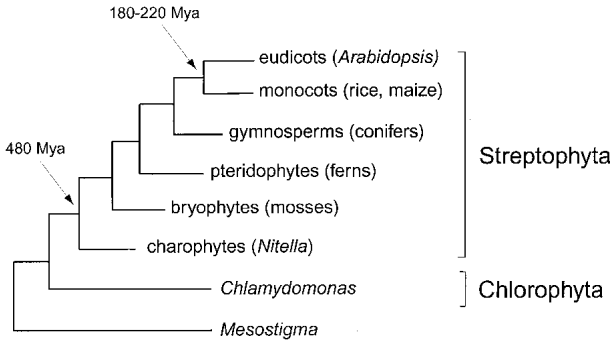


**Figure 1** An abbreviated phylogeny of green plants.

genome structure will be for tracing evolutionary relationships. Nevertheless, the structure of the *Arabidopsis* genome will undoubtedly lead to a renewed appreciation of the importance of polyploidy in plant evolution (Soltis and Soltis 2000).

Knowing the identity of all 25,000 or so genes in *Arabidopsis* now makes it possible to determine their function not only in *Arabidopsis* but in other plant species as well. This is a large task because of the number of genes involved and the potential for functional redundancy but—at least in *Arabidopsis*—is technically not very difficult. Loss-of-function mutations can be readily identified by screening for transgene insertions by PCR (Krysan et al. 1996), and large populations of T-DNA transformed lines have been created specifically for this purpose. A variety of methods also exist for ectopically expressing genes in a regulated or unregulated fashion. Microarrays have already proven their value for determining global patterns of gene expression in *Arabidopsis* (Harmer et al. 2000), and will be an important tool for characterizing the effect of loss-of-function and gain-of-function mutations. The goal of determining the function of every gene in *Arabidopsis* by 2010 (Chory et al. 2000; Somerville and Dangl 2000) is therefore well within reach.

Information about the types of genes present in *Arabidopsis* is of widespread interest because gene function is often well conserved among flowering plants. Nevertheless, because many plants are highly polyploid and have experienced varying degrees of gene amplification, it is not necessarily a straightforward matter to identify orthologous genes in different species. It is also impossible to predict whether regulatory sequences have been conserved along with coding sequences. Analyses of the MADS box gene family illustrate this point. For example, Kramer and Irish (1998) showed that the evolution of two members of this family, *PI* and *AP3*, was accompanied by several duplication events and involved significant changes in gene expression patterns as well (Kramer and Irish 1999). Whether these changes are associated with changes in gene function is unknown. This phenomenon will be

encountered over and over again as investigators try to move from *Arabidopsis* to other members of the plant kingdom and is a major challenge for future research.

The biggest obstacle to using *Arabidopsis* to identify genes in economically important plants is that most important crop plants in the world are monocots, and are therefore distantly related to *Arabidopsis*. In order of harvested weight, the top 10 crops in 1999 were: sugarcane, maize, rice, wheat, potatoes, cassava, soybeans, sweet potatoes, and barley (see http://apps.fao.org/). Of these, sugarcane, maize, rice, wheat, and barley are monocots and members of the grass family. Ready access to the genomes of these species will come with the sequencing of the rice genome, expected to be completed in 10 years if not sooner (http://rgp.dna.affrc.go.jp/rgp/News/Newsletter.html). In the meantime, gene identification in these species is facilitated by the growing number of maize and rice ESTs and by other gene identification strategies that are being explored courtesy of a significant increase in federal funding for plant genomics. The value of sequencing the maize genome is controversial. In contrast to rice, maize has a large genome in which small islands of unique sequence are surrounded by oceans of repetitive DNA (Bennetzen et al. 1998); consequently, there is relatively little to be gained by sequencing more than only maize genes and their flanking 5′ and 3′ regions. On the other hand, the value of having a complete record of all the genes in the maize genome is undeniable, and many different ways to accomplish this goal are currently being explored (see http://plantgenome.sdsc.edu/).

After *Arabidopsis*, rice, and perhaps maize, what next? Although there is considerable interest in sequencing other crop species, it would be fundamentally more interesting (and, in the long run, perhaps more useful) to determine the genomic sequence of a unicellular ancestor of higher plants. This would provide a reference point for comparing the genomic organization of divergent plant lineages and would reveal the genetic basis for the morphological changes that accompanied the evolution of multicellularity (Graham et al. 2000). Flowering plants are members of the Streptophyta, a group that consists of all land plants and their closest algal relatives, the charophytes (Figure 1). Recent phylogenetic analyses based on ribosomal (Melkonian et al. 1995), actin (Bhattacharya et al. 1998), and chloroplast (Lemieux et al. 2000) gene sequences suggest that the unicellular biflagellate *Mesostigma viridae* is the most likely representative of the group from which streptophytes evolved. Consistent with its basal position, *Mesostigma* possesses a single actin gene (*Arabidopsis* has at least 10; McDowell et al. 1996) and has a larger complement of chloroplast genes than any other green alga or land plant. Efforts are already underway to sequence the genome of the unicellular chlorophyte, *Chlamydomonas*. These evolutionary arguments suggest that the *Mesostigma* genome also has much to say about plant function and evolution, and should be given serious consideration.

With 25,000 mostly-uncharacterized genes, the *Arabidopsis* genome will keep plant biologists busy for a long time. Given the traditional interest in plant diversity, it will be interesting to see whether this information will be used primarily to explore the biology of *Arabidopsis*, or as a starting point for forays into the far reaches of the plant kingdom. One thing is clear: With a host of interesting problems to study, an unrivalled set of molecular genetic tools, and now a sequenced genome, for plant biologists the fun has just begun.

## ACKNOWLEDGMENTS

I am grateful to Maja Bucan, Tony Cashmore, and members of my laboratory for helpful comments on this manuscript, and to Linda Graham and Claude Lemieux for information about *Mesostigma viridae*.

## REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. *Science* **287:** 2185–2195.

*Arabidopsis* Genome Initiative 2000. *Nature* **408:** 796–815.

Bennetzen, J.L., San Miguel, P., Chen, M.S., Tikhonov, A., Francki, M., and Avramova, Z. 1998. *Proc. Natl. Acad. Sci.* **95:** 1975–1978.

Bhattacharya, D., Weber, K., An, S.S., and Berning-Koch, W. 1998. *J. Mol. Evol.* **47:** 544–550.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. 2000. *Plant Cell* **12:** 1093–1101.

*Caenorhabditis elegans* Sequencing Consortium 1998. *Science* **282:** 2012–2018.

Chory, J., Ecker, J.R., Briggs, S., Caboche, M., Coruzzi, G.M., Cook, D., Dangl, J., Grant, S., Guerinot, M.L., Henikoff, S., et al. 2000. *Plant Physiol.* **123:** 423–426.

Dean, C. and Schmidt, R. 1995. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **46:** 395–418.

Doebley, J. and Lukens, L. 1998. *Plant Cell* **10:** 1075–1082.

Fink, G. 1998. *Genetics* **149:** 473–477.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. *Science* **274:** 546–567.

Graham, L.E., Cook, M.E., and Busse, J.S. 2000. *Proc. Natl. Acad. Sci.* **97:** 4535–4540.

Grant, D., Cregan, P., and Shoemaker, R.C. 2000. *Proc. Natl. Acad. Sci.* **97:** 4168–4173.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A., and Kay, S.A. 2000. *Science* **290:** 2110–2113.

Kowaleski, S.P., Lan, T.H., Feldman, K.A., and Paterson, A.H. 1994. *Genetics* **138:** 499–510.

Kramer, E.M., Dorit, R.L. and Irish, V.F. 1998. *Genetics* **149:** 765–783.

Kramer, E.M. and Irish, V.F. 1999. *Nature* **399:** 144–148.

Krysan, P.J., Young, J.C., Tax, F., and Sussman, M. R. 1996. *Proc. Natl. Acad. Sci.* **93:** 8145–8150.

Ku, H.M., Vision, T., Liu, J., and Tanksley, S.D. 2000. *Proc. Natl. Acad. Sci.* **97:** 9121–9126.

Lemieux, C., Otis, C., and Turmel, M. 2000. *Nature* **403:** 649–652.

Leutwiler, L.S., Hough-Evans, B.R., and Meyerowitz, E.M. 1984. *Mol. Gen. Genet.* **194:** 15–23.

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M., et al. 1999. *Nature* **402:** 761–768.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G.,

Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K.D., Terryn, N., et al. 1999. *Nature* **402:** 769–777.
McDowell, J.M., Huang, S., McKinney, E.C., Huang, S., and Meagher, R.B. 1996. *Genetics* **142:** 587–602.
Melkonian, M., Marrin, B., and Surek, B. 1995. In *Biodiversity and Evolution*, (eds. R. Arai, M. Kato, and Y. Doi), pp. 153–176. Tokyo: The National Science Museum Foundation.
Meyerowitz, E.M. 1999. *Trends Biochem. Sci.* **24:** M65-M68.
Paterson, A.H. 1996. *Nat. Genet.* **14:** 380–382.
Pruitt, R.E. and Meyerowitz, E.M. 1986. *J. Mol. Biol.* **187:** 169–183.
Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., et al. 2000. *Science* **290:** 2105–2110.
Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blocker, H., Perez-Alonso, M., Obermaier, B., et al. 2000. *Nature* **408:** 820–822.
Soltis, P.S. and Soltis, D.E. 2000. *Proc. Natl. Acad. Sci.* **97:** 7051–7057.
Somerville, C. and Dangl, J. 2000. *Science* **290:** 2077–2078.
Song, K., Lu, P., Tang, K. and Osborn, T. 1995. *Proc. Natl. Acad. Sci.* **92:** 7719–7723.
Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., et al. 2000. *Nature* **408:** 823–826.
Theologis, A., Ecker, J.R., Palm, C.J., Federspiel, N.A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C.L., et al. 2000. *Nature* **408:** 816–820.
Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. *Science* **290:** 2114–2117.
Wang, D.Y.C., Kumar, S., and Hedges, S.B. 1999. *Proc. R. Soc. Lond. B.* **266:** 163–171.