How many genes are there in plants (... and why are they there)?

Lieven Sterck[1], Stephane Rombauts[1], Klaas Vandepoele[1], Pierre Rouzé[2] and Yves Van de Peer[1]

Annotation of the first few complete plant genomes has revealed that plants have many genes. For *Arabidopsis*, over 26 500 gene loci have been predicted, whereas for rice, the number adds up to 41 000. Recent analysis of the poplar genome suggests more than 45 000 genes, and partial sequence data from *Medicago* and *Lotus* also suggest that these plants contain more than 40 000 genes. Nevertheless, estimations suggest that ancestral angiosperms had no more than 12 000–14 000 genes. One explanation for the large increase in gene number during angiosperm evolution is gene duplication. It has been shown previously that the retention of duplicates following small- and large-scale duplication events in plants is substantial. Taking into account the function of genes that have been duplicated, we are now beginning to understand why many plant genes might have been retained, and how their retention might be linked to the typical lifestyle of plants.

Addresses
[1] Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[2] Laboratoire Associé de L'Institut National de la Recherche Agronomique, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Corresponding author: Van de Peer, Yves (yves.vandepeer@psb.ugent.be)

Current Opinion in Plant Biology 2007, **10**:199–203

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Stefan Jansson and Edward S Buckler

Available online 7th February 2007

1369-5266/$ – see front matter

DOI 10.1016/j.pbi.2007.01.004

Introduction

How to count the number of genes in eukaryotic genomes, much less the number of proteins that they encode, is not self-evident [1–4]. Nevertheless, in the past few years, much has been learned from major genome annotation projects, and estimates of gene number are generally becoming more realistic [5,6]. What has become clear from recent plant genome-sequencing projects is that plants seem to have lots of genes: studies often report more than 40 000 [7,8••,9•,10••]. There might be several reasons for this. A first explanation that has been put forward many times has to do with the typical lifestyle of a plant. Plants are sessile and cannot escape enemies or uncomfortable conditions. They are stuck in place and have therefore developed many strategies that improve their chances of survival when faced with grazing herbivores (including insects and snails), pathogens (viral, bacterial and fungal), varying climates, competing neighbour species, and other forms of stress. In addition, because they do not move, many plants have invented either efficient reproductive strategies that rely on external factors, such as wind and water, or ways to build colourful and scented flowers that attract pollen- and nectar-collecting animals to effect efficient mating and seed dispersal. In other words, plants must make tens of thousands of chemical compounds, which they use to ward off competition from other plants, to fight infections, and to respond generally to the environment.

A second reason why plants have so many genes might be gene duplication, or more precisely gene retention following gene duplication. Gene duplication and retention in plants has been extensive and gene families are generally larger in plants than in animals. Furthermore, most (if not all) plant species have experienced at least one (and probably more) whole-genome duplications in their evolutionary past [11,12•]. Many of the genes created through these major events have been retained in extant plant genomes [13••]. Here, we briefly discuss what is known about the number of genes in those plant species whose whole-genome sequences have been determined, and comment on possible reasons for the large number of genes in these genomes. When discussing gene numbers, we consider protein-coding gene loci rather than the number of transcripts a gene potentially encodes. Non-protein-coding genes are not discussed here, although it has been shown that many regions of the genome that were previously considered inactive or featureless might actually contain many sites of RNA activity [14,15].

How many genes are there in plants?

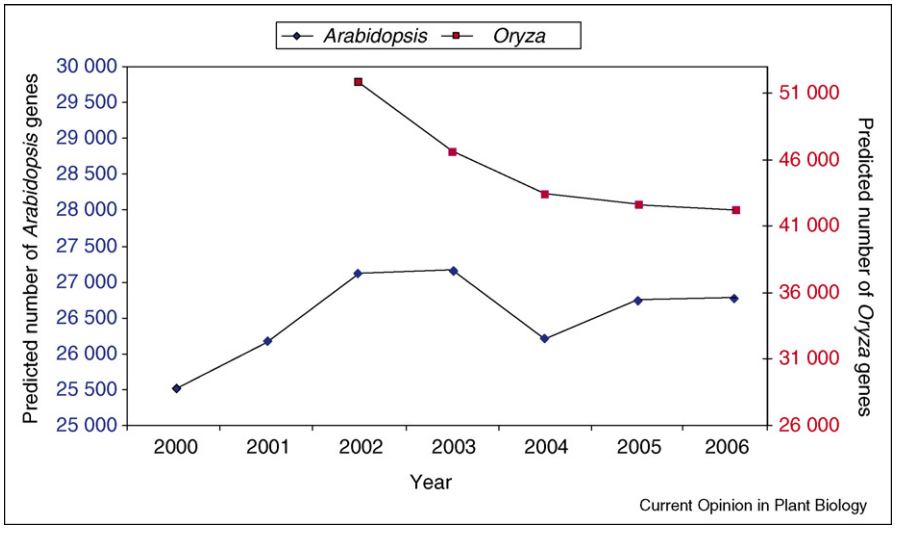
The caveats in gene prediction have been extensively discussed elsewhere and are not the subject of this paper. Suffice it to say that, although great progress has been made in the development of sophisticated gene finders and gene-prediction platforms (e.g. [16–20]), gene prediction and genome annotation are notoriously difficult [3–5]. Because the annotation community is well aware of this, gene models are continuously being re-evaluated on

the basis of novel data and feedback from experts, and software is being improved and retrained before the whole annotation pipeline is applied again [21•]. In addition, increasing numbers of transposable elements are being identified [5,22], and hence these can be masked in annotation processes. More and more expressed sequence tags (ESTs), full-length cDNAs and genomes are being sequenced, allowing the use of comparative approaches to reveal regions of sequence conservation and hence indicating the presence of genes that were missed by intrinsic prediction [23]. It is also important to realize that, for many genomes, particularly during the first stages of annotation, there are a considerable number of so-called 'hypothetical genes'. The structure of these genes has been predicted solely on the basis of intrinsic gene prediction, with no support from either expressed sequence matches or homologs of nucleic acids or proteins from other species. For instance, at the completion of the *Arabidopsis thaliana* genome in 2000, 5690 of the annotated genes were designated as 'hypothetical' [24,25], but with increased database content, improved annotation, and proof of expression (e.g. [25]), this number has decreased over time to just 732.

As a result of ongoing annotation efforts, predicted gene numbers continue to change (Figure 1). The genomes of *Arabidopsis* and rice, have both been available to the scientific community for a couple of years. When the first draft of the *Arabidopsis* genome sequence was published, about 25 500 protein-coding genes were reported [24]. In subsequent releases of the genome, this number increased to more than 27 000, mainly because of the availability of many more ESTs and improved gene prediction software. These new data and improved gene prediction revealed that some genes were fused in the initial modelling, while smaller genes were missed. Currently, 31 407 genes are documented in the TAIR6 (The Arabidopsis Information Resource6) release ([26]; http://www.Arabidopsis.org/), of which 26 751 are annotated as protein-coding genes, 3818 as pseudogenes, and 838 as non-coding RNA genes. For the rice genome, more than 50 000 genes were predicted upon publication of its draft sequence [27,28]. In further releases of its annotation, this number has dropped significantly to slightly more than 41 000 genes (http://www.tigr.org/). This decrease in the number of genes is mainly due to the discovery and removal of an increasing number of transposable elements [5]. Although all predictions of gene number are thus still subject to change, it is clear that *Arabidopsis* has a considerably smaller number of genes than rice (see also [29]). It would be premature to conclude, however, that dicots in general have smaller numbers of genes than monocots [8••]. For example, the number of genes in the poplar genome is estimated to be over 45 000 [10••], whereas partial sequence information from the *Medicago* and *Lotus* genome projects also suggests more than 40 000 genes [9•].

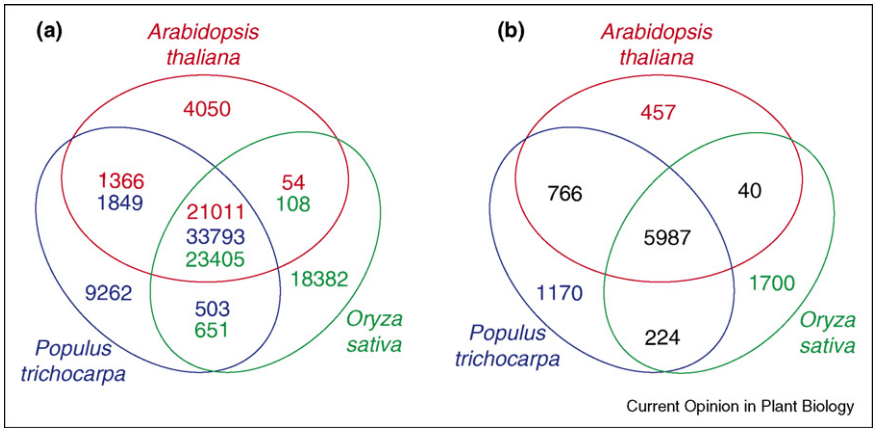
Figure 2 shows a Venn-diagram representing the number of shared and unique genes in the genomes of *Arabidopsis*, poplar, and rice. *Populus* has more protein-coding gene loci than *Arabidopsis*, with an average of 1.4–1.6 putative *Populus* homologs for each *Arabidopsis* gene. The relative frequency of protein domains in the two genomes is highly similar [10••], however, confirming that the greater gene number in poplar is largely due to the expansion of gene families. Rice also has many more genes than *Arabidopsis*, but unlike the situation in poplar, these are primarily genes for which no homolog exists in *Arabidopsis*, and might therefore be monocot specific ([29]; Figure 2). Although we only consider gene loci

Figure 1



Change in the predicted number of protein-coding gene loci for *Arabidopsis thaliana* (left axis) and *Oryza sativa* (right axis).

Figure 2



Venn-diagram representation of **(a)** genes that are shared or unique and **(b)** gene families for *Arabidopsis*, poplar, and rice. Gene families were constructed with MCLBLASTLINE (inflation factor: 2.2) based on BLASTP (E-value $< e^{-3}$) analyses. Genes are classified as unique when they do not belong to gene families that are shared with the other genomes.

here, it should be stressed that the proteome of plants might be much larger than the number of predicted protein-coding loci. Recent analyses suggested that more than 20% of plant genes might be alternatively spliced [25,30•], although it remains to be seen to what extent all of these splice variants are biologically functional.

Why are there so many genes in plants?

One of the most striking features of angiosperms is that many have experienced one or more episodes of polyploidy in their ancestry [12•,31]. Apart from species that are currently polyploid, which include most crops, others are considered to have paleopolyploid genomes. When the sequencing of the flowering plant *Arabidopsis* genome started, this model plant, with its small genome, was not expected to be an ancient polyploid. Five years after the release of its genome sequence [24], however, there is compelling evidence that the genome of *Arabidopsis*, or rather that of its ancestors, has been duplicated three times during the past 150–200 million years [13••,32,33].

Recently, we developed an evolutionary model that simulates the birth and death dynamics of genes on the basis of the age distribution of duplicated genes in the *Arabidopsis* genome [13••]. We took into account both a continuous mode of small-scale gene duplications and the three major genome-wide duplications that the *Arabidopsis* genome has undergone in its evolutionary past. When different functional classes of genes are considered, our study showed that gene families that are involved in transcriptional regulation, signal transduction, and development have all expanded considerably following the whole-genome duplications. Similar conclusions were reached by others who studied the retention of genes after genome duplication events [34,35]. However, few regulatory and developmental gene duplicates appear to have survived small-scale duplication events. This is in agreement with the 'gene balance' hypothesis, which states that the retention of genes that could have strong dosage effects, such as transcription factors, will be selected against if they are copied without their partners in a regulatory or protein-interaction network [13••,36]. On the other hand, if genes that encode products that cooperate in a complex pathway or network are duplicated together, as is the case in whole-genome duplications, gene dosage effects might be avoided by retaining all of the genes in that particular complex or network.

We also showed that genes that are involved in secondary metabolism or in responses to biotic stimuli, such as pathogen attack, tend to be preserved regardless of the mode of duplication [13••]. The finding that such genes have a good chance of retention following either small- or large-scale gene duplications probably reflects the continuous interaction between plants and animals, fungi, or plant pathogens, which imposes a constant need for adaptation. In other words, whatever the mechanism of duplication, novel genes that are important for fast adaptation to changing environments are often retained and quickly put to use by plants [37].

We can use the number of genes that have been retained following small- and large-scale duplication events to estimate the number of genes present in the ancestral angiosperm genome. Assuming a continuous rate of gene birth and three whole-genome duplications during the evolution of *Arabidopsis* and its predecessors, we estimate that the ancestor of the angiosperms had no more than 14 000 genes [13••]. A comparison of the *Arabidopsis* and poplar gene sets infers a similar number of around 12 000 genes in the conserved gene complement of their common ancestor

[10••]. So, if *Arabidopsis* and poplar both started out with about 12 000 or 14 000 genes and both experienced three genome duplications, two of which have been shared, how can poplar have more than 45 000 genes whereas *Arabidopsis* has only 26 500? This is particularly puzzling as it is believed that the youngest genome duplication in *Arabidopsis* occurred more recently (24–40 Mya, [38]) than that in poplar (60–65 Mya, [10••]). If the youngest genome duplication in *Arabidopsis* is more recent than that in poplar, we would expect *Arabidopsis* to have lost fewer genes (i.e. to have formed fewer pseudogenes) than poplar as gene decay is a function of time [13••,39]. One explanation for the smaller number of genes in *Arabidopsis* is that this species could have lost an unexpectedly large set of genes since its divergence from poplar. There are indeed some indications that this might be the case [8••,40]. For instance, there are 224 gene families that are present in poplar and rice but absent from *Arabidopsis* (Figure 2). About 30% percent of these gene families are also present in the partially determined *Medicago* genome (L Sterck *et al.*, unpublished). Alternatively, the greater number of genes in poplar might be explained by its lower rate of evolution. Because poplar is a long-lived vegetatively propagated species, it has the potential to contribute gametes to multiple generations. *Arabidopsis* plants have an annual lifespan, whereas a single *Populus* genotype can persist as a clone on the landscape for millennia. Recurrent contributions of ancient gametes from very old individual trees could potentially account for the markedly reduced rates of sequence evolution [10••] and thus also gene loss seen in *Populus*. This implies that many poplar genes might still be on the track to pseudogenisation.

Rice has probably also experienced several genome-wide duplications [41], although convincing evidence can only be found for the most recent one, which occurred after the split of monocots and dicots (120–150 Mya) but before the divergence of the grasses (50–70 Mya) [41,42]. There is, however, evidence of additional segmental duplications and massive ongoing individual gene duplications in rice [7], which are at least partly responsible for the large number of genes in rice.

Conclusions

When discussing genomes with fellow scientists, their first question is usually, 'How many genes'? The abstracts of papers that publish the first drafts of genome sequences also often mention the estimated number of genes. Our interest in the number of genes in a genome is probably a relic from the days when we were convinced that this number was correlated with the complexity of its host. In the meantime, we have learned better. The fact that man has only about twice the number of genes of the worm *Caenorhabditis elegans*, which in turn has more genes than the more complex fly *Drosophila*, was sobering in this respect. We have come to realize that the number of gene loci is far from being the sole contributor to genomic and biological complexity.

Nevertheless, the number of genes in plant genomes is very high. The sessile lifestyle of most plants could be part of the explanation, particularly if we link it to gene duplication, which has been rampant in plant genomes. Until we have a more complete picture that is based on the careful annotation of many more genomes, it will be hard to judge the extent to which gene number is related to the rate of evolution, the number of major duplication events, ecology, and plant biology in general.

Acknowledgements

We would like to thank all members of the Bioinformatics and Evolutionary Genomics group for stimulating discussions. KV is a postdoctoral fellow of the Fund for Scientific Research, Flanders.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Petsko GA: **Count me out**. *Genome Biol* 2000, **1**: comment1006.

2. Zhang MQ: **Computational prediction of eukaryotic protein-coding genes**. *Nat Rev Genet* 2002, **3**:698-710.

3. Mathé C, Sagot MF, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses**. *Nucleic Acids Res* 2002, **30**:4103-4117.

4. Roderic G, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E *et al.*: **EGASP: the human ENCODE genome annotation assessment project**. *Genome Biol* 2006, **7(Suppl I)**:S21-S31.

5. Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W: **Consistent over-estimation of gene number in complex plant genomes**. *Curr Opin Plant Biol* 2004, **7**:1-5.

6. Town CD: **Annotating the genome of *Medicago truncatula***. *Curr Opin Plant Biol* 2006, **9**:122-127.

7. Yu J, Wang J, Lin W, Li S, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J *et al.*: **The genomes of *Oryza sativa*: a history of duplications**. *PLoS Biol* 2005, **3**:e38.

8. •• Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA: **Differential methylation of genes and repeats in land plants**. *Genome Res* 2005, **15**:1431-1440.
The hypomethylated fraction of plant genomes is usually enriched in genes and can be selectively cloned using methylation filtration. The authors discuss the results of pilot methylation filtration sequencing projects in several plant species. They estimated the gene content of different plant genomes by analyzing their whole-genome shotgun sequence set. In general, they report that most plant species have large numbers of genes.

9. • Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T *et al.*: **Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes**. *Proc Natl Acad Sci USA* 2006, **103**:14959-14964.
A comparative study of the partial genome sequences of two legumes. Gene prediction and annotation of approximately half of the sequenced genomes suggest the existence of more than 40 000 genes in both species. There is also evidence that a genome duplication occurred before the divergence of *Medicago* and *Lotus*.

10. •• Tuskan G, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.*: **The**

genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray ex Brayshaw). *Science* 2006, **313**:1596-1604.
The authors publish the poplar genome sequence, the second one of a dicot plant. There is evidence for three whole-genome duplications, two of which were most probably shared with *Arabidopsis*, whereas one occurred more recently, after the divergence of *Arabidopsis* and poplar. Although both species have experienced three whole-genome duplications, poplar has retained many more protein-coding genes than *Arabidopsis*. On average, poplar has from 1.4 to 1.6 copies for every *Arabidopsis* gene, although the number of different protein domains is similar in each genome.

11. De Bodt S, Maere S, Van de Peer Y: **Genome duplication and the origin of angiosperms**. *Trends Ecol Evol* 2005, **20**:591-597.

12. • Ciu L, Wall PK, Heebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A *et al.*: **Widespread genome duplications throughout the history of flowering plants**. *Genome Res* 2006, **16**:738-749.
Genome duplications have been demonstrated for core eudicots, such as *Arabidopsis* and poplar, and for Poaceae, such as rice. However, this study uses EST data to show that genome duplications have also occurred in more basal angiosperm lineages. The authors also provide evidence for ancient polyploidy in the Gnetales, but there is no evidence for genome duplications in pine.

13. •• Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y: **Modeling gene and genome duplications in eukaryotes**. *Proc Natl Acad Sci USA* 2005, **102**:5454-5459.
This paper describes a mathematical model that simulates the birth and death dynamics of genes. By simulating age distributions of duplicated genes in the *Arabidopsis* genome, and then distinguishing between different classes of genes, the authors show that regulatory genes are mainly retained following whole-genome duplications. In contrast, there is selection against the retention of such genes when they are duplicated individually. Other genes, for instance those involved in responses to biotic stress or in secondary metabolism, have similar probabilities of being retained however they are duplicated.

14. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Beddnarek S: **Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays**. *Proc Natl Acad Sci USA* 2006, **102**:4453-4458.

15. Lu C, Tej SS, Luo S, Haudenschild CD, Meyer BC, Green PJ: **Elucidation of the small RNA component of the transcriptome**. *Science* 2005, **309**:1567-1569.

16. Schiex T, Moisan A, Rouzé P: **Eugene: a eukaryotic gene finder that combines several sources of evidence**. *Lect Notes Comput Sci* 2001, **2066**:111-125.

17. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R: **Comparative gene prediction in human and mouse**. *Genome Res* 2003, **13**:108-117.

18. Majoros WH, Pertea M, Slazberg SL: **TigrScan and GlimmerHMM: two open-source *ab initio* eukaryotic gene-finders**. *Bioinformatics* 2004, **20**:2878-2879.

19. Allen JE, Salzberg SL: **JIGSAW: integration of multiple sources of evidence for gene prediction**. *Bioinformatics* 2005, **21**:3596-3603.

20. Majoros WH, Pertea M, Delcher AL, Salzberg SL: **Efficient decoding algorithms for generalized hidden Markov model gene finders**. *BMC Bioinformatics* 2005, **6**:16.

21. • Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK Jr, Maiti R, Chan AP, Yu C, Farzad M, Wu D, White O, Town CD: **Complete reannotation of the *Arabidopsis genome*: methods, tools, protocols and the final release**. *BMC Biol* 2005, **3**:7.
The authors describe the efforts to re-annotate the *Arabidopsis* genome since its original publication. To further improve the annotations, new pipelines with an emphasis on utilizing full-length cDNAs and EST sequences have been set up to automate gene structure annotation.

22. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D: **Combined evidence annotation of transposable elements in genome sequences**. *PLoS Comp Biol* 2005, **1**:166-175.

23. Windsor AJ, Mitchell-Olds T: **Comparative genomics as a tool for gene discovery**. *Curr Opin Biotechnol* 2006, **17**:1-7.

24. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:796-815.

25. Xiao Y-L, Smith SR, Ishmael N, Redman JC, Kumar N, Monaghan EL, Ayele M, Haas BJ, Wu HC, Town CD: **Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2 reveals numerous transcript variants**. *Plant Physiol* 2005, **139**:1323-1337.

26. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M: **The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community**. *Nucleic Acids Res* 2003, **31**:224-228.

27. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)**. *Science* 2002, **296**:92-100.

28. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**:79-92.

29. Vandepoele K, Van de Peer Y: **Exploring the plant transcriptome through phylogenetic profiling**. *Plant Physiol* 2005, **137**:31-42.

30. • Wang B-B, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *Proc Natl Acad Sci USA* 2006, **103**:7175-7180.
In this study, the authors show that more than 20% of both *Arabidopsis* and rice genes with EST or cDNA evidence are alternatively spliced. Although not all splice variants will be biologically functional, alternative splicing undoubtedly plays an important role in adding to plant proteome diversity.

31. Tate JA, Soltis DE, Soltis PS: **Polyploidy in plants**. In *The Evolution of the Genome.* Edited by Gregory TR. Elsevier Academic Press; 2005:371-426.

32. Simillion C, Vandepoele K, Van Montagu M, Zabeau M, Van de Peer Y: **The hidden duplication past of *Arabidopsis thaliana***. *Proc Natl Acad Sci USA* 2002, **99**:13627-13632.

33. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**:433-438.

34. Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution**. *Plant Cell* 2004, **16**:1679-1691.

35. Seoighe C, Gehring C: **Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome**. *Trends Genet* 2004, **20**:461-464.

36. Freeling M, Thomas BC: **Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity**. *Genome Res* 2006, **16**:805-814.

37. Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y: **Nonrandom divergence of gene expression following gene and genome duplications in plants**. *Genome Biol* 2006, **7**:R13.

38. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome**. *Genome Res* 2003, **13**:137-144.

39. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151-1155.

40. Allen KD: **Assaying gene content in *Arabidopsis***. *Proc Natl Acad Sci USA* 2002, **99**:9568-9572.

41. Paterson AH, Bowers JE, Van de Peer Y, Vandepoele K: **Ancient duplication of cereal genomes**. *New Phytol* 2005, **165**:658-661.

42. Vandepoele K, Simillion C, Van de Peer Y: **Evidence that rice, and other cereals, are ancient aneuploids**. *Plant Cell* 2003, **15**:2192-2202.