the plant journal

SEB Society for Experimental Biology

The Plant Journal (2011) **66**, 58–65 doi: 10.1111/j.1365-313X.2011.04521.x

THE PLANT GENOME: AN EVOLUTIONARY VIEW ON STRUCTURE AND FUNCTION

Journey through the past: 150 million years of plant genome evolution

Sebastian Proost[1,2], Pedro Pattyn[1,2], Tom Gerats[3] and Yves Van de Peer[1,2,*]

[1]*Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium,*

[2]*Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium, and*

[3]*Radboud University, IWWR/Plant Genetics, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands*

Received 15 December 2010; revised 28 January 2011; accepted 31 January 2011.

[*]For correspondence (fax +32 9 3313807; e-mail yves.vandepeer@psb.vib-ugent.be).

SUMMARY

The genome sequence of the plant model organism *Arabidopsis thaliana* was presented in December of the year 2000. Since then, the 125 Mb sequence has revealed many of its evolutionary secrets. Through comparative analyses with other plant genomes, we know that the genome of *A. thaliana*, or better that of its ancestors, has undergone at least three whole genome duplications during the last 120 or so million years. The first duplication seems to have occurred at the dawn of dicot evolution, while the later duplications probably occurred <70 million years ago (Ma). One of those younger genome-wide duplications might be linked to the K-T extinction. Following these duplication events, the ancestral *A. thaliana* genome was hugely rearranged and gene copies have been massively lost. During the last 10 million years of its evolution, almost half of its genome was lost due to hundreds of thousands of small deletions. Here, we reconstruct plant genome evolution from the early angiosperm ancestor to the current *A. thaliana* genome, covering about 150 million years of evolution characterized by gene and genome duplications, genome rearrangements and genome reduction.

Keywords: Arabidopsis genome, genome evolution, comparative genomics, whole genome duplication.

SETTING THE STAGE

One determining factor that enabled the development of life on Earth was the realization of an information codex, documenting all instructions for creating and maintaining life in its most fundamental form. There are reasonable arguments to state that life originated through RNA, for one because of its combination of catalytic and information-storing abilities (Gilbert, 1986). Next, it is believed, the machinery to synthesize proteins was developed and this enhanced the catalytic properties and possibilities, not only for metabolic processes but also for refining information storage and replication. The chemically much more stable DNA was involved only later to maintain and transmit the genetic blueprints of all of life's structural and functional components. The genome thus has evolved over evolutionary time, perhaps from initially being the hard- and software package for self-replication, to the archive of the dazzling diversity of today's interacting and interfering millions of life forms. Parallel to that, where positive mutations early in evolution could often be a crude and easy hit, nowadays life forms and their interactions are so complex and mutually dependant that it becomes harder and harder for life to invent 'something totally new.'

Some of the major steps in the evolution from the early, prokaryotic life forms towards the complex eukaryotes have been the change from a circular to linear chromosomes, for which telomeres and centromeres had to be developed; the shift from prokaryotes (without membrane-enclosed organelles) to eukaryotes at around 2.7 Ga (Brocks *et al.*, 1999); the development of multicellularity during the so-called Cambrian explosion, 550 Ma, in which all major extant clades of multicellular animals appear, basically at the same moment. Land plants evolved a bit later, at around 400 Ma and showed a similarly explosive brief window of time for developing the angiosperms, around 140 Ma; their radiation was so fast and massive as to provoke Charles Darwin to call it an 'abominable mystery.' It is on the angiosperms and more in particular the dicots that we will mainly focus our journey through the past.

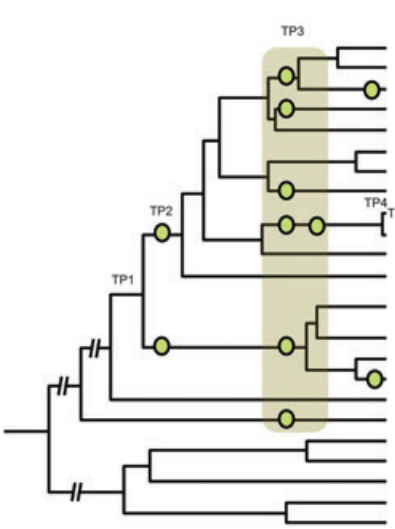
THE ANCESTRAL ANGIOSPERM GENOME: PLANT LIFE WITH 14 000 (OR FEWER) GENES (TP1)

For several decades, *A. thaliana* has been an excellent plant model organism for reasons well known (Koornneef and Meinke, 2010). Additionally, various techniques are available to genetically engineer *A. thaliana* (Bent, 2000). Furthermore, within the family of the Brassicaceae, many species are of major economical value. Important food crops include broccoli, cabbage (both *Brassica oleracea* ssp.) and mustard (*Brassica rapa/nigra*, *Sinapis alba*), while rapeseed (*Brassica napus*) is used to produce oils and more recently became a source for biodiesel. All this contributed to the popularity of *A. thaliana* in plant laboratories worldwide.

Last but not least, there is the small genome size of *A. thaliana.* With the recent advances in sequencing technologies, determining a genome sequence can be considered almost routine. A decade ago, however, genome sequencing was still a daunting, very expensive and laborious task and the size of the genome to be sequenced was a major determinant in whether or not a genome project was initiated. Coincidentally, with a size of about 125 Mb, the genome of *A. thaliana* was also one of the smallest plant genomes known and therefore an ideal target for sequencing (Arabidopsis Genome Initiative, 2000).

Analysis of the *A. thaliana* genome, and comparison with other plant genomes that have been determined subsequently, unveiled a very complex evolutionary history of the genome and that of its dicot ancestors. Although being a superb model system for plant geneticists, the genome of *A. thaliana* actually might be rather exceptional, with its many genome duplications, huge amount of gene losses, and recent genome shrinkage. Here, covering some 150 million years of angiosperm evolution, we discuss some milestones in the evolution of the *A. thaliana* genome and that of its ancestors, which eventually have led to the genome we know today.

In earlier studies, based on a mathematical model that simulates the birth and death of genes through small- and large-scale gene duplication events, we estimated that the ancestral angiosperm genome contained no more than 14 000 genes (Maere *et al.*, 2005). Although this was solely based on the analysis of the *A. thaliana* genome, similar values have been obtained through the comparison of different plant genomes. For instance, comparing the *A. thaliana* and poplar (*Populus trichocarpa*) gene sets suggested an ancestral gene count of 12 000 (Tuskan *et al.*, 2006), whereas clustering of homologous genes from *A. thaliana*, rice and 32 other plant species delineated approximately 12 400 ancestral genes (Vandepoele and Van de Peer, 2005). Recently, counting the number of genes that show cross-species synteny between the genomes of *A. thaliana*, grapevine (*Vitis vinifera*), papaya (*Carica papaya*) and poplar, suggested 10 000–13 000 ancestral angiosperm genes (Tang *et al.*, 2008b). In conclusion, it is probably safe to say that the ancestral angiosperm genome contained around



Species	#Chromosomes	Genome size (Mb)	#Protein coding genes	#Tandems	#Tandem clusters	#Tandems/#tandem clusters	#Protein coding genes (without tandems)	AVG intron length (bp)	#Introns per gene	AVG intergenic length (kb)
Lotus japonicus	6	560.0	43064	4092	1685	2.43	40657	352	1.64	14.5
Medicago truncatula	8	307.6	53423	19526	6860	2.85	40757	414	2.41	4.4
Glycine max	20	973.5	46382	9966	3677	2.71	40093	420	4.82	17.4
Malus domestica	17	742.3	63546	7305	2886	2.53	59127	405	3.76	8.3
Cucumis sativus	7	243.5	26682	-	-	-	-	483	3.39	-
Manihot esculenta	-	416.8	47443	13577	5362	2.53	39228	389	3.24	7.1
Ricinus communis	-	350.6	31221	5137	1963	2.62	28047	379	3.14	7.7
Populus trichocarpa	19	417.3	41377	10858	3929	2.76	34448	328	3.67	7.5
Arabidopsis thaliana	5	119.7	27291	6017	2204	2.73	23478	159	4.11	1.8
Arabidopsis lyrata	8	206.7	32670	7035	2634	2.67	28269	204	3.86	4.3
Carica papaya	-	331.4	27944	5306	2162	2.45	24800	498	3.05	10.2
Vitis vinifera	19	487.1	26419	7357	2412	3.05	21474	973	4.92	12.4
Oryza sativa ssp. japonica	12	372.9	42103	9370	3051	3.07	35784	406	3.00	3.9
Oryza sativa ssp. indica	12	427.7	49139	13362	4801	2.78	40578	399	3.32	4.2
Brachypodium distachyon	5	272.1	25532	4976	1893	2.63	22449	397	4.17	7.6
Sorghum bicolor	10	739.2	34525	8648	3020	2.86	28897	388	3.61	17.6
Zea mays	10	2061.7	30567	5249	2113	2.48	27431	509	3.64	65.4
Selaginella moellendorffii	-	212.8	34697	8997	3217	2.80	28917	102	4.67	4.4
Physcomitrella patens	-	480.2	28012	2520	1145	2.20	26637	276	4.24	11.0
Ostreococcus lucimarinus	21	13.2	7805	329	161	2.04	7637	161	0.30	0.5
Ostreococcus tauri	20	12.6	7933	485	226	2.15	7674	85	0.80	0.2
Micromonas sp. RCC299	17	21.1	10147	515	227	2.27	9859	170	0.54	0.5
Volvox carteri	-	137.8	15544	2399	969	2.48	14114	508	6.72	3.7
Chlamydomonas reinhardtii	17	112.5	16720	2577	1062	2.43	15205	335	6.24	3.1

Figure 1. Schematic and highly pruned phylogenetic tree of green algae and land plants for which the genome sequence has been determined. The background colour of cells indicates whether values are small (red), intermediate (yellow) or high (green) compared to the average value in the same column. Dots on the tree denote whole genome duplications. TPx denote specific time points discussed in the text. Raw data are derived from PLAZA 2.0 (Proost *et al.*, 2009). AVG, Average.

12 000–14 000 genes. Gene counts in extant angiosperm genomes are all considerably larger [see Figure 1; data derived from PLAZA 2.0 (Proost *et al.*, 2009)], due to the continuous process of gene duplication (Lynch and Conery, 2000) and, in numerous cases, genome duplications (see further).

Although the moss *Physcomitrella patens* seems to contain a number of genes that is comparable to that of many angiosperms, probably also due to a genome duplication event, the gene content is considerably different (Rensing *et al.*, 2008). Unicellular green algae on the other hand contain much fewer genes, as might be expected from their much simpler morphology, lifestyle, and ecology. *Volvox carteri* (Prochnik *et al.*, 2010) and *Chlamydomonas reinhardtii* (Merchant *et al.*, 2007) contain more than 15 000 and 16 000 genes, respectively, while the picoeukaryotic algae *Micromonas* and *Ostreococcus* contain about 10 000 and 8000 genes, respectively. It is interesting to note that the difference in gene count between the prasinophytes *Ostreococcus* sp. (Palenik *et al.*, 2007) and *Micromonas* sp. (Worden *et al.*, 2009) and the Chlorophyceae *Volvox carteri* and *Chlamydomonas reinhardtii* seems to be mainly due to duplicated genes present in the latter two species, but generally missing in the former ones (Figure 1).

THE HEXAPLOID ANCESTOR OF EUDICOT PLANTS: FROM 7 TO 21 CHROMOSOMES? (TP2)

Early analysis of the *A. thaliana* genome unveiled several rounds of Whole Genome Duplications (WGDs), although the exact number and timing has been disputed (Vision *et al.*, 2000; Simillion *et al.*, 2002; Blanc *et al.*, 2003; Bowers *et al.*, 2003). For instance, it was initially suggested that one of the WGDs detected in *A. thaliana* occurred before the radiation of most eudicots, and that the oldest WGD predated the divergence of dicots and monocots (Simillion *et al.*, 2002; Bowers *et al.*, 2003). By comparison with additional whole plant genomes however, a more complete picture has emerged. In particular the genomes of grapevine and papaya revealed conclusive evidence regarding the exact number and timing of WGDs that occurred early in the history of angiosperms (Jaillon *et al.*, 2007; Ming *et al.*, 2008). Grapevine is an early-diverging rosid and regions in the grapevine genome typically show homology with two other regions elsewhere in the same genome. Because of this triplicate genome structure, it was concluded that, most likely, three ancestral genomes had contributed to the grapevine lineage (Jaillon *et al.*, 2007). The recently released papaya genome shows a similar triplicate genome structure (Ming *et al.*, 2008), although papaya is not closely related to grapevine. Instead, it belongs to the order Brassicales and is more closely related to *A. thaliana* from which it diverged approximately 70 Ma (Wikström *et al.*, 2001; Ming *et al.*, 2008). Therefore, the most plausible and parsimonious explanation would be that the triplicate genome structure is ancient and shared between many, if not all eudicots. This is further supported by analysis of partial genome data of the asterid *Coffea* (Cenci *et al.*, 2010) and EST data of several other Asteraceae (Barker *et al.*, 2008), as well as by the recent completion of two additional rosid genomes, soybean (*Glycine max*; Schmutz *et al.*, 2010) and apple (*Malus domestica*; Velasco *et al.*, 2010). By comparing the pattern of gene losses in homeologous segments in papaya and grapevine, it was observed that two of three were more fractionated, suggesting that a first duplication event generated a tetraploid, which then hybridized with a diploid to generate a triploid. This triploid then underwent yet another whole genome duplication event to generate a hexaploid, giving rise to the triplicate genome structure we still find in species such as grapevine and papaya (Lyons *et al.*, 2008). Uncovering the triplicate genome structure in other plant genomes is more difficult because of additional WGD events that have occurred in several of these lineages (Van de Peer *et al.*, 2009a).

The extant grapevine genome consists of 19 chromosomes, most of which are clearly syntenic to two other chromosomes, hence the triplicate genome structure. Furthermore, two chromosomes show synteny to two different chromosomes, indicating chromosome fusions (Jaillon *et al.*, 2007). This particular structure would suggest that, about 120 Ma, the ancestral pre-hexaploid genome from which all dicots have evolved, consisted of seven chromosomes. This would also suggest that, subsequent to the hexaploidy event, the ancestral post-hexaploid genome would have consisted of 21 chromosomes (Jaillon *et al.*, 2007; Abrouk *et al.*, 2010). Possibly, amongst the ones available at this moment, the grapevine genome is the genome that still resembles that ancestral chromosomal state most, due to its slow rate of evolution (Jaillon *et al.*, 2007).

There is some evidence, albeit mostly circumstantial, that these early duplications can be linked to the origin and fast diversification of angiosperms (De Bodt *et al.*, 2005; Soltis *et al.*, 2008; Van de Peer *et al.*, 2009b). Gene and genome duplications potentially facilitate reproductive isolation (Lynch and Conery, 2000; Scannell *et al.*, 2006; Semon and Wolfe, 2007; Bikard *et al.*, 2009) and increase the diversifying potential of species thereby providing putative selective advantages over their diploid progenitors (Osborn *et al.*, 2003; Rieseberg *et al.*, 2003, 2007; Comai, 2005). Although their exact timing is uncertain, the hexaploidization event early in the evolution of flowering plants might have facilitated the emergence of flowers and specialized pollination strategies (Stuessy, 2004). This in turn might have been one of the crucial factors in the rapid diversification and speciation of flowering plants in the Early Cretaceous (Crepet, 2000; De Bodt *et al.*, 2005; Soltis *et al.*, 2008, 2009) and, if true, make the abominable mystery Darwin referred to somewhat less of a mystery.

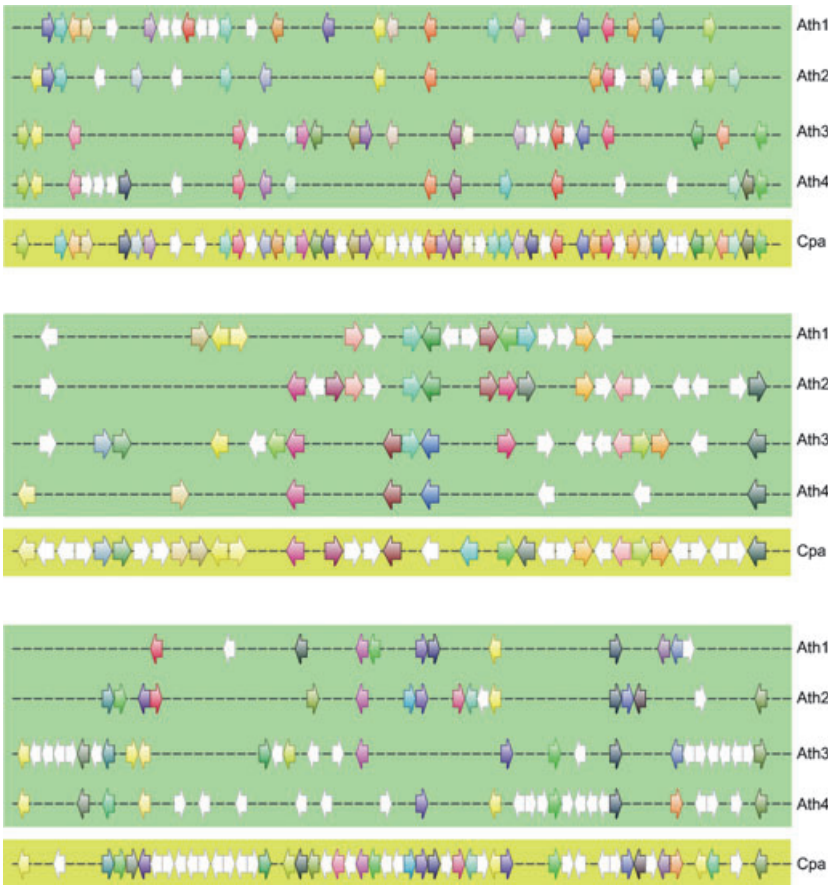


Figure 2. Collinearity between papaya and duplicated regions in *Arabidopsis thaliana*. In general, one region in papaya corresponds with four homologous regions in *A. thaliana*, providing strong evidence for two WGDs in Arabidopsis since its divergence from papaya, approximately 70 Ma. Ath: *Arabidopsis thaliana*; Cpa: *Carica papaya*.

TWO MORE GENOME DUPLICATIONS FOR ARABIDOPSIS (TP3)

Apart from the hexaploidy shared by most eudicots, many plant lineages show traces of additional, independent and more recent genome duplications (Blanc and Wolfe, 2004; Schlueter *et al.*, 2004; Cui *et al.*, 2006; Barker *et al.*, 2008, 2009; Lescot *et al.*, 2008). Interestingly, many independent WGDs, such as those in the cereals, the legumes, the Solanaceae, the Compositae, cotton (*Gossypium hirsutum*), poplar, banana (*Musa* sp.), and apple (*Malus domestica*) appear to have occurred somewhere between 50 and 70 Ma (Paterson *et al.*, 2004; Tuskan *et al.*, 2006; Lescot *et al.*, 2008; Fawcett *et al.*, 2009). Recently, it has been suggested that these duplication events might have coincided with the Cretaceous-Tertiary (K-T) extinction, the most recent large-scale mass extinction that wiped out around 80% of plant and animal species, including the dinosaurs (Fawcett *et al.*, 2009; Van de Peer *et al.*, 2009b).

Also the ancestors of *A. thaliana* seem to have undergone two additional genome duplications. Again, this has been uncovered through comparison with a close(r) relative, namely papaya, which has not shared these genome duplications. Figure 2 shows several sets of homologous regions in the genomes of papaya and *A. thaliana.* As can be observed, the one genome copy in papaya corresponds with four copies in *A. thaliana*, providing convincing support for two genome duplications in the lineage leading to *A. thaliana* since their divergence from papaya, about 70 Ma (Wikström *et al.*, 2001; Ming *et al.*, 2008; Tang *et al.*, 2008a,b). These findings were unexpected as other methods, relying on fossil evidence and phylogenetic trees to calibrate molecular clocks, placed both duplications considerably earlier. However, fossils for the Brassicales are rare and therefore few reliable age constraints could be used. Only recently, more advanced methods have been developed that can account for uncertainties in tree topology and allow evolutionary rates to be uncorrelated across the tree (Beilstein *et al.*, 2010). Recent age estimates now also place one WGD very close to the divergence from papaya and the most recent WGD within a window of 23–43 Ma (Barker *et al.*, 2009; Fawcett *et al.*, 2009).

From Figure 2, it also becomes clear why inferring the number of WGDs proved difficult using only the *A. thaliana* genome. Homologous segments in *A. thaliana* are often highly degenerated due to extensive gene loss. Indeed, as previously noted, high frequencies of gene loss [or gene fractionation *sensu* (Freeling *et al.*, 2008)] reduce collinearity resulting in duplicated regions that share very few, if any, homologous genes (Vandepoele *et al.*, 2002). Nevertheless,

by comparing chromosomal segments across multiple genomes, and in particular with genomes that have not shared the duplication event(s), such highly degenerated regions can often still be unveiled to be homologous (Van de Peer, 2004; Lyons *et al.*, 2008; Tang *et al.*, 2008a).

Using the papaya genome, it also became possible to estimate how much gene translocation has occurred in *A. thaliana*, since their divergence. Starting from collinear regions between both species, the chromosomal positions of *A. thaliana* genes were scored based on the conservation of homologous neighboring genes in papaya (Freeling *et al.*, 2008). Although the frequency of translocation varied among different gene families and functional categories, Freeling *et al.* estimated that about 25% of all *A. thaliana* genes had translocated since the origin of the Brassicales. Therefore, both massive gene loss and gene translocations seem to be responsible for the highly degenerated patterns of collinearity observed in intra-genome *A. thaliana* comparisons (Figure 2).

Previously, we estimated that the number of genes created by the hexaploidy event and surviving until today amounted to about 800. Furthermore, we estimated that the number of genes that have survived both of the more recent WGDs in *A. thaliana* is about 6700. The number of genes created through continuous small-scale duplications since the eudicot ancestor has been estimated to be about 5300 (Maere *et al.*, 2005, Table S3). When we add these numbers to the 14 000 genes assumed to have been present in the ancestor of the angiosperms (see above), we obtain a number (26 800) that is very close to the actual number of genes that currently has been annotated for the *A. thaliana* genome (Figure 1). It should be noted though that these values will be different for different plant species, dependent on the rate genes get duplicated and lost again (see further).

RECENT GENOME REDUCTION IN *ARABIDOPSIS THALIANA* (TPS 4 AND 5)

As far as we know, *A. thaliana* has remnants of more rounds of WGDs than any other dicot, maybe with the exception of soybean. Soybean has also undergone, apart from a legume-specific WGD that it shares with *Medicago* and *Lotus*, an additional genome duplication, about 13 Ma. Nevertheless, the current gene number of *A. thaliana* is comparable to the pre-13 Ma-duplication proto-soybean gene number, which was estimated at about 30 000 (Schmutz *et al.*, 2010). Because the most recent WGD in soybean has added another 16 000 genes to its gene complement, totalling the number at 46 000, the number of genes in *A. thaliana* (27 291), for a comparable number of WGDs, is very small and comparable to grapevine, papaya, and cucumber (Huang *et al.*, 2009), none of which have undergone additional genome duplications (see Figure 1). Apparently, *A. thaliana* has lost a larger set of genes during its evolution. This might, at least partly, be ascribed to its relatively high nucleotide substitution rate, compared with other species (Tang *et al.*, 2008b; Fawcett *et al.*, 2009). It has been shown that substitution rates are indeed dependent on generation time, population size, and metabolic rate, and can differ substantially between species (Lartillot and Poujol, 2011). A high substitution rate means that genes following duplication accumulate more mutations, which in turn may speed up pseudogenization. By contrast, the large number of genes found in species such as poplar and apple (Figure 1) could be explained by the combination of relatively recent WGDs and a slow(er) rate of evolution.

However, the large amount of gene loss in the *A. thaliana* lineage might not only be due to the loss of more gene duplicates accumulating deleterious mutations. By identifying pre- and post speciation paralogs between *A. thaliana* and its close relative *A. lyrata*, the genome sequence of which became available recently (Hu *et al.*, 2011), gene birth and loss events can be inferred, which allows to enumerate the contribution of these processes to the difference in gene count between both species [see also (Shiu *et al.*, 2004)]. This way, the last common ancestor of *A. thaliana* and *A. lyrata* was estimated to harbor about 30 500 genes. This value should be treated as a lower bound since genes that have been deleted in both species cannot be identified by the approach used. Based on phylogenies of pre- and post speciation paralogs, we estimated that, since their divergence, *A. thaliana* lost about 5700 genes compared with only about 3100 in *A. lyrata.* On the other hand, approximately 700 and 1800 new genes were created through post speciation duplications in *A. thaliana* and *A. lyrata*, respectively. Relative to the ancestral gene number, *A. thaliana* thus experienced a net loss of about 5000 genes compared with about 1300 for *A. lyrata.* It can therefore be concluded that gene deletion in *A. thaliana*, rather than enhanced duplication in *A. lyrata*, is responsible for the smaller gene set in *A. thaliana.*

Besides the number of genes, as mentioned previously, also the genome size of *A. thaliana* is small. For instance, *A. lyrata*, believed to have diverged from *A. thaliana* about 10–13 Ma (Beilstein *et al.*, 2010; Hu *et al.*, 2011), has a genome size of about 207 Mb. This means that, in a little more than 10 million years time, *A. thaliana* apparently lost about half of its genome. Various mechanisms appear to have caused this strong reduction in genome size. Part of this genome shrinkage can be explained by the elimination of three centromeres and six telomeres during chromosome fusion events that have lead to the current chromosome number of five in *A. thaliana* from the ancestral karyotype with eight chromosomes (Lysak *et al.*, 2006; Schranz *et al.*, 2006). Collinearity between both Arabidopsis species reveals this process, apart from additional reciprocal translocations, chromosome 1 in *A. thaliana* is a fusion of chromosome 1 and 2 of *A. lyrata*, chromosome 2 resulted from a fusion between chromosomes 3 and 4, and large

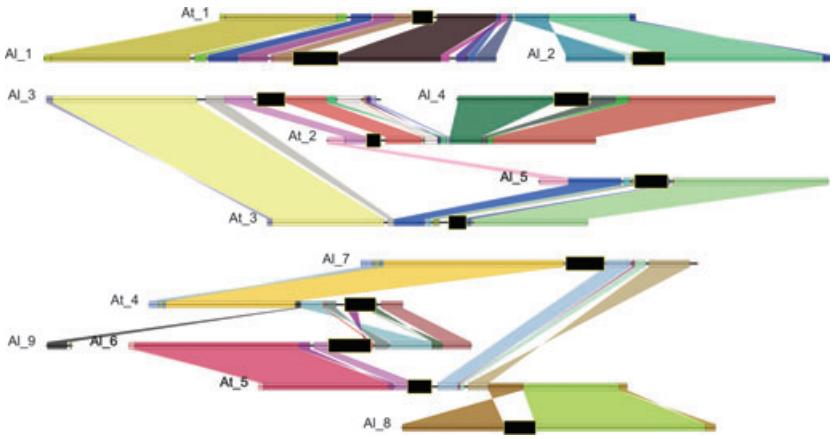


Figure 3. Collinearity between the five *A. thaliana* (At_1–5) and the eight *A. lyrata* chromosomes (Al_1–8). An additional *A. lyrata* scaffold nine shows significant collinearity but could not be assigned to one of the eight chromosomes due to a lack of markers. Colinear regions have the same colour and are connected through coloured areas. Inversions are represented by sandglass-like shapes. Centromeres are indicated by black boxes.

portions of chromosome 5 correspond to chromosome 6 and 8 (Figure 3). Therefore, the difference in chromosome number between both species can be explained by three chromosome fusions while the reduction of the *A. thaliana* genome size is explained at least in part by the loss of three centromeres and three pairs of telomeres. A similar reduction in chromosome number is also observed in other Brassicaceae, such as *Pachycladon*, *Stenopetalum nutans/lineare* and *Ballantinia antipoda.* Though in these studies, due to the absence of a full genome sequence, comparative chromosome painting was used to detect homeologous regions (Mandakova *et al.*, 2010a,b).

So, although genome reduction in *A. thaliana* can be partially attributed to the loss of DNA from large-scale rearrangements, the main cause lies in hundreds of thousands of small deletions found throughout the genome (Hu *et al.*, 2011). Although these microdeletions occurred primarily in non-coding DNA and transposons, many are also present in protein-coding multi-gene families, probably explaining the smaller number of protein-coding genes present in the *A. thaliana* genome, as discussed before. Using the *A. lyrata* genome to determine the derived state among a set of insertion and deletion polymorphisms found throughout the genome of 95 *A. thaliana* ecotypes (Nordborg *et al.*, 2005), Hu *et al.* (2011) uncovered more than 2600 fixed and more than 850 segregating deletions, compared with almost 1900 fixed and almost 100 segregating insertions, a clear excess of deletions over insertions. Furthermore, deletions are on average longer than insertions. If no selection were involved, and if this pattern were only due to mutational bias favoring deletions, deletion and insertion polymorphisms should have similar allele frequencies in the *A. thaliana* ecotypes. However, segregating insertions are, on average, found in fewer individuals than are deletions or single-nucleotide polymorphisms. Deletions are often found in the majority of individuals, and many are approaching fixation in *A. thaliana*, which seems to suggest that deletions are favored over insertions because of selection, rather than simple mutational bias, thus leading to a smaller genome.

So far, we have mainly focused on protein coding genes when discussing plant genome evolution. However, there are two main types of non-coding DNA, namely introns and transposable elements (TEs), which make a huge contribution to the size and structure of angiosperm genomes. The functions of both introns and TEs are still uncertain and vividly discussed. There is roughly a 1000-fold difference in genome size in flowering plants, mainly due to differences in copy number of a whole range of TEs. The proliferation process keeps going on: several species have doubled their genome over the past 5 million years (Ramsey and Schemske, 1998; Otto and Whitton, 2000; Soltis and Soltis, 2009). Up to around 90% of the elements usually are class I TEs, retrotransposons that transpose via an RNA intermediate. Class II elements transpose via a cut-and-paste mechanism and are much less abundant (Wicker *et al.*, 2007). In order to better understand the evolution of genome size in relation to its structure and function, the dynamics of TE behavior will have to be explored much deeper (Bennetzen, 2005). Changes in the balance between insertion and excision rates may be occurring under the influence of natural selection and cause changes in the distribution of the various elements.

CONCLUSION

In roughly 10 years time, we have developed a surprisingly detailed view of the evolutionary history of the *A. thaliana* genome and that of its ancestors. This only could have been done through the comparison of genomes of related plant species. It is to be expected that the current wave of emerging technologies will further improve our understanding of genome sequences and how they evolve over time. For instance, with the price per megabase significantly lower than 10 years ago, many re-sequencing studies have now become feasible. The first 500 *A. thaliana* genome sequences as part of the 1001 Arabidopsis genome initiative have already been determined and will be released soon

(D. Weigel, personal communication). Studying these additional genomes will undoubtedly further provide invaluable information on variations in the genome between very closely related species and will give a detailed picture on how genomes evolve at very small timescales. Resequencing of Arabidopsis genomes already provides lots of information on genome evolution at the nucleotide level, because single nucleotide polymorphisms and small insertions/deletions can be identified, as well as copy number variations (Clark *et al.*, 2007; Nordborg and Weigel, 2008; DeBolt, 2010). Sequencing closely related genomes can pinpoint specific genomic traits associated with an organism's lifestyle as well as provide insights in the genomic background of speciation. However, not only resequencing of known genomes will be done. Additionally, many genomes of various new plant families will be sequenced in the near future. Undoubtedly, these will reveal invaluable insights into how plant genomes have evolved through time.

ACKNOWLEDGEMENTS

S.P. and P.P. are indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). This work is supported by the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

REFERENCES

Abrouk, M., Murat, F., Pont, C. *et al.* (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J. and Rieseberg, L.H. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455.

Barker, M.S., Vogel, H. and Schranz, M.E. (2009) Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome Biol. Evol.* **1**, 391–399.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **107**, 18724–18728.

Bennetzen, J.L. (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* **15**, 621–627.

Bent, A.F. (2000) Arabidopsis *in planta* transformation. Uses, mechanisms, and prospects for transformation of other species. *Plant Physiol.* **124**, 1540–1547.

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M.J. and Loudet, O. (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, **323**, 623–626.

Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667–1678.

Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* **13**, 137–144.

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.

Brocks, J.J., Logan, G.A., Buick, R. and Summons, R.E. (1999) Archean molecular fossils and the early rise of eukaryotes. *Science*, **285**, 1033–1036.

Cenci, A., Combes, M.C. and Lashermes, P. (2010) Comparative sequence analyses indicate that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol. Genet. Genomics*, **283**, 493–501.

Clark, R.M., Schweikert, G., Toomajian, C. *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, **317**, 338–342.

Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846.

Crepet, W.L. (2000) Progress in understanding angiosperm history, success, and relationships: Darwin's abominably 'perplexing phenomenon'. *Proc. Natl Acad. Sci. USA*, **97**, 12939–12941.

Cui, L., Wall, P.K., Leebens-Mack, J.H. *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749.

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

DeBolt, S. (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol. Evol.* **2**, 441–453.

Fawcett, J.A., Maere, S. and Van de Peer, Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA*, **106**, 5737–5742.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R. and Lisch, D. (2008) Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res.* **18**, 1924–1937.

Gilbert, W. (1986) Origin of life: the RNA world. *Nature*, **319**, 618.

Hu, T.T., Pattyn, P., Bakker, E.G. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* (in press).

Huang, S., Li, R., Zhang, Z. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281.

Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Koornneef, M. and Meinke, D. (2010) The development of Arabidopsis as a model plant. *Plant J.* **61**, 909–921.

Lartillot, N. and Poujol, R. (2010) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* **28**, 729–744.

Lescot, M., Piffanelli, P., Ciampi, A.Y. *et al.* (2008) Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. *BMC Genomics*, **9**, 58.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Lyons, E., Pedersen, B., Kane, J. and Freeling, M. (2008) The value of non-model genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* **1**, 181–190.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc. Natl Acad. Sci. USA*, **103**, 5224–5229.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA*, **102**, 5454–5459.

Mandakova, T., Heenan, P.B. and Lysak, M.A. (2010a) Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evol. Biol.* **10**, 367.

Mandakova, T., Joly, S., Krzywinski, M., Mummenhoff, K. and Lysak, M.A. (2010b) Fast diploidization in close mesopolyploid relatives of Arabidopsis. *Plant Cell*, **22**, 2277–2290.

Merchant, S.S., Prochnik, S.E., Vallon, O. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.

Ming, R., Hou, S., Feng, Y. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.

Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature*, **456**, 720–723.

Nordborg, M., Hu, T.T., Ishino, Y. *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196.

Osborn, T.C., Pires, J.C., Birchler, J.A. *et al.* (2003) Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* **19**, 141–147.

Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437.

Palenik, B., Grimwood, J., Aerts, A. *et al.* (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA*, **104**, 7705–7710.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA*, **101**, 9903–9908.

Prochnik, S.E., Umen, J., Nedelcu, A.M. *et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science*, **329**, 223–226.

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y. and Vandepoele, K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.

Ramsey, J. and Schemske, D.W. (1998) Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* **29**, 467–501.

Rensing, S.A., Lang, D., Zimmer, A.D. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A. and Lexer, C. (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Rieseberg, L.H., Kim, S.C., Randell, R.A., Whitney, K.D., Gross, B.L., Lexer, C. and Clay, K. (2007) Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, **129**, 149–165.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. and Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341–345.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, **47**, 868–876.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schranz, M.E., Lysak, M.A. and Mitchell-Olds, T. (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542.

Semon, M. and Wolfe, K.H. (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* **23**, 108–112.

Shiu, S.H., Karlowski, W.M., Pan, R., Tzeng, Y.H., Mayer, K.F. and Li, W.H. (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell*, **16**, 1220–1234.

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

Soltis, P.S. and Soltis, D.E. (2009) The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* **60**, 561–588.

Soltis, D.E., Bell, C.D., Kim, S. and Soltis, P.S. (2008) Origin and early evolution of angiosperms. *Ann. NY Acad. Sci.* **1133**, 3–25.

Soltis, D.E., Albert, V.A., Leebens-Mack, J. *et al.* (2009) Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348.

Stuessy, T.F. (2004) A transitional-combinatorial theory for the origin of angiosperms. *Taxon*, **53**, 3–16.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008a) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M. and Paterson, A.H. (2008b) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954.

Tuskan, G.A., Difazio, S., Jansson, S. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr & Gray). *Science*, **313**, 1596–1604.

Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* **5**, 752–763.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. and Vandepoele, K. (2009a) The flowering world: a tale of duplications. *Trends Plant Sci.* **14**, 680–688.

Van de Peer, Y., Maere, S. and Meyer, A. (2009b) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732.

Vandepoele, K. and Van de Peer, Y. (2005) Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol.* **137**, 31–42.

Vandepoele, K., Simillion, C. and Van de Peer, Y. (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.* **18**, 606–608.

Velasco, R., Zharkikh, A., Affourtit, J. *et al.* (2010) The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.* **42**, 833–839.

Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in Arabidopsis. *Science*, **290**, 2114–2117.

Wicker, T., Sabot, F., Hua-Van, A. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* **8**, 973–982.

Wikström, N., Savolainen, V. and Chase, M.W. (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Lond. B Biol. Sci.* **268**, 2211–2220.

Worden, A.Z., Lee, J.H., Mock, T. *et al.* (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science*, **324**, 268–272.