Evolution of Plant Genomes

Phillip McClean
April, 2012

Introduction

The genomes of modern plant species are quite variable. For example, the size of some are quite small, such as the ~150 megabase (Mb) *Arabidopsis thaliana* genome. Others are quite large, for example the 18,000 Mb hexaploidy wheat genome. (See http://data.kew.org/cvalues/ for a searchable database containing the genome size of many plant species.) Understanding the evolutionary history of genomes is essential for several reasons. From an applied genetics perspective, evolutionary history is critical for the application of comparative genomics for gene discovery. For example, the Arabidopsis terminal flower 1 (*tfl*1) gene in encodes a factor the controls the indeterminacy/determinacy phenotype. Additional research using *tfl*1 as a reference gene, discovered a homolog of this gene also controls the phenotype in other dicot species such as snapdragon (*Antirrhinum*) and pea (*Pisum sativum*) and in the monocot rice (*Oryza sativum*). In each case, the mutations in the gene results in a determinate phenotype.

So the relevant question is to what degree are functional genes in one plant species conserved in another species. To answer this type of question, it is important to trace the evolutionary events that lead to the current organization of plant genomes.

Polyploidy and the Construction of Plant Genomes

Background. A common event in the evolution of nearly all (if not all) plant species is a whole genome duplication (WGD). WGDs occur when the entire genome is doubled in size by either duplicating the same genome or two diploid species merging. For example, if during mitosis the chromatids migrate to separate daughter cells, rather moving to only one cell, that cell will be a tetrapolid contaning 2x the normal number of chromosomes. If this duplicate cell then gives rise to a cell involved in reproduction, the resulting gamete will have 2x the normal number of cells. And if this gamete unites with another gamete with 2x number of chromosomes, then the offspring will be tetraploid.

Tetraploid are one class of organism called polyploids. A polyploidy is an organism that contains extra sets of chromosomes. For example, the cultivated potato and alfalfa species are tetraploid. The success of any polyploidy depends upon its ability to generate balanced gametes. A balanced gamete is one that has the proper number of chromosomes relative to other gametes in the species. Only gametes with the same number of gametes can successfully merge.

This is just one class of polyploids. Another class is called allopolyploids. There occur when two species with very similar chromosomal structure and number intermate. Following a chromosomal doubling event, the resulting organism will have a number of chromosomes equal to the sum of the number of chromosomes from each of the parent species. Good examples of

allolopolyploid species are the tetraploid durum wheat (x=14) and the hexaploid bread wheat (x=21). Durum wheat arose from the union of two diploid species (x=7) species, while bread wheat arose by the mating of a diploid wheat species with the tetraploid wheat species.

Constructing the A. thaliana genome as a model for eudicot genome evolution. The availability of its whole genome sequence allowed researchers to study the duplication history of the *A. thaliana* genome. In particular, signatures of ancestral duplication events could be inferred. First the researcher uses a blastp analysis (protein vs. protein comparison) to identify those pairs of genes that meet a specific criteria (E-value < -10 used in Fig. 1) that suggests they are ancestrally related. Then these duplicates are mapped based on their relative position in the genome. This is typically displayed using a dot blot. Blocks, consisting of a cluster of linear arrayed dots, that form a diagonal in the dot blot, represent duplicate genes that are considered to be signatures of a duplication event.

Let's discuss Fig. 1 below. This was from an early paper that compared the proteins among each of the five *A. thaliana* chromosomes. Of interest are the red and green diagonals in the upper right panel. Do you see the fragment labeled α3 in the chromosome 1 vs. chromosome 1 block? This is a signature of a duplicated block of genes. Here we have genes that have the same conserved order found near the two ends of the *A. thaliana* chromosome 1. You will notice a longer block called α5 that is also duplicated on chromosome 1. Now go to chromosome 3. Here there is a block called α8. Again, this block is a consecutive set of genes shared between the two chromosomes. The largest conserved block, α11, contains genes from the ends of chromsomes 3 and 2. All totaled, there are 27 major duplicated blocks. Because of their strong signals of similarity across longer lengths of the genome, these are considered to be signals of a recent duplication.

So how does this relate to the mechanism of genome construction? This is rather simple. At some point in the recent past, the entire genome of *A. thaliana* was duplicated. Following that event individual chromosomes were broken, and the individual fragments were rearranged into new chromosomes. That rearrangement resulted in new chromosomes that were comprised of blocks of DNA from the progenitor species.

Next we need to consider the structure of the progenitor genome and how it was affected structural by the duplication event. To address this question, researchers typically look to another species that is evolutionary close. For *A. thaliana*, the species used was *A. lyrata*. This species has eight chromosomes compared to *A. thaliana* which has five. To determine the evolutionary history, genetic maps, developed using shared loci were compared. Fig. 2 shows this comparison.

This result shows that the five *A. thaliana* chromosomes are actually constructed from fragments of an ancestor with a chromosome number of eight, similar to that found in *A. lyrata*. If you look at At Chr 1 you can see that one end consists of a block of DNA from AlyLG1 and another block fro AlyLG2. Similarly, At Chr II was constructed from blocks of AlyLG3 and AlyLG4. Therefore, even though the two species have different chromosome numbers, they consist of the same chromosomal blocks.

Figure 1. Dot blot display revealing duplication events. (from Bowers et al. 2003. Nature 422:433)

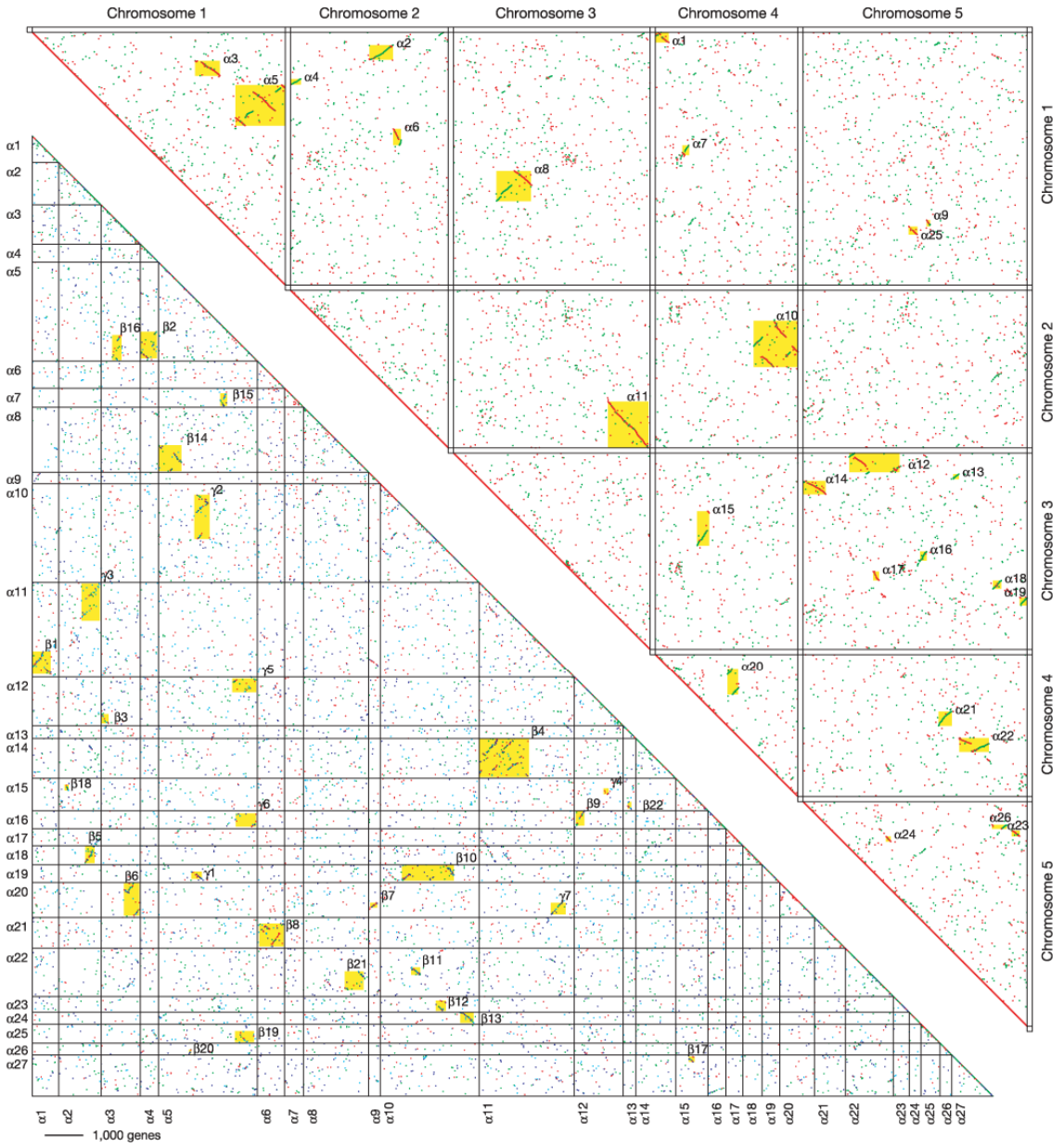


Figure 1 Arrangement of duplicated protein-encoding genes in *Arabidopsis thaliana*. Top right, a, duplications. Both x and y axes represent 26,028 genes in their chromosomal order. The best-matching gene pairs are plotted, colour-coded to indicate same (red) or opposite (green) transcriptional orientations. For further analysis, 57 adjacent duplicated regions with opposite orientation and order explicable by localized inversions were excluded, and combined into 26 'large' duplications (α1–α26) that each included ≥1% (260) of the genes. Eight shorter duplications were present (lower left, b, and g, duplications. Both x and y axes represent 21,749 genes, in an inferred ancestral order that accounts for the composition of the 26 large duplications (at left and bottom). Twenty-nine duplications (see text) are highlighted. Colours show how the four modern *Arabidopsis* chromosome segments contributed to duplications, distinguishing contributions to the duplicated segments at left and bottom respectively from the: (1) lower-numbered chromosomes (red); (2) higher- and lower-numbered chromosomes (light blue); (3) lower- and higher-numbered chromosomes (dark blue); (4) higher-numbered chromosomes (green). Higher-resolution versions of the figure and lists of gene orders are available (see Supplementary Information).

Figure 2. Comparative physical map of *A. thaliana* and the genetic map of *A. lyrata*. (from: Yogeeswaran et al. Genome Research 15:505)

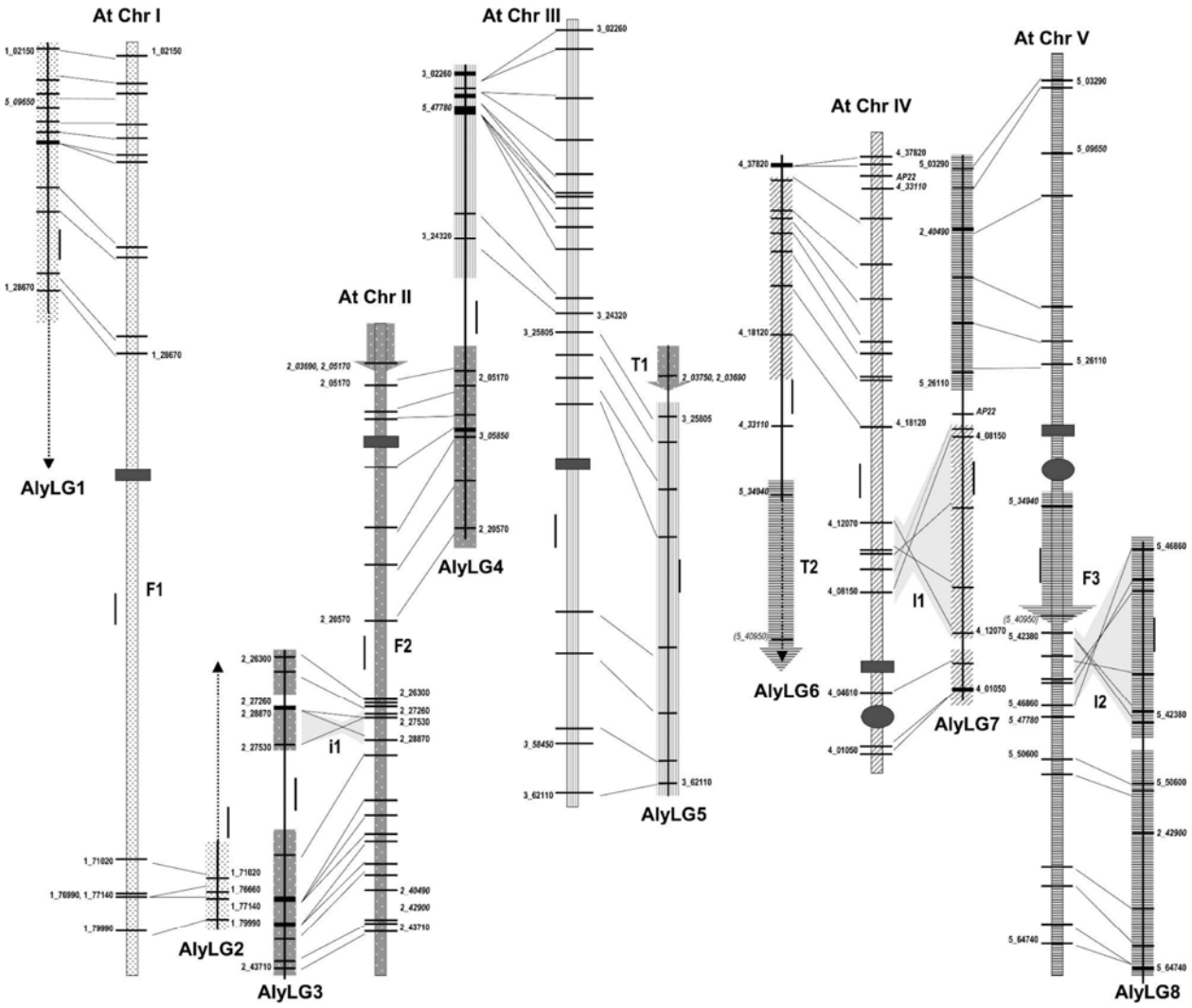


Figure 2. Colinearity of A. l. lyrata linkage map with the A. thaliana genome. A. thaliana chromosomes (At Chr I –V) are represented as patterned bars (drawn to scale, 1 unit = 1 Mbp; gray rectangles, centromeres; gray circles , heterochromatic knobs). A. l. lyrata linkage groups (Aly LG 1 –8) are shown in black (drawn to scale, 1 unit = 5cM). Sixteen colinear blocks are highlig hted with the same pattern as the At chromosome to which they correspond. Markers defining the ends of each colinear block are shown on the map in blac k lettering. Markers mapping with LOD score less than 3.0 are featured in parentheses. Italicized markers map to translocated or nonsyntenic re gions in A. l. lyrata . Translocations T1 and T2 are highlighted by arrows whose patterns correspond to the At chromosome where their colinear region lies . Major inversions I1 and I2 and minor inversion i1 are highlighted in light gray. Three chromosomal fusions are denoted as F1-F3.

Early duplication events. Fig. 1 also provides evidence of more ancient duplications. The authors reoriented the 27 α duplications. For example, notice block α5. This consists of two duplicates blocks in the same order and two in an opposite orientation. These four blocks were ordered to represent the presumed ancestral order. Then the 27 reoriented blocks were subjected to the same analysis that uncovered the α blocks. Here two types of blocks were discovered. These consist of 22 β and 7 γ events. The β events are thought to represent another duplication event in the *A. thaliana* lineage.

So what about the γ event? This has been more controversial. Some have argued it occurred early in the angiosperm lineage, while others suggest it occurred after the split of monocots and dicots. The grapevine genome provided important evidence that appears to have resolved this question. Grape is considered an ancestor of the rosids, the group of species that include *A. thaliana*. Using the same dot blot approach, it was shown that most regions of the grape genome share a common set of genes with two other regions of the genome. That is depicted below in Fig. 3. This would suggest that the grape genome has a hexaploid history. How about other species, can this signal of hexaploidy be detected. Next, the grape and poplar genomes were compared using just those regions that were triplicated in grape. In Fig. 4, it is clear that each of the triplicated regions are found in two copies in poplar. This means the hexaploid ancestry concept is correct, and that the poplar genome has undergone an additional WGD after its divergence from the grape lineage.

To address the issue of a shared duplication event in the dicot and monocot lineages, a similar comparison of the orthologs between grape and rice. If rice shared the hexaploid ancestry, a three-to-three relationship would have been observed. Rather it was observed that a three-to-one relationship existed between these two genes. This implies that monocots, as represented by rice, did not share the hexaploid history. (**Note**: See Tang et al. 2008. Genome Research18:1944 for an alternative perspective.)

This research has been summarized in the following manner. The first event was a mating between diploid to generate tetraploid species. This species was next mated to another diploid to produce the ancestral hexaploid. All subsequent eudicots would presumably have signatures of this event in their genome history.

Monocot genome evolution. As with eudicots, the monocots also have evidence of a duplication history. An early study (see Fig. 5) compared rice and maize. By using maize (y-axis) as the reference, it can be seen that most sequences are indeed found in two copies in rice. In addition, using rice chromosomes as a reference, some blocks are found three or four times in maize. (See rice chromosome 1 as a reference of three blocks, and rice chromosome 4 as a reference as an example of four blocks.) This figure allows us to conclude: 1) there was a WGD in the history of monocots; and 2) an additional duplication occurred in the maize lineage.

The publication of the rice, sorghum, Brachypodium (a model grass species), and maize genome sequences has lead to a unified model of grass evolution. 56-73 MYA, an ancestral grass species containing five chromosomes was duplicated to generate a genome with ten chromosomes. Then two different pairs of the ten chromosomes underwent genome intrachromosomal breakage/translocation/fusion events that constructed two additional chromosomes. (See Fig. 6, where A7 and A10 in the paleotetraploid fractionated and constructed the A7, A10, and A3 chromosomes. A similar event occurred between A4 and A6 to also form A2.) These basic set of 12 chromosomes formed the basis of all of the modern grasses.

By convention, the rice genome is thought to represent the ancient paleotetraploid. This basic set of chromosomes served as the building blocks for other genomes. As seen in Fig. 6, Brachypodium, the Poideae (representing the wheat lineage), and the Panicoideae (representing

Figure 3. Dot blot representation of duplicate regions of the grapevine genome. (from: Jaillon et al. 2007. Nature 449:463)

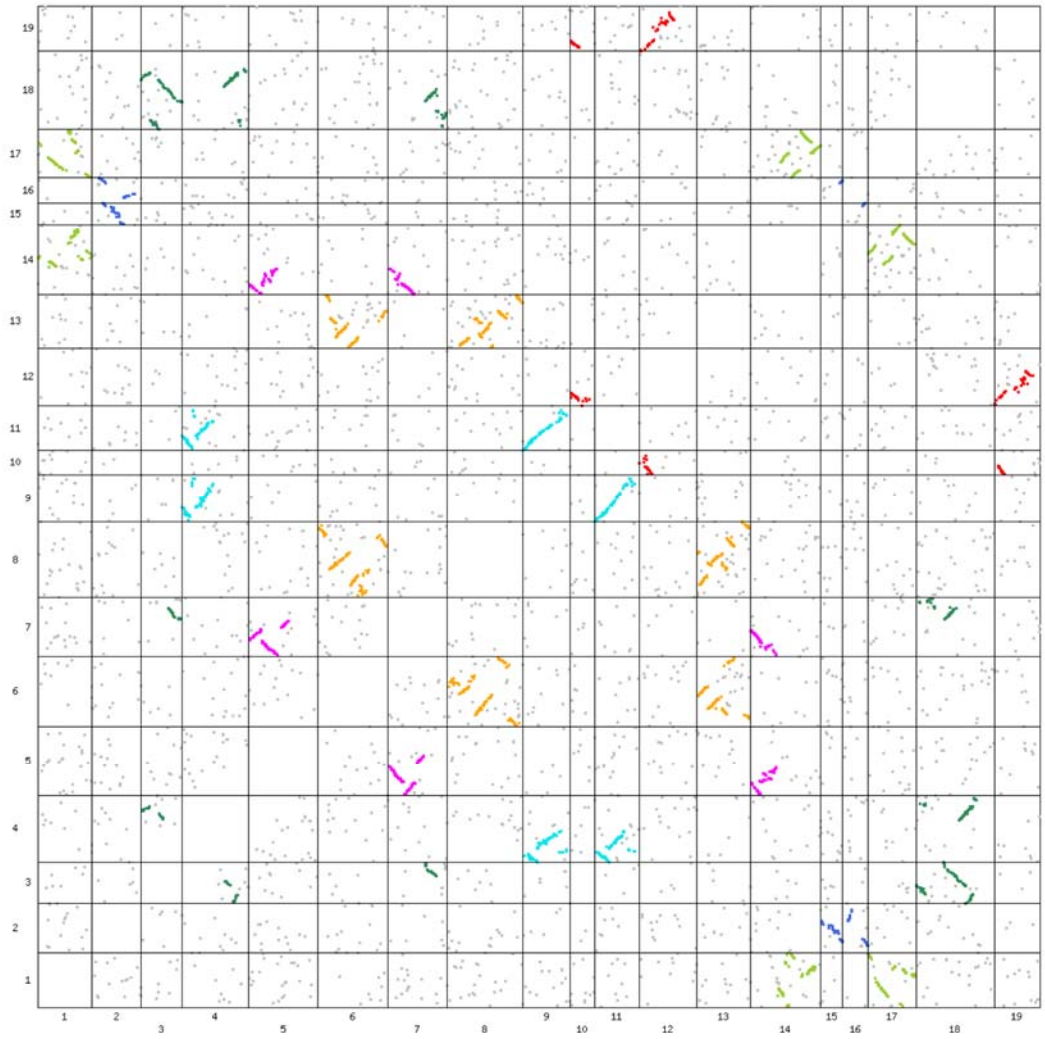


Figure S5. The grape genome originated from a polyploidy event that joined three ancestral genomes. The nineteen chromosomes of grape are represented on both the x and y axis. Dots represent the positions of paralogous pairs of genes. For clarity, intrachromosomal paralogs are not shown. Clusters of paralogs form a succession of dots, that indicate that the gene order of the ancestral genome was locally maintained. These clusters are painted in seven colours. Each colour marks paralogous blocks, that were colinear in the ancestors of the three constituents of the grape genome. Some regions are not painted in triplicate in this grid, either because a whole region is not visible in synteny with two others in the present-day grape genome (too many rearrangements or gene loss), or because one or two syntenic regions lie in supercontigs which are still not anchored.

Figure 4. Comparison of the triplicated blocks and the Poplar genome. (from: Jaillon et al. 2007. Nature 449:463)

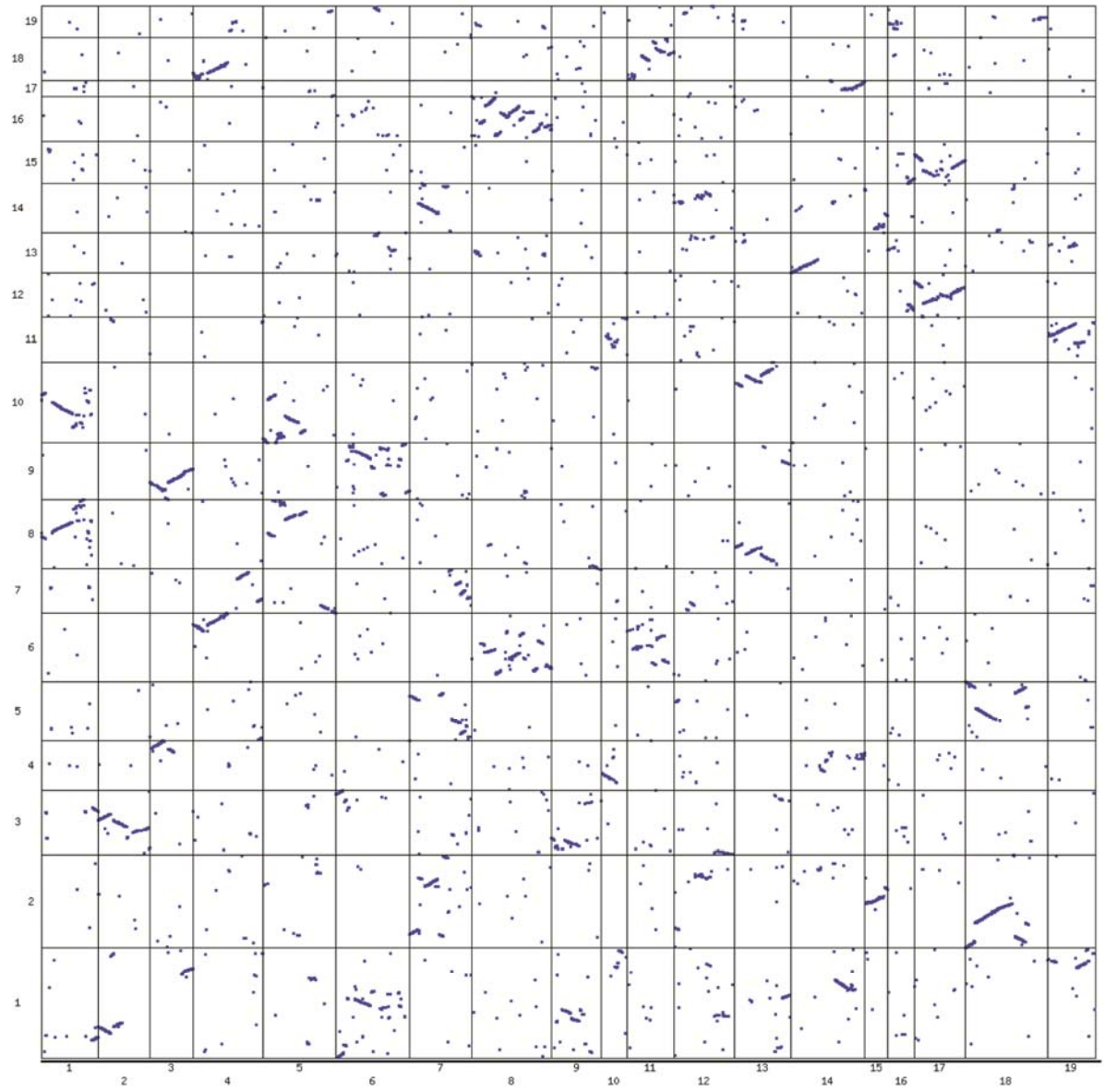


Figure S6. The distribution of 8,604 orthologous genes between *Vitis vinifera* (x axis) and *Populus trichocarpa* (y axis) chromosomes.

Figure 5. A comparison of maize and rice duplication events. (from: Wei et al. (2007) PLoS Genetics 3(7):e123, 1254)

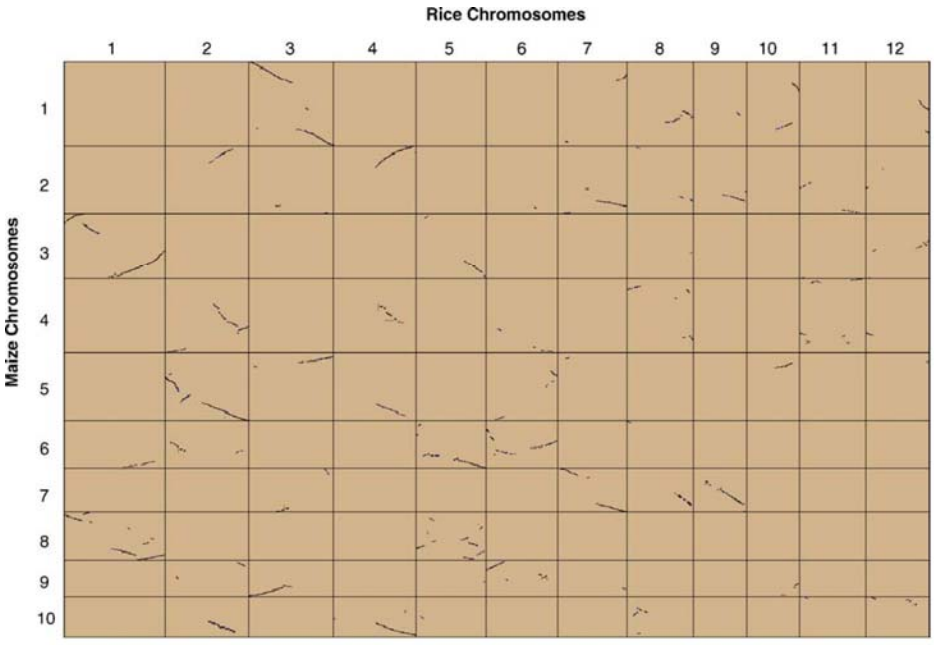


Figure 1. Dotplot Analysis of the Integrated Maize Map against Rice Pseudomolecules

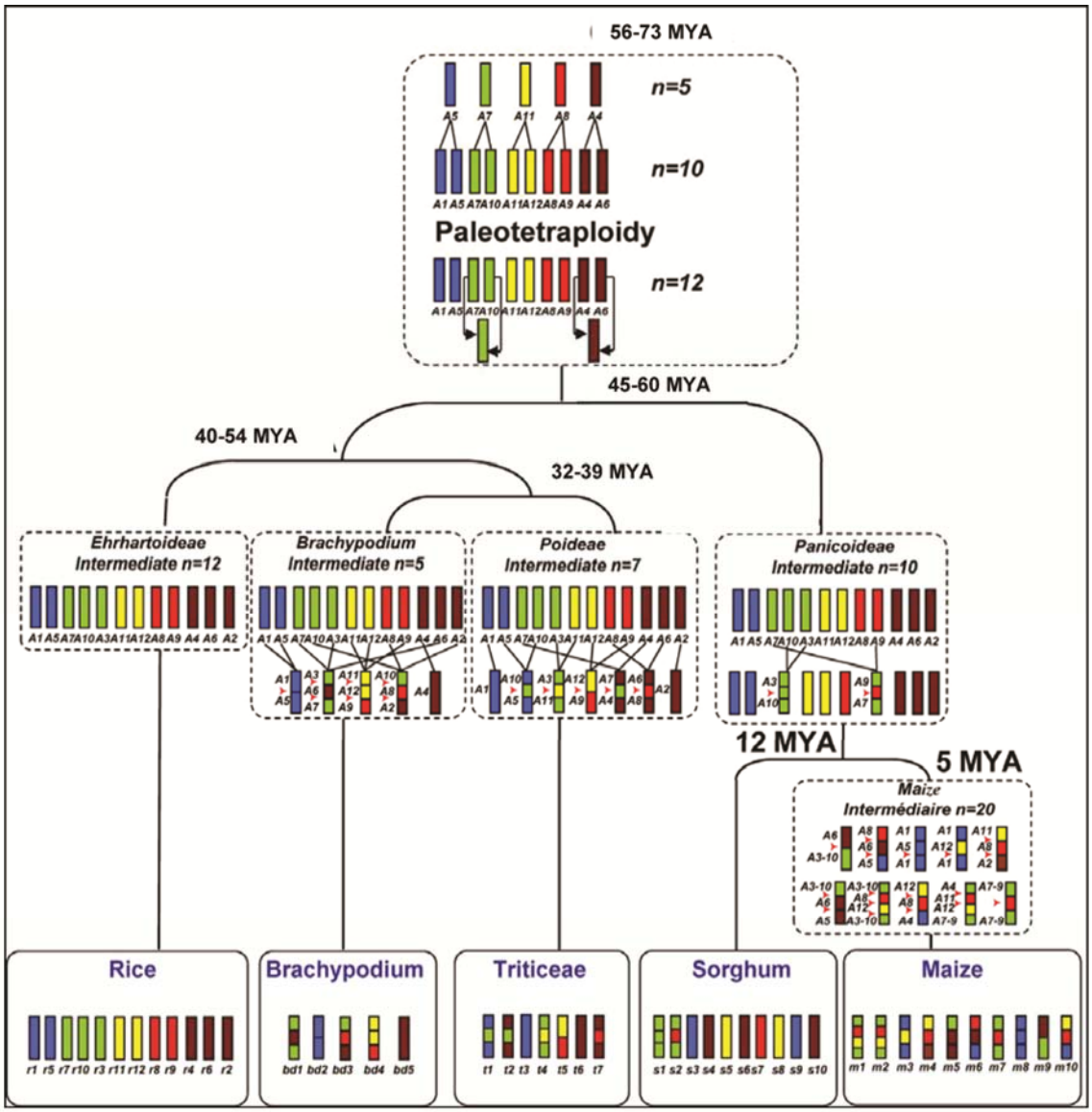
Synteny blocks were detected, and background noise was filtered with SyMAP [37]. The interactive dotplot can be viewed at http://www.agcol.arizona.edu/symap. When clicking the related synteny block, the detailed window with contig number will pop up. The viewer can select the preferred area and double click the selection, and then a graphic alignment is displayed.

doi:10.1371/journal.pgen.0030123.g001

the maize/sorghum lineage), each arose by a series of breakage/translocation/fusion events involving chromosomal fragments from the n=12 ancestor. While the other lineages had more complex patterns of evolution, the Panicoideae had the simplest history by arising from only four breaks. This chromosomal pattern is currently maintained by the sorghum genome structure. The maize genome then underwent additional evolution that included another whole genome duplication event and additional breakage/translocation/fusion events.

Summary. Plant genomes, unlike animal and fungal genoemes, have a long history of genome duplications. These are collectively illustrated in Fig.7. (Based on the evidence above, the γ event should be moved to the origin of the eudicot lineage.) This figure clearly shows the significant role of WGD in development of plant species. The fact that many of the duplications appear 55-70 MYA is interesting. This is at the transition point between the Cretaceous and Tertiary periods. At about this time, there was a mass extinction of species. Some think that duplications gave plants the needed gene repertoire needed to survive this event and ultimately flourish on earth (see Fawcett et al. 2009. PNAS USA 106:5737). Recently, Jaio et al (2011, Nature 473:97) extended this analysis to include species deeper into the plant phylogeny. Here the added to addition duplication events, one associated with ancestral seed plants (ζ at ~330 MYA) and a second associated with ancestral angiosperms (ε at ~220 MYA). Collectively, the evidence is clear that genome duplications are a major event in the evolutionary history of plants.

Figure 6. A unified model of grass genome evolution. (from: Vogel et al. 2010. Nature 463:763.)



Supplementary Figure 18. Grass chromosome evolution model. The monocot chromosomes (r1-r12 for rice, t1-t7 for Triticeae, bd1-bd5 for *Brachypodium,* s1-s10 for sorghum, and m1-m10 for maize) are represented with a five colour code to illustrate the evolution of segments from a common ancestor with five proto-chromosomes and a n=12 intermediate as described in [62], and are named according to the rice nomenclature. The events that have shaped the structure of the 5 different grass genomes including the 7 *Brachypodium* chromosome nested insertion events during their evolution from the common ancestor are indicated as whole genome duplication, ancestral chromosome translocations and fusions, and lineage- specific nested chromosome insertions.

The Gene-based Evolution of Duplicated Genes

If duplications are a major signature of plant genomes, then the number of genes should correspond to the number of rounds of duplication. Table 1 list the number of genes found within each of the species for which a complete genome sequence is currently available. If the hexoploidy concept is true for dicots, and grape only contains this hexaploid event, then it can be estimated that the ancestral dicot contained ~10,000 genes (=30,000/30). Following this conclusion, and based on the fact that poplar underwent an additional duplication, it should contain 60,000 genes, and *A. thaliana* which underwent two duplications relative to the shared ancestor should contain 120,000 genes. And since this clearly is not the case as Table 1 shows.

A similar argument can be made for the monocots. Given that rice, Brachypodium, and sorghum only contain a duplication event, then it can be argued that the basic number of monocot genes is 15,000 (=30,000/2). Maize has undergone an additional duplication event, but has undergone a reduction to ~30,000 gene suggests that it was necessary to reduce the number of genes to ensure the success of the species.

Diploidization. The polploid past was a surprising result of the Arabidopsis and rice genome sequencing. It was surprising because these species were selected for sequencing because of their small genome sizes. So what has been the consequences of this polyploidy? One obvious consequence of polyploidy is the doubling or tripling of the number of chromosomes. This is clearly evident for monocots. So what about the fate of the additional gene set that are the product of the WGD? It is generally thought that a species cannot maintain the entire set of duplicate chromosomes and genes because it provides the genetic material that can generate deleterious mutations that compromises the fitness of a genome. Therefore it is important for the duplicate genome to transition back to its original state. This process is call ***diploidization***. To revert back to the diploid state many duplicate genes must be eliminated from the gene set. But a recently duplicated genome, such as soybean, appears to be able to withstand the extra copies since its gene number is roughly twice that of the basic set of 30,000 genes detected in the hexoploid ancestral eudioct.

Several genomic events appear to be associated with diploidization. Firstly, the duplicate genome must change its chromosome pairing pattern. Immediately after the duplications, the four chromosomes will pair and form quadravalents. This chromosomal structure must be changed so that bivalents are formed. This will result in a doubling of the chromosome number as seen for the monocot lineage. Once bivalents are formed, the gene sets can evolve by processes such as deletions and chromosomal rearrangements. Then the duplicate genes can undergo specific changes. A most common fate is death of the additional copies of the genes. These loses can accompany such events as chromosomal breakage and rearrangements. Following these events, a new basic set of chromosomes and genes will have appeared.

The fate of all duplicate genes is not similar. By analyzing different genes, it appears that some are retained as multicopy (up to the ploidy level for that species). Other genes seem to be reduced to only a single copy. Therefore, some genes can be considered to be "deletion resistant" while others are "duplication resistant." Annotation data suggest many of the

Figure 7. A summary of the duplication history of plants. (from Van de Peer et al. 2009. Trends in Plant Sciences 14:680)

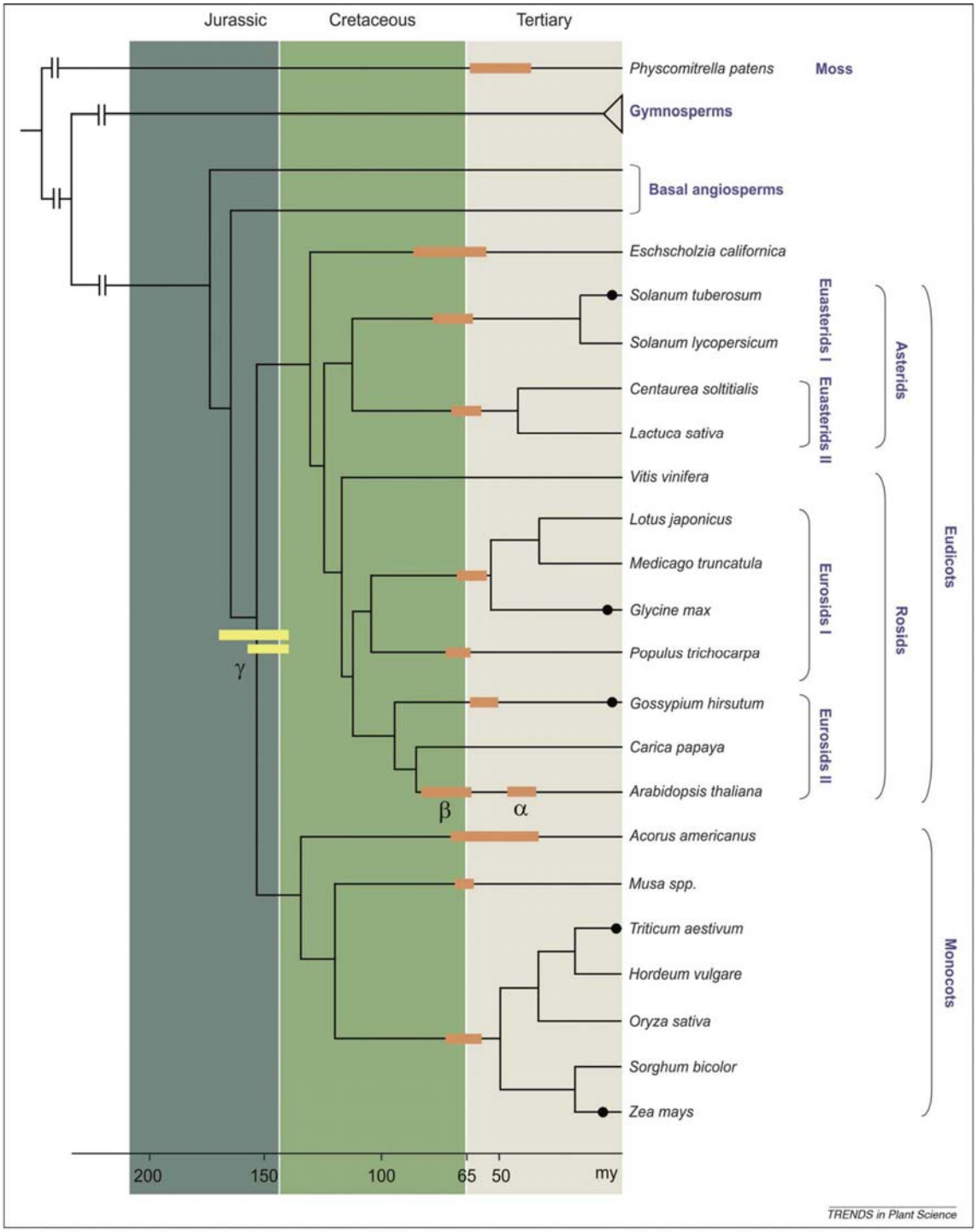
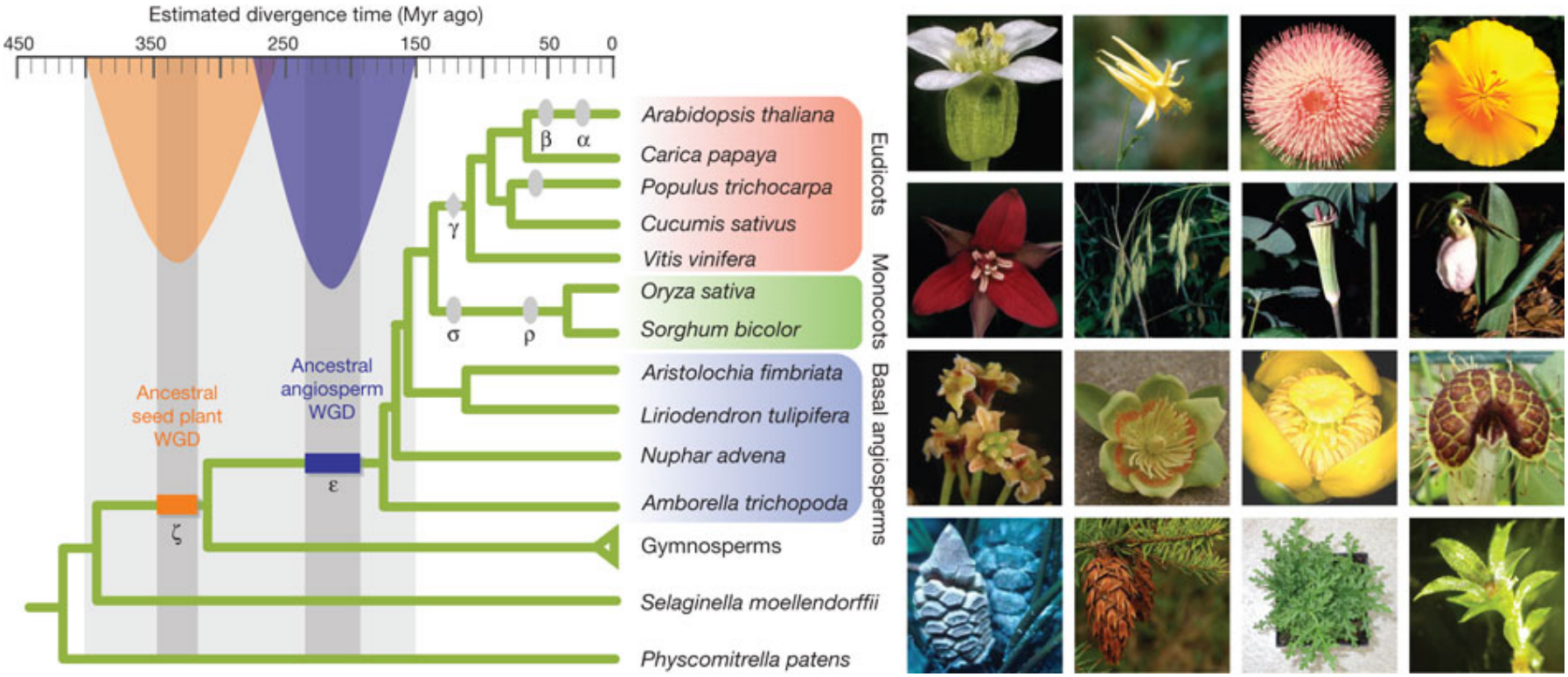


Figure 2 . Phylogenetic tree offl owering plants (eudicots and monocots). WGDs, inferred from recent studies [28–30], are indicated by horizontal bars. Yellow bars denote the hexaploidy event. More recent WGDs appear to be clustered around the KT boundary [29]. The black dots indicate recent polyploidy events [~1–2 mya in cotton (Gossypium hirsutum), < 10 mya in potato (Solanum tuberosum), ~10–15 mya in soybean (Glycine max), ~10 mya in maize (Zea mays), and < 1 mya in wheat (Triticum aestivum)]. Alpha, beta and gamma denote the generally accepted duplication events in Arabidopsis [5–7,36] (see main text for details). Modified with permission from [29].

Figure 8: Ancestral polyploidy events in seed plants and angiosperms. [Jiao et al (2011) Nature 473:97]



Legend from Figure 3 of Nature article: Ancestral polyploidy events in seed plants and angiosperms. Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see text). The diamond refers to the triplication event probably shared by all core eudicots. Horizontal bars denote confidence regions for ancestral seed plant WGD and ancestral angiosperm WGD, and are drawn to reflect upper and lower bounds of mean estimates from Fig. 2 (more orthogroups) and Supplementary Fig. 5 (more taxa). The photographs provide examples of the reproductive diversity of eudicots (top row, left to right: *Arabidopsis thaliana*, *Aquilegia chrysantha*, *Cirsium pumilum*, *Eschscholzia californica*), monocots (second row, left to right: *Trillium erectum*, *Bromus kalmii*, *Arisaema triphyllum*, *Cypripedium acaule*), basal angiosperms (third row, left to right: *Amborella trichopoda*, *Liriodendron tulipifera*, *Nuphar advena*, *Aristolochia fimbriata*), gymnosperms (fourth row, first and second from left: *Zamia vazquezii*, *Pseudotsuga menziesii*) and the outgroups *Selaginella moellendorfii* (vegetative; fourth row, third from left) and *Physcomitrella patens* (fourth row, right). See Supplementary Table 4 for photo credits.

Table 1. The estimated number of genes in sequenced plant genomes.

Species	Estimated # of Genes (from www.phytozome.net)
Eudicots	
Cucumber	21,491
Cassava	47,164
Poplar	41,000
Medicago	50,692
Soybean	66,153
Arabidopsis	27,343
Papaya	27,332
Grape	30,434
Mimulus	25,530
Monocots	
Sorghum	34,496
Maize	32,540
Brachypodium	25,532
Rice	31,500

"duplication resistant" genes encode enzymes or genes of unknown function. Similarly, the "deletion resistant" genes mainly encode transcription factors.

Developing new functions. Many of the duplicate set of genes cannot be maintained for long because deleterious mutations. Rather the collection of duplicate genes are modified. These gene changes will provide new or altered functions, and these new functions may lead to the evolution of the species, potentially to a higher level of fitness. One example of an evolutionary modification of duplicate genes is called ***neofunctionalization***. Here one duplicate gene maintains its original function, while evolution can act on the second copy and generate a new function that may increase the adaptability of an individual. A second process called ***subfunctionalization*** modifies the duplicates. The basic structure of both copies can be altered so that the expression pattern of the gene changes. Typically, this change results in a higher level of the protein the gene encodes. Alternately, the function of the original gene is maintained, but the structure of both copies is significantly changed. The collective changes to the copies though retains the function of the original gene.

Synteny: The Result of WGD and Reconstructing Plant Genomes

A major result of the duplication history is the syntenic relationships among plant species. Synteny is the maintenance of gene order between two species. The classic approach to synteny studies was based on shared markers mapped onto two different species. In this manner, the macrosyntenic relationships can be determined. Macrosynteny is detected by the large scale

Figure 9. Macrosynteny between tomato and eggplant, including a QTL for a shared domestication trait. (from: Doganlar et al. 2002. Genetics 161:1713.)

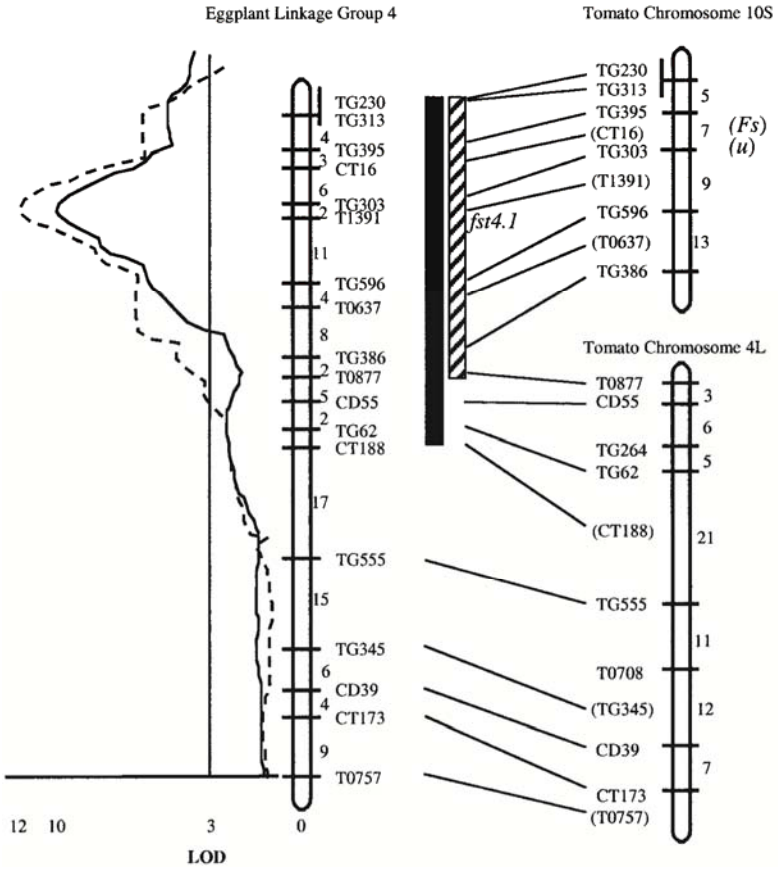


Figure 3.—Comparative mapping off ruit stripe locus on eggplant linkage group 4. Simple interval analysis for fst4.1 is shown to the left of the molecular map of eggplant linkage group 4 (solid line for NY data, dashed line for FR data). Bars to the right of the linkage group represent the position of the QTL as determined by single-point regression analysis ($P \leq 0.05$; see Table 1 for details; solid bar for NY data, hatched bar for FR data). Molecular maps for tomato chromosome arms are from Tanksley et al (1992).

chromosomal blocks shared by two species. Fig. 9 is an example of macrosynteny between tomato and eggplant. Here we can see that eggplant linkage group 4 is evolutionarily related to tomato linkage groups 10S and 4L. This figure shows the highly conserved marker order over many centimorgans of the two genomes.

The genetic mapping of shared loci is the first method of comparing species. This is actually the only way to compare species that have not been sequenced. There are many examples of synteny mapping in plants. The power of synteny mapping is best revealed by the discovery of loci from two species that control the same phenotype and map to the same genetic location. Fig. 9 shows that a major QTL for fruit striping is found on one end of eggplant linkage 4. Previous work with tomato showed that a major gene for this trait was located on linkage group 10. The fact that multiple loci are shared in the same macrosyntenic order suggests that the same ancestral gene is controlling this trait in these two Solanceae species.

So how can synteny be used to leverage knowledge in one species for gene discovery in a second species? If phenotypic traits have been mapped extensively in one species, that information can point a researcher working on a second species to the likely location of a similar gene in that second species. This leverage is a great aid for genetic discovery, especially for species in which the discovery of important genetic factors has been limited by a lack of funding.