General Steps in Sequencing a Plant (and other) Genome

1. Create sequencing libraries of different insert sizes

- 2kb
    - Bulk of sequencing is performed on these libraries
- 10kb
    - Used for linking contigs during assembly
- 40kb
    - Used to link larger contigs assembly
- Bacterial artificial chromosomes
    - Used to link ever larger contigs assembly

2. Paired-end sequencing data collected for libraries

3. Contigs created by looking for overlapping reads

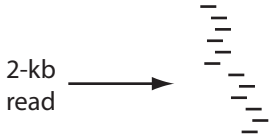
4. Contigs assembled based on homology to 10kb, 40kb and BAC sequence data; these large assemblies are called **scaffolds**

5. Pseudochromosomes are assembled based on homology of scaffolds to the markers located on a high-density genetic map

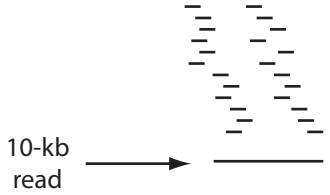
Scaffold Assembly

Building a Scaffold Using Paired-end Reads of Different Sized Sequences

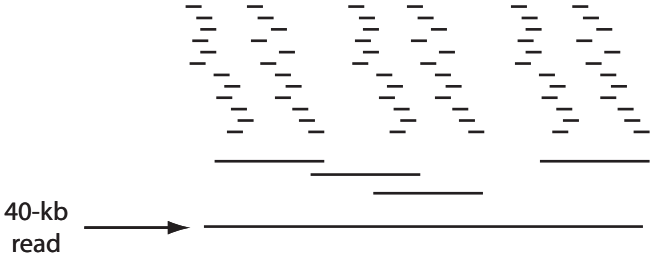
Step 1: Build a contig with overlapping 2-kb paired-end reads



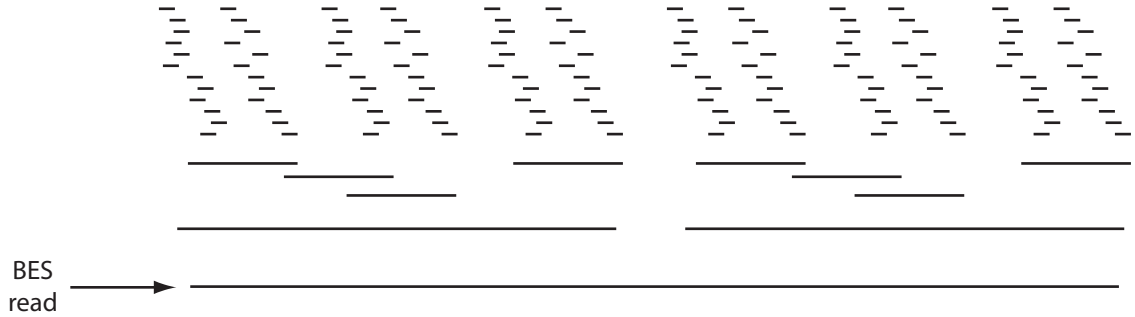
Step 2: Link two contigs with 10-kb paired-end reads



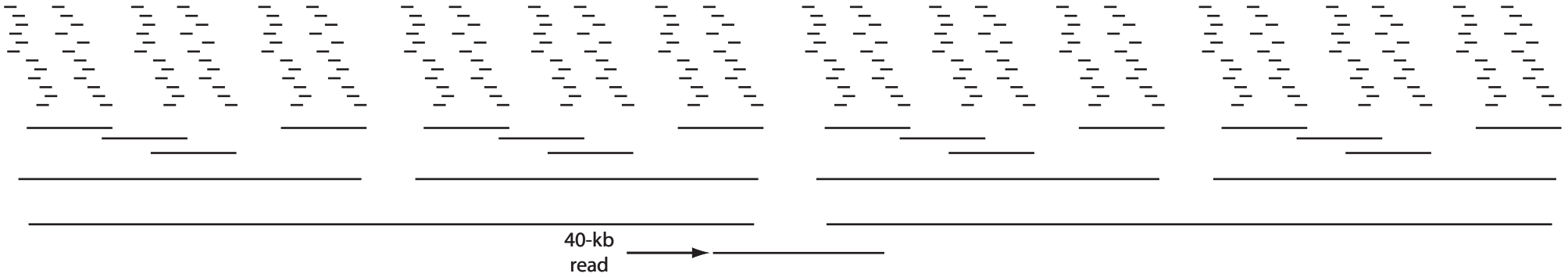
Step 3: Link three 10-kb contigs with 40-kb paired-end reads



Step 4: Link two 40-kb contigs with 100-kb BAC end sequences (BES)



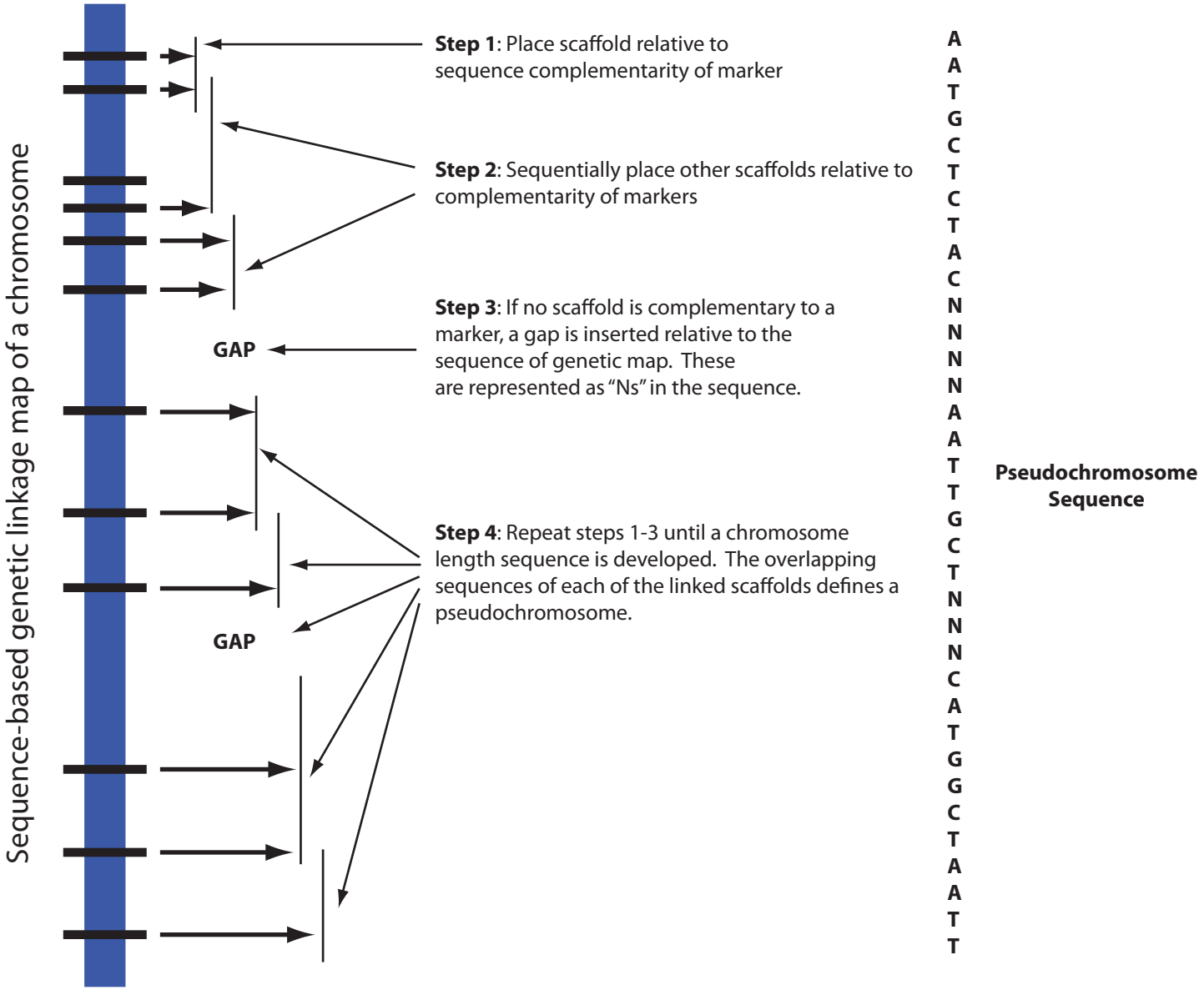
Step 5: Here link two100-kb BAC sized contigs with a 40-kb paired-end read; other sized reads can also be used for this linking



Step 6: Continue linking larger blocks of sequences until the block can not be linked with another block. This block is defined as a scaffold.

Genome Assembly

Linking Scaffolds to a Dense Genetic Map



Phaseolus vulgaris

Summary Genome Sequencing and Assembly

Production Information

- Sequence technology: Sanger, Roche 454, Illumina
- Number of libraries: 21 (15 paired, 6 unpaired)
- Total Reads: 49,214,786 (10,696,722 successful paired-end reads; 2.3% failed)
- Coverage: 21.02x total (18.64X linear, 3.38X paired-end)

Assembly Information

Summary information	**Statistic**
Main genome scaffold total	708
Main genome contig total	41,391
Main genome scaffold sequence total	521.1 Mb
Main genome contig sequence total	472.5 Mb (9.3% gap)
Main genome scaffold N50/L50	5/50.4 Mb
Main genome contig N50/L50	3,273/39.5 Mb
Number of scaffolds > 50 Kb	28
% main genome in scaffolds >50 Kb	99.3%

Estimated genome coverage from Kew Gardens C-value Database

- *P. vulgaris* = 0.6 picograms
  - 1 pg = 978 megabases
    - *P. vulgaris* **= 586.8 Mb**

Coverage

- 521.1 Mb/586.8 Mb = **88.8% coverage**

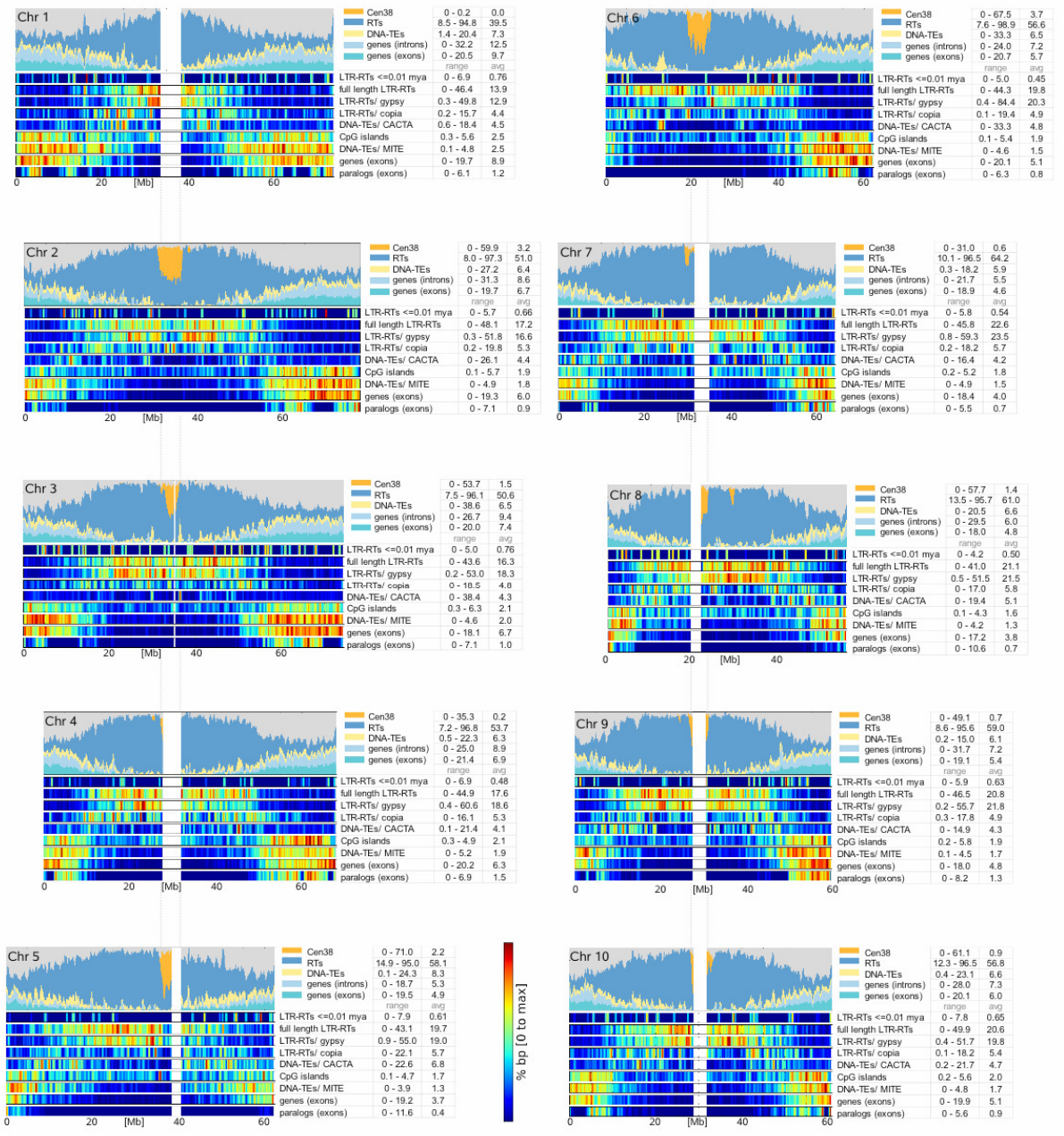
Assembly Notes

- **Initial assembly**
  - Arachne assembler
  - Assembly checked using
    - Genetic map
      - 7,015 markers used
    - Soybean/common bean synteny used to detect misjoins
      - 71 breaks were applied to initial assembly

- **Final assembly**
  - Genetic map and soybean/common bean synteny applied to data
  - 248 joins applied to broken assembly
  - 11 pseudochromosomes assembled
    - 98.8% of assembled sequence is found in the pseudochromsomes

Species name	Common name	Genotype	Year	Publication	Technical method	# Chrom	Est. genome size/assembled size (Mb)	Repeat content (%)	Chrom size range (Mb)	# genes/ transcripts	Contig N50/L50 (#/kb)	Scaffold N50/L50 (#/kb)	Genome duplication history
Arabidopsis thaliana	Arabidopsis	Columbia	2000	Nature 408:796	HSS/S	5	125/135	20[1]	18-29	27,416/ 35,386			Eudicot 3x + Brassicaceae (2x+2x)
Oryza sativa	Rice	Nipponbare	2005	Nature 436:793	HSS/S	12	430/371	45[1]	23-43	39,049/ 49,061			Poales (2x+2x)
Populus trichocarpa	Poplar	Nisqually 1	2006	Science 313:1596	WGS/S	19	485/423	40[1]	11-36	41,335/ 73,013	??/126	??/3,100	Eudicot 3x + (2x)
Vitus vinifera	Grape	PN40024	2007	Nature 449:463	WGS/S	19	475/487	22[1]	10-22	/ 26,346	??/126	??/2,065	Eudicot 3x
Carica papaya	Papaya	Sunup	2008	Nature 452:991	WGS/S	9	372/370	52		27,332/ 27,996	??/11	??/1,000	Eudicot 3x
Sorghum bicolor	Sorghum	BTx623	2009	Nature 457:551	WGS/S	10	818/727[2]	63[1]	50-70	33,032/ 39,441	958/195	6/62,400	Poales (2x+2x)
Zea mays	Maize	B73	2009	Science 326:1112	HSS/S	10	/3,234	84	150-301	39,475/ 137,208			Poales (2x+2x) + (2x)
Cucumis sativus	Cucumber	9930	2009	Nat Genet 41:1275	WGS/S,I	7	??/244	22[1]		21,491/ 32,528	??/227	??/1,140	Eudicot 3x
Glycine max	Soybean	Williams 82	2010	Nature 463:178	WGS/S	20	1115/978	57	37-62	56,044/ 88,647	1,492/189	10/47,800	Eudicot 3x + Legume 2x + (2x)
B. distachyon	Brachypodium	Bd21	2010	Nature 463:763	WGS/S	5	272/275	28	25-75	26,552/ 31,029	252/348	3/59,300	Poales (2x+2x)
Ricinus communis	Castor bean	Hale	2010	Nat Biotech 28:951	WGS/S, 454	10	320/326	~50		31,237/??	??/21	??/497	Eudicot 3x
Malus x domestica	Apple	Golden Delcious	2010	Nat Genet 42:833	WGS/S	17	742/604	36	21-47	63,538/ 63,541	16,171/13	102/1,542	Eudicot 3x + Rosaceae 2x
Jatropha curcas	Jatropha		2010	DNA Res 18:65	WGS/S		380/285	37		40,929/??	??/4		
Theobroma cacao	Cocao	B97-61/B2	2011	Nat Genet 43:101	WGS/S, 454, I	10	430/362	24	12-31	29,452/ 44,405		??/5,624	Eudicot 3x
Fragaria vesca	Strawberry	H4x4	2011	Nat Genet 43:109	WGS/S, 454, I, So	7	240/220	23		32,831/??		??/1,300	Eudicot 3x
Arabidopsis lyrata	Lyrata	MN47	2011	Nat Genet 43:476	WGS/S	8	??/207	30	19-33	32,670/??	1,309/5,200		Eudicot 3x + Brassicaceae (2x+2x)
Phoenix dactylifera	Date palm	Khalas	2011	Nat Biotech 29:521	WGS/I	18	658/381	29		28,890/??	??/6	??/30	
Solanum tuberosum	Potato	DM1-3 516 R44	2011	Nature 475:189	WGS/S, 454, I, So	12	844/727	62		35,119/ 51,472	6,446/31	121/1,782	Eudicot 3x + Solanaceae 3x
Thellungiella parvula	Thellungiella		2011	Nat Genet 43:913	WGS/454, I	7	160/137	8		30,419/??		8/5,290	
Cucumis sativus	Cucumber	B10	2011	PLoS ONE 6:e22728	WGS/S, 454	7	??/323			26,587/??	??/23	??/323	Eudicot 3x
Brassica rapa	Cappage	Chiifu-401-42	2011	Nat Genet 43:1035	WGS/I	10	??/283	40		41,174/??	2,778/27	39/1,971	Brassicaceae 2x + (2x)
Cajanus cajan	Pigeon pea	ICPL 87119		Nat Biotech 30:83	WGS/S, I	11	808/606	52	10-48	40,071	7815/23	380/516	Eudicot 3x + Legume 2x
Medicago truncatula	Medicago		2011	Nature 480:520	WGS/S, 454, I	8	454/384		35-57	44,135/ 45,888		53/1270	Eudicot 3x + Legume 2x
Setaria italica	Foxtail millet	Yugu 1	2012	Nat Biotech 30:555	WGS/S	9	451/406	40	24-48	35,471/ 40,599	982/126	4/47,300	

Species name	Common name	Genotype	Year	Publication	Technical method	# Chrom	Est. genome size/assembled size (Mb)	Repeat content (%)	Chrom size range (Mb)	# genes/ transcripts	Contig N50/L50 (#/kb)	Scaffold N50/L50 (#/kb)	Genome duplication history
Solanum lycopersicon	Tomato	Heinz 1706	2012	Nature 485:635	WGS/S,So	12	900/760	63	45-65	34,727/??			Eudicot 3x + Solanaceae 3x
Linum usitatissimum	Flax	CDC Bethune	2012	Pl Journal 72:461	WGS/I	15	373/318	24		43,484	4,427/20	132/693	Eudicot 3x + (2x)
Musa acuminata	Banana	DH-Pahang, ITC1511	2012	Nature 488:213	WGS/S, 454, I	11	??/523	44	22-35	36,542	/43	/1,311	Zingiberales 2x + (2x + 2x)
Gossypium raimondii	Cotton (B genome diploid)		2012	Nat Genet 44:1098	WGS/I	13	775/567	57	25-69	40,976/??	4,918/45	2,284/95	Eudicot 3x + Gossypium 2x
Azadirachta indica	Neem	Local tree	2012	BMC Genomics 13:464	WGS/I		??/364	13		20,169/??	??/0.7	??/452	
Gossypium raimondii	Cotton (D genome diploid)		2012	Nature 492:423	WGS/S, 454, I	13	880/738	61	35-70	37,505/ 77,267	1596/136	6/62,200	Eudicot 3x + Gossypium 2x
Prunus mume	Chinese plum	2 genotypes	2012	Nature Communications 3:1318	WGS/I	8	??/237	45		31,390/??	2009/32	120/578	
Pyrus bretschneideri	Pear		2013	Genome Research	HSS+WGS/I	17	528/512	53	11-43	42,812/??	??/36	??698	Eudicot 3x + Rosaceae 2x
Cirtullus lanatus	Watermelon	97103	2013	Nat Genet 45:51	WGS/I	11	425/354	45	24-34	24,828/??	??/26	??/2380	Eudicot 3x
Morus notabilis	Mulberry		2013	Nature Communications 4:2445	WGS/I	7	357/330	47		29,338/??	2,638/34	245/390	Eudicot 3x
Phaseolus vulgaris	Common bean	G19833	2014	Nat Genet (in press)	WGS/S, 454,I	11	587/521	45	32-60	27,197/ 31,688	3,273/40	5/50	Eudicot 3x + Legume 2x


Chr 1
Chr 6
Chr 2
Chr 7
Chr 3
Chr 8
Chr 4
Chr 9
Chr 5
Chr 10
% bp [0 to max]
[Mb]

Evolution of Plant Genomes

Introduction

Modern plant genomes are quite variable

- ~150 megabase (Mb) *Arabidopsis thaliana* genome.
- 18,000 Mb hexaploidy wheat genome.

Why understanding the evolutionary history of genomes?

- Applied genetics perspective
- Application of comparative genomics for gene discovery.
    - Arabidopsis terminal flower 1 (*tfl*1)
        - Encodes a transcription factor
        - It controls indeterminacy/determinacy phenotype
        - Arabidopsis *tfl*1 as a reference gene
            - Homolog of this gene also controls the phenotype in other
                - Dicot species
                    - Snapdragon (*Antirrhinum*)
                    - Pea (*Pisum sativum*)
                - Monocot
                    - Rice (*Oryza sativum*)
            - Mutations all results in a determinate phenotype

The relevant question

- To what degree are functional genes in one plant species conserved in another species?
    - Important to trace
        - Evolutionary events
        - Related to current organization of plant genomes

Polyploidy and the Construction of Plant Genomes

Whole genome duplication (WGD)

- Common event in the evolution of plant species
  - Entire genome doubles in size
  - Duplicates the same genome
- Two related diploid species merge
  - During mitosis
    - Chromatids migrate to separate daughter cells
  - If they movie to only one cell
    - The cell will be a tetrapolid
- If the 2x duplicate cell is involved in reproduction
  - Resulting gamete
    - 2x the normal number of cells
      - If 2x gamete unites
        - Offspring will be tetraploid

Polyploidy

- An organism that contains extra sets of chromosomes.
  - Tetraploids
    - Cultivated potato
    - Alfalfa
- For a success of any polyploidy
  - It must generate balanced gametes.
    - The same number of chromosomes as other gametes
- Embryos from gametes with the same number of gametes
  - Successfully survive

Other Polyploids

- Allopolyploids
  - Two species with very similar chromosomal structure and number intermate.
  - After chromosomal doubling organism, genome will have
    - Number of chromosomes equal to the sum of the number of chromosomes from each of the parent species.
- Examples of allolopolyploid species
  - Tetraploid durum wheat (x=14)
  - Hexaploid bread wheat (x=21).
- Durum wheat arose from
  - Union of two diploid species (x=7) species
- Bread wheat arose from
  - Diploid wheat species with the tetraploid wheat species

Constructing the *A. thaliana* genome as a model for eudicot genome evolution

- With the whole genome sequence
  - Study the duplication history of the *A. thaliana* genome.
  - Ancestral duplication signatures could be inferred
    - Blastp analysis
      - Protein vs. protein comparison
      - Identifies gene pairs
        - E-value < -10 used in Fig. 1
      - ***Suggests genes are ancestrally related***
    - Duplicates are mapped relative position in the genome
      - Displayed using a dot blot
      - Blocks observed
        - Linear arrayed dots
        - Form a diagonal in the dot blot,
          - Signatures of a duplication event

Figure 1A

- Early comparison of the proteins in the *A. thaliana* genome
  - Red and green diagonals in the upper right panel
    - Block α3
      - Chromosome 1 vs. chromosome 1 block
      - Signature of a duplicated block of genes
      - Genes that have the same conserved order
      - At two ends of the *A. thaliana* chromosome 1
    - Block α5
      - Another pairs of duplicated genes on chromosome 1
    - Block α8
      - Shared block on chromosomes 1 and 3
    - Block, α11
    - Largest block
    - Ends of chromsomes 3 and 2
  - Total
    - 27 major duplicated blocks
      - Strong signals
      - ***Signals of a recent duplication***

So how does this relate to the mechanism of genome construction?

- *A. thaliana* underwent a WGD
  - Chromosomes were broken
  - Rearranged into new chromosomes
  - New chromosomes developed
    - ***Represent blocks of DNA from the progenitor species***

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

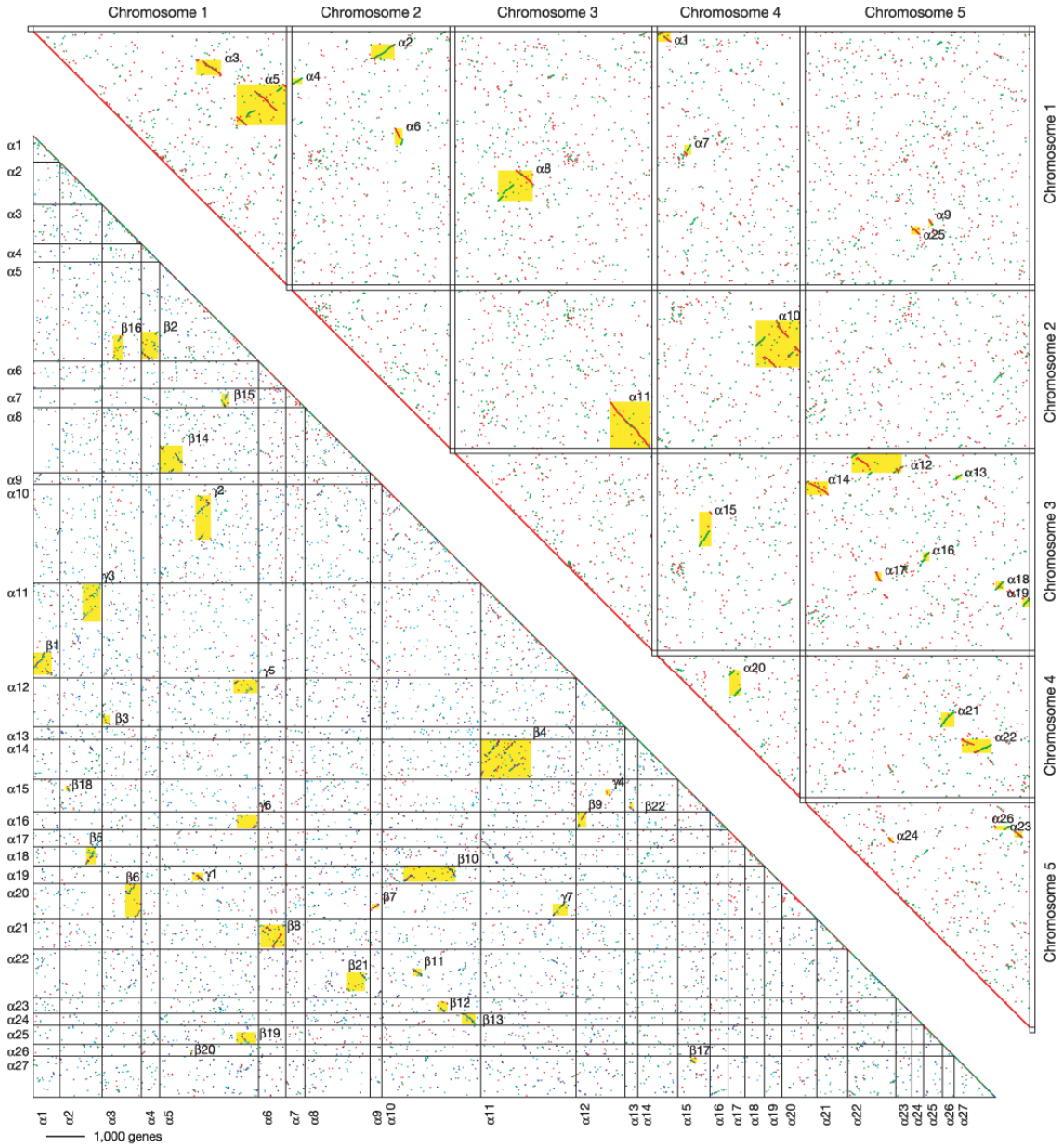


Figure 1 Arrangement of duplicated protein-encoding genes in *Arabidopsis thaliana*. The top right (*x* and *y* axes represent 26,028 genes in their chromosomal order. The best-matching gene pairs are plotted, colour-coded to indicate same (red) or opposite (green) transcriptional orientations. For further analysis, 57 adjacent duplicated regions with opposite orientation and order explicable by localized inversions were combined into 26 'large' duplications (α1–α26) that each included ≥1% (260) of the genes. Eight shorter duplications were also detected (lower left, and duplications. Both *x* and *y* axes represent 21,749 genes, in an inferred ancestral order that accounts for the duplications (see text) are highlighted. Colours show how the four ancestral chromosome segments contribute to the large duplications, distinguishing contributions to the 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).

Progenitor *Arabidopsis* genome

- How it was modified by the duplication event
- Compare to species that is evolutionary close.
  - *A. lyrata*
    - 8 chromosomes
  - *A. thaliana*
    - 5 chromosomes
- Genetic maps developed using shared loci were

Fig. 2

- Five *A. thaliana* chromosomes
  - Constructed from ancestral genome with eight chromosomes
- At Chr 1
  - Blocks of AlyLG1 + AlyLG2
- At Chr II
  - Blocks of AlyLG3 + AlyLG4.
- Conclusion
  - ***Two species with different chromosome numbers consist of the same chromosomal blocks***

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

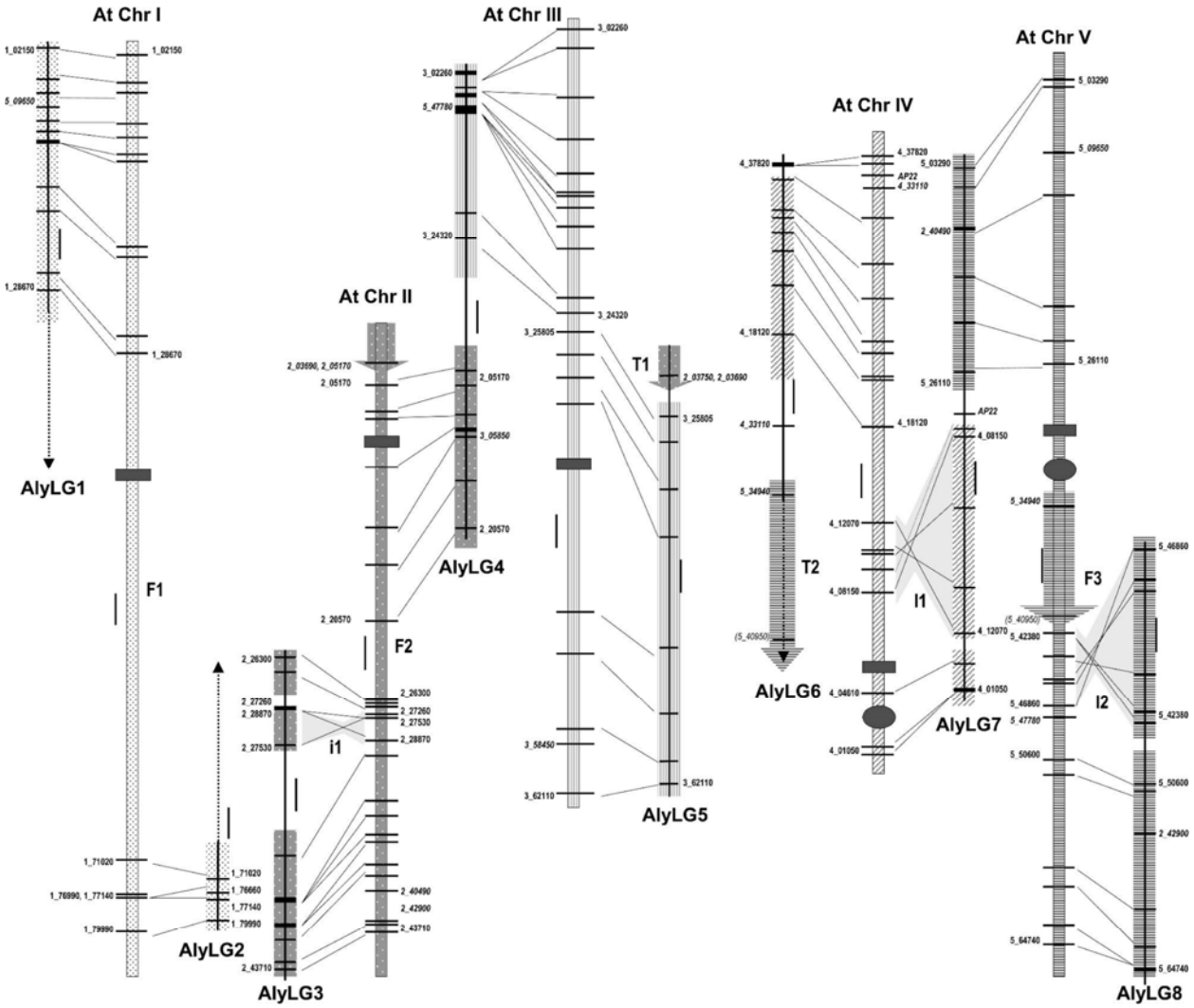


Figure 2. Colinearity of A. l. lyrata linkage map with the A. thaliana genome. A. thaliana chromosomes (At Chr I –V) are represented as patterned bars (drawn to scale, 1 unit = 1 Mbp; gray rectangles, centromeres; gray circles , heterochromatic knobs). A. l. lyrata linkage groups (Aly LG 1 –8) are shown 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.

Fig. 1B – Early duplication events

- Shows evidence of more ancient duplications
  - 27 α duplications reoriented
    - Notice block α5
    - Two duplicates blocks in the same order
    - Two in an opposite orientation
      - Presumed ancestral order derived from these four blocks
    - Same procedure that uncovered the α blocks.
      - Two types of blocks discovered.
        - 22 β blocks
          - ***Another duplication event in the A. thaliana lineage***

The 7 γ blocks

- Controversial
  - Hypothesis 1
    - Early duplication in the angiosperm lineage
  - Hypothesis 2
    - Duplication after the split of monocots and dicots
- Grapevine genome sequenced
  - Evidence from the genome appears to have resolved this question
    - Grape
      - Ancestor of the rosids
        - Group of species included *A. thaliana*.
    - Blast and dot blot analysis of grape genome

Figure 3

- Any genes shared with two other regions of the genome
  - Grape genome has a hexaploid history
- How about other species
  - Signal of hexaploidy is detected
    - **Figure 4**
      - Grape and poplar genomes were compared
      - Only triplicated regions in grape used
        - Triplicated regions
          - Two copies in poplar
  - ***Hexaploid ancestry concept is supported***
  - ***Poplar under went an additional WGD after its divergence from the grape lineage***

Shared duplications in dicot and monocot analysed

- Grape and rice orthologs analyzed
  - Hypothesis 1
    - Rice shared the hexaploid ancestry
      - 3-to-3 relationship
        - Not observed
  - Hypothesis
    - Rice does not share the same hexaploid ancestry
      - 3-to-1 relationship observed
  - Conclusion
    - ***Monocots and dicots do not share the same hexaploid history.***

(**Note**: See Tang et al. 2008. Genome Research18:1944 for an alternative perspective.)

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

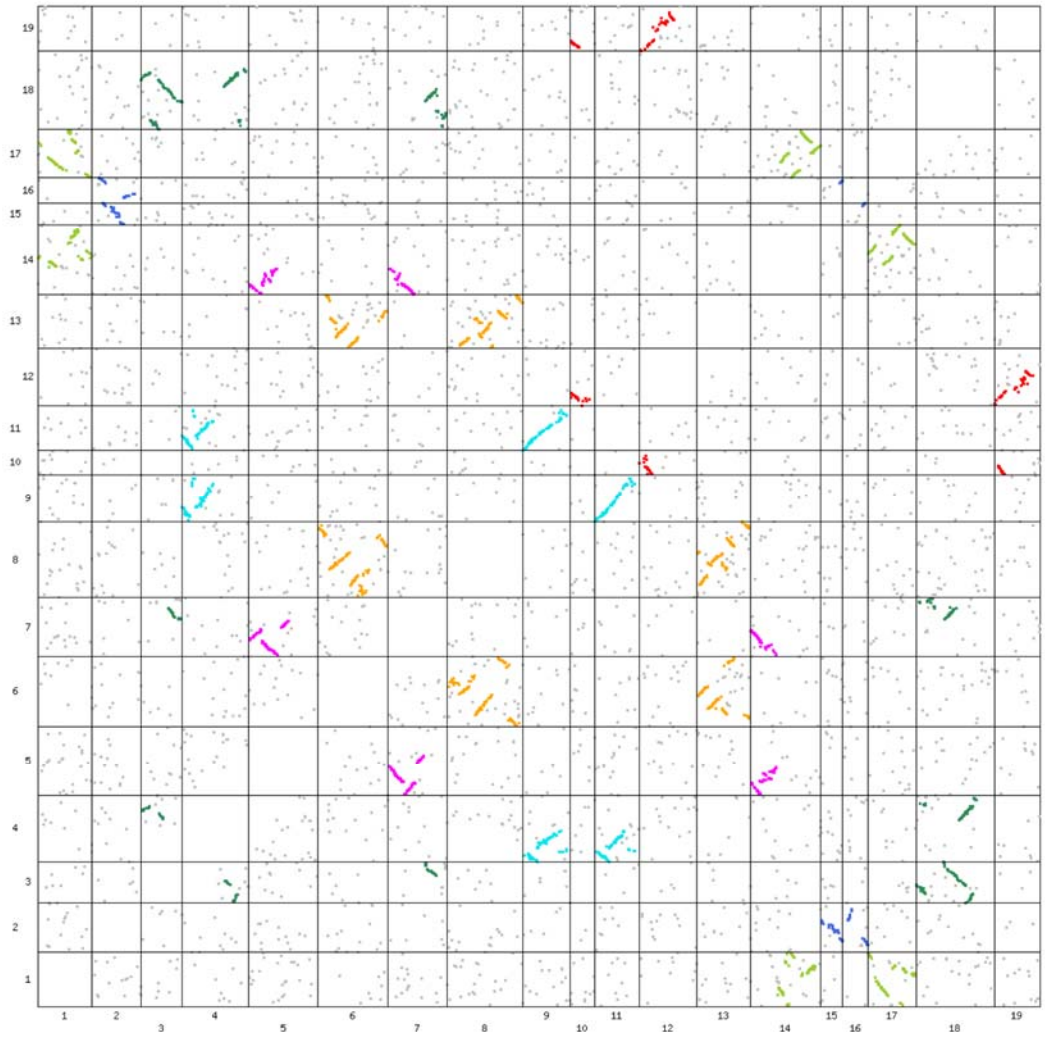


Figure S5. The grape genome originated from a polyploidy event that joined three ancestral genomes. The nineteen chromosomes of grape are represented on both the x and y axis. Dots represent the positions of paralogous pairs of genes. For clarity, intrachromosomal paralogs are not shown. Clusters of 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)

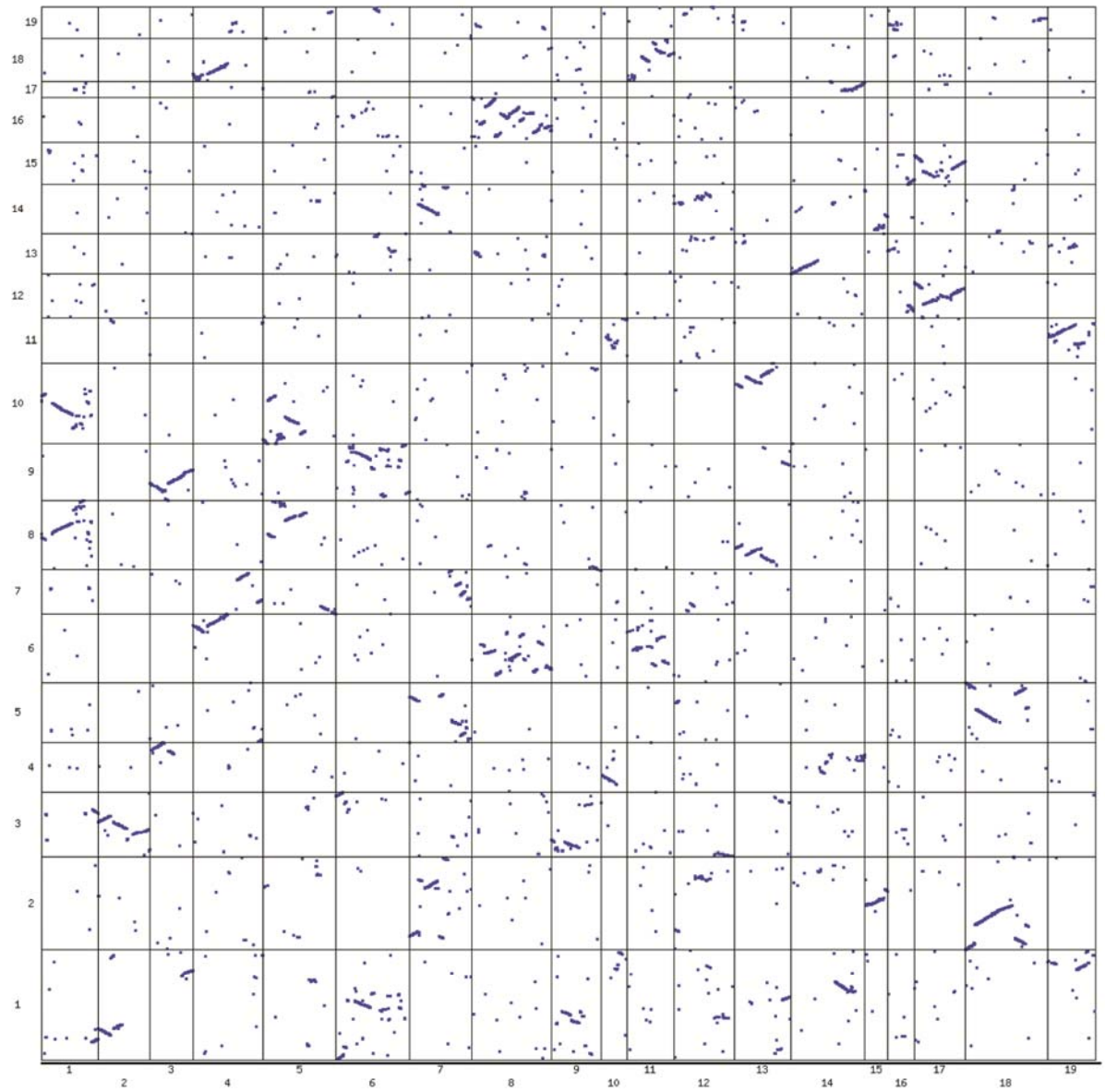


Figure S6. The distribution of 8,604 orthologous genes between *Vitis vinifera* (x axis) and *Populus trichocarpa* (y axis) chromosomes.

Summary of Eudicot Evolution

- Two diploid mate
  - Tetraploid species developed
- Tetraploid species mated to another diploid
  - Produce the ancestral hexaploid
    - All subsequent eudicots derived from this ancestor
      - ***Signatures of the same duplications***
        - ***Should be observed in their genome history***

Monocot genome evolution.

- Monocots also have a duplication history.
  - **Figure 5**
    - Compared rice and maize.
      - Maize chromosomes (y-axis) as the reference
        - Most rice genes found in two copies
      - Rice chromosomes (x-axis) as the reference
        - Blocks found three or four times in maize.
    - Conclusion
      - ***WGD event in the history of monocots***
      - ***An additional duplication occurred in the maize lineage.***

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

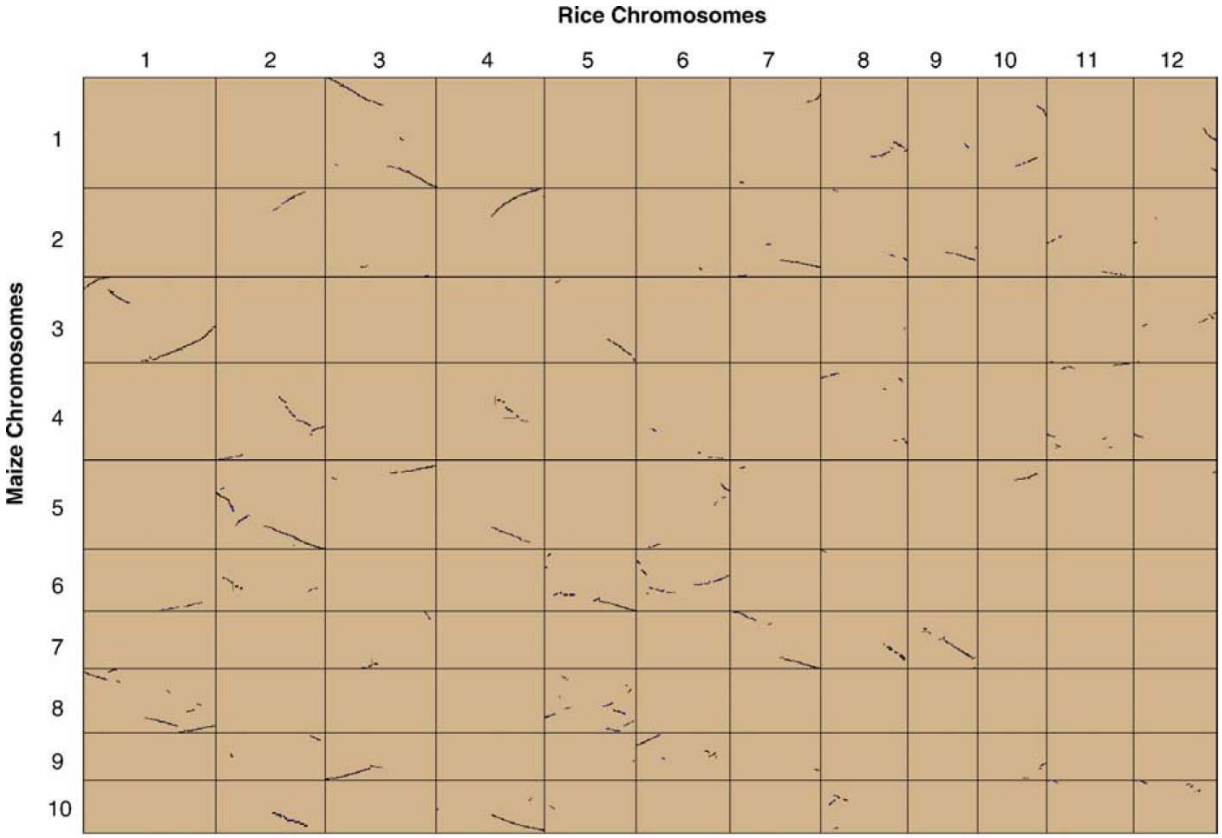


Figure 1. Dotplot Analysis of the Integrated Maize Map against Rice Pseudomolecules

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

Unified model of grass evolution – developing the ancestor

- Based on sequences of genome sequences of
  - Rice
  - Sorghum
  - Brachypodium (a model grass species)
  - Maize
- 56-73 MYA
  - Ancestral grass species containing five chromosomes
    - Duplicated
    - Genome with ten chromosomes appeared
  - Then
    - A4 and A6 fractionated
      - Chromosomes A4, A6, and A2 appear
    - A7 and A10 fractionated
      - Chromosomes A7, A10, and A3 appear
  - Paleopolyploid developed
    - 12 chromosomes
      - ***Progenitor of all of the modern grasses***

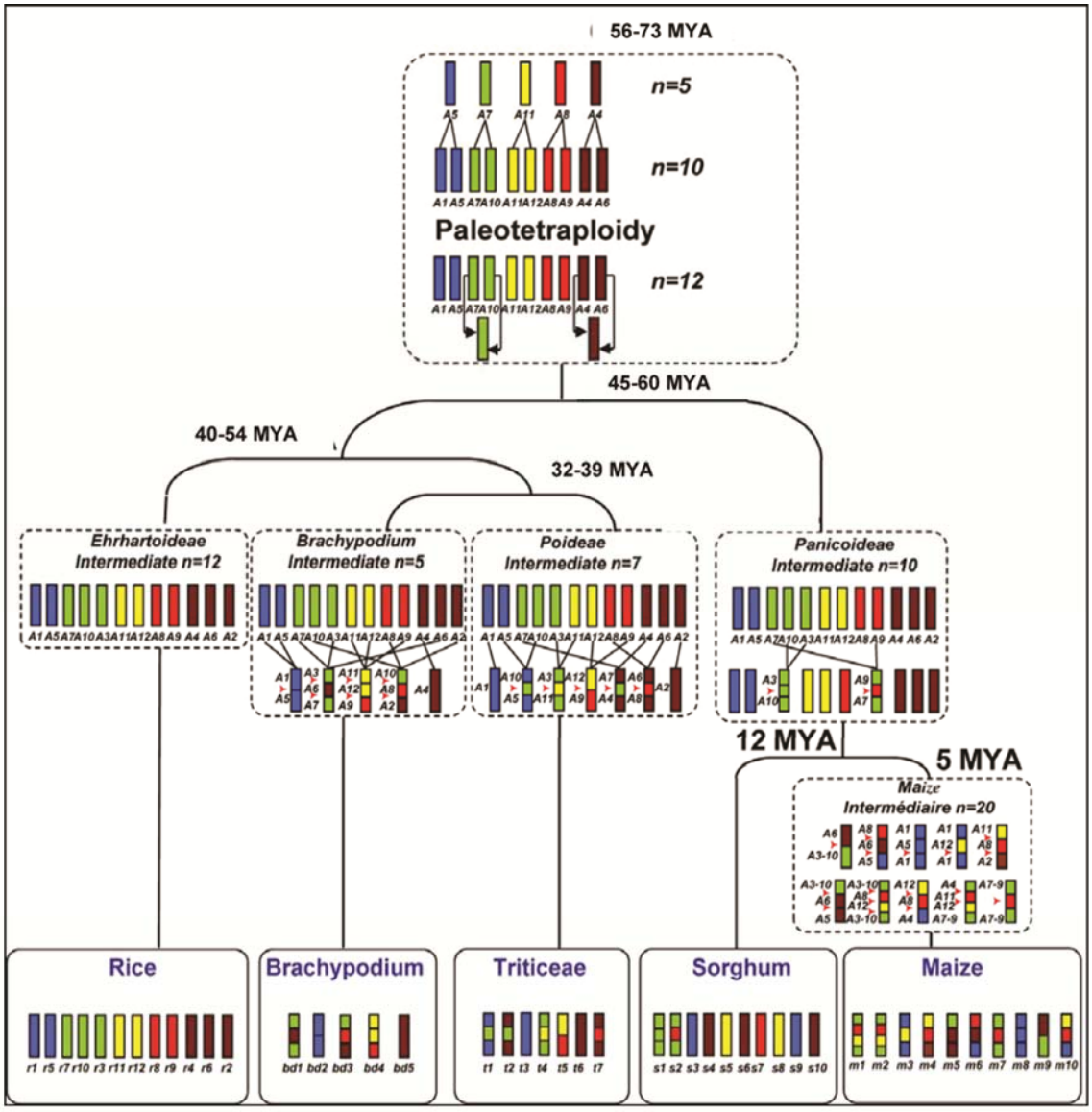
Unified model of grass evolution – developing the lineages

- Rice genome structure
  - Represents the ancient paleotetraploid.
    - Basic set of chromosomes
      - Building blocks for other genomes

Figure 6

- Breakage/translocation/fusion events
  - Involve chromosomal fragments from the n=12 ancestor.
    - Developed
      - Brachypodium
      - Poideae (representing the wheat lineage)
      - Panicoideae (representing the maize/sorghum lineage)
      - Panicoideae
        - Simplest history
        - Arose from only four breaks
      - Other lineages
        - More complex patterns of evolution
          - Maize genome
            - Underwent additional duplication
            - Additional breakage/translocation/fusion events
            - ***Constructed the modern maize chromosomes***

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.

Summary

- Plant genomes
  - A long history of genome duplications
    - Unlike animal and fungal genoemes,
- **Figure 7**
  - Illustrates the duplication history
    - (The γ event should be moved to the origin of the eudicot lineage.)
  - Significant role of WGD in development of plant species
    - Many duplications appear 55-70 MYA
      - Transition point
        - Cretaceous and Tertiary periods
          - Mass extinction of species
      - Hypothesis
        - Duplications gave plants the needed gene repertoire
          - ***To survive this extinction***
          - ***Flourish on earth***

(see Fawcett et al. 2009. PNAS USA 106:5737)

- **Figure 8**
  - Additional species were analyzed
  - Extended the analysis to deeper phylogeny
  - Additional duplication events determined
    - Ancestral seed plants
      - ζ at ~330 MYA
    - Ancestral angiosperms
      - ε at ~220 MYA

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

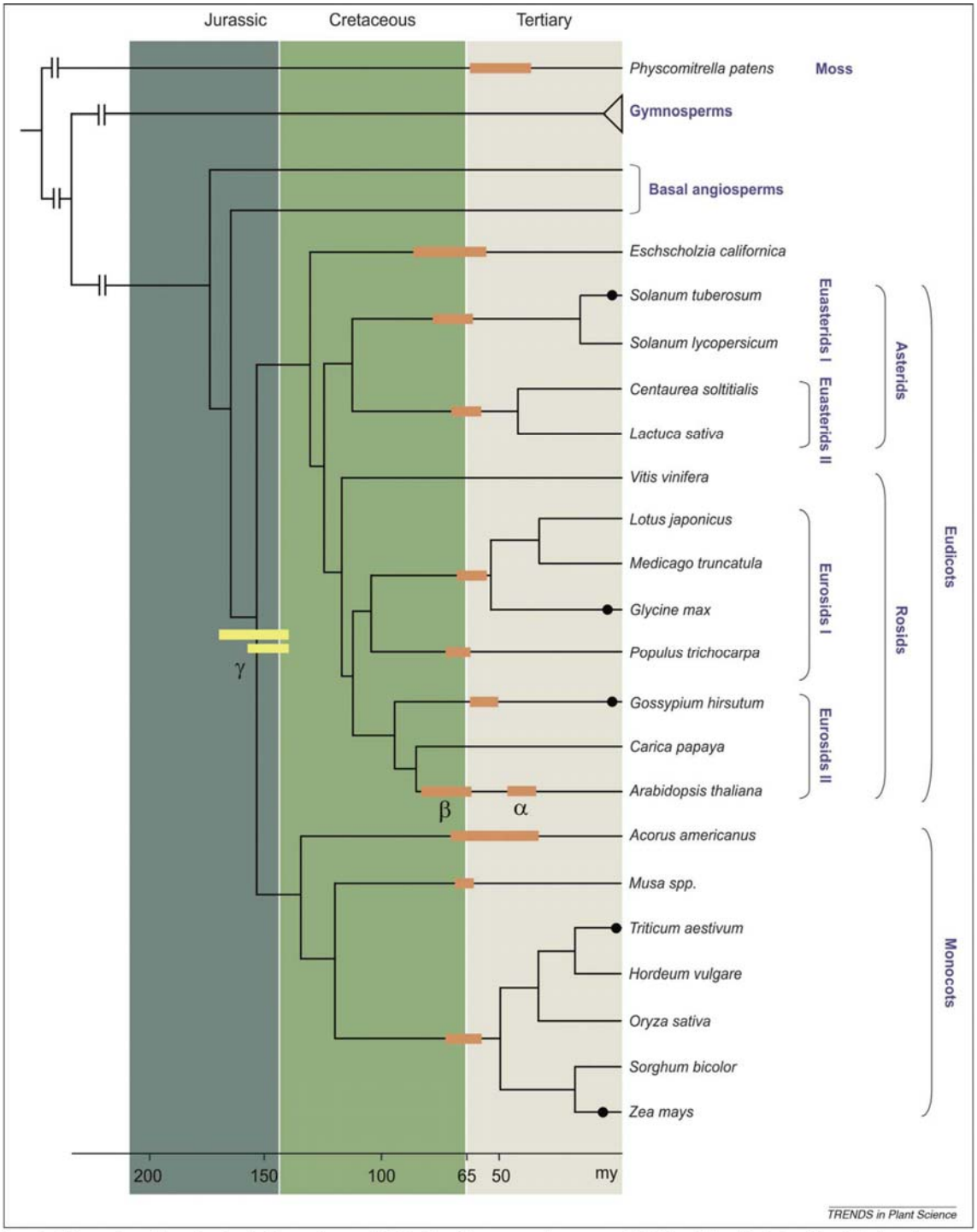
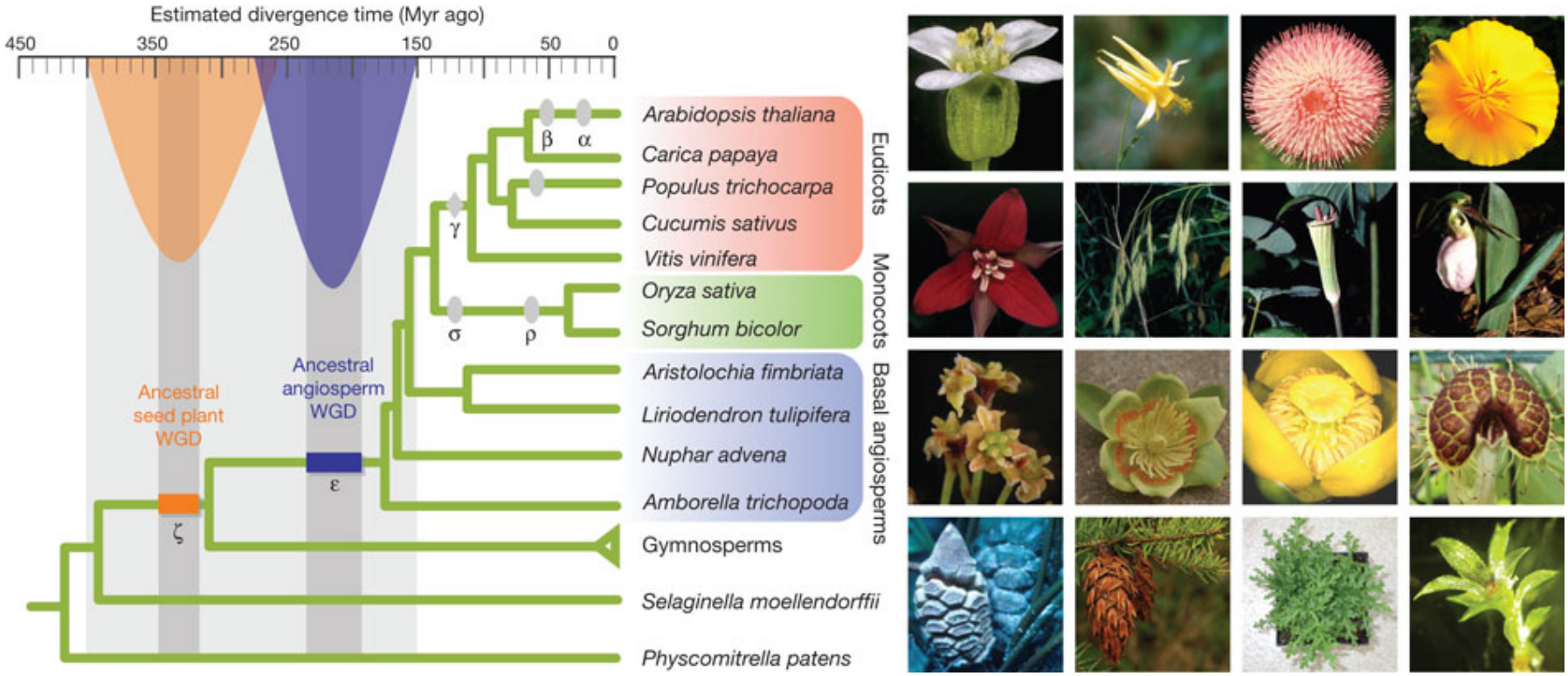


Figure 2 . Phylogenetic tree offl owering plants (eudicots and monocots). WGDs, inferred from recent studies [28–30], are indicated by horizontal bars. Yellow bars denote the hexaploidy event. More recent WGDs appear to be clustered around the KT boundary [29]. The black dots indicate recent polyploidy 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]

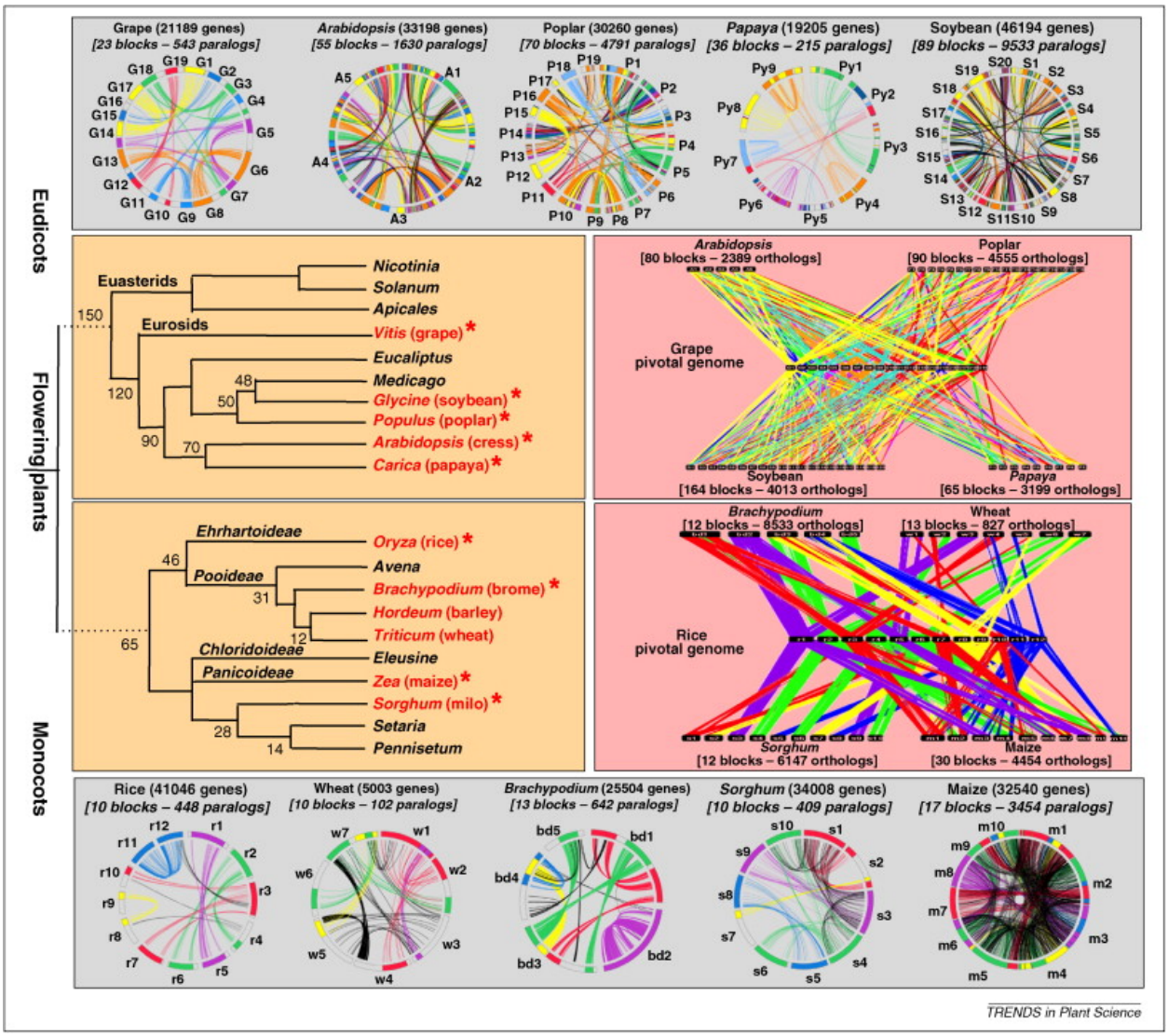


Original figure legend from manuscript. Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see text). The diamond refers to the triplication 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.

Dicot Paleohistory
[From: Trends in Plant Science (2010) 15:479; Nature Genetics (2011) 43:101]

- A duplication history is common in both dicots and monocots
    - Revealed by comparisons within different species

- Rice used as a reference for monocots
- Grape used as a reference for dicots

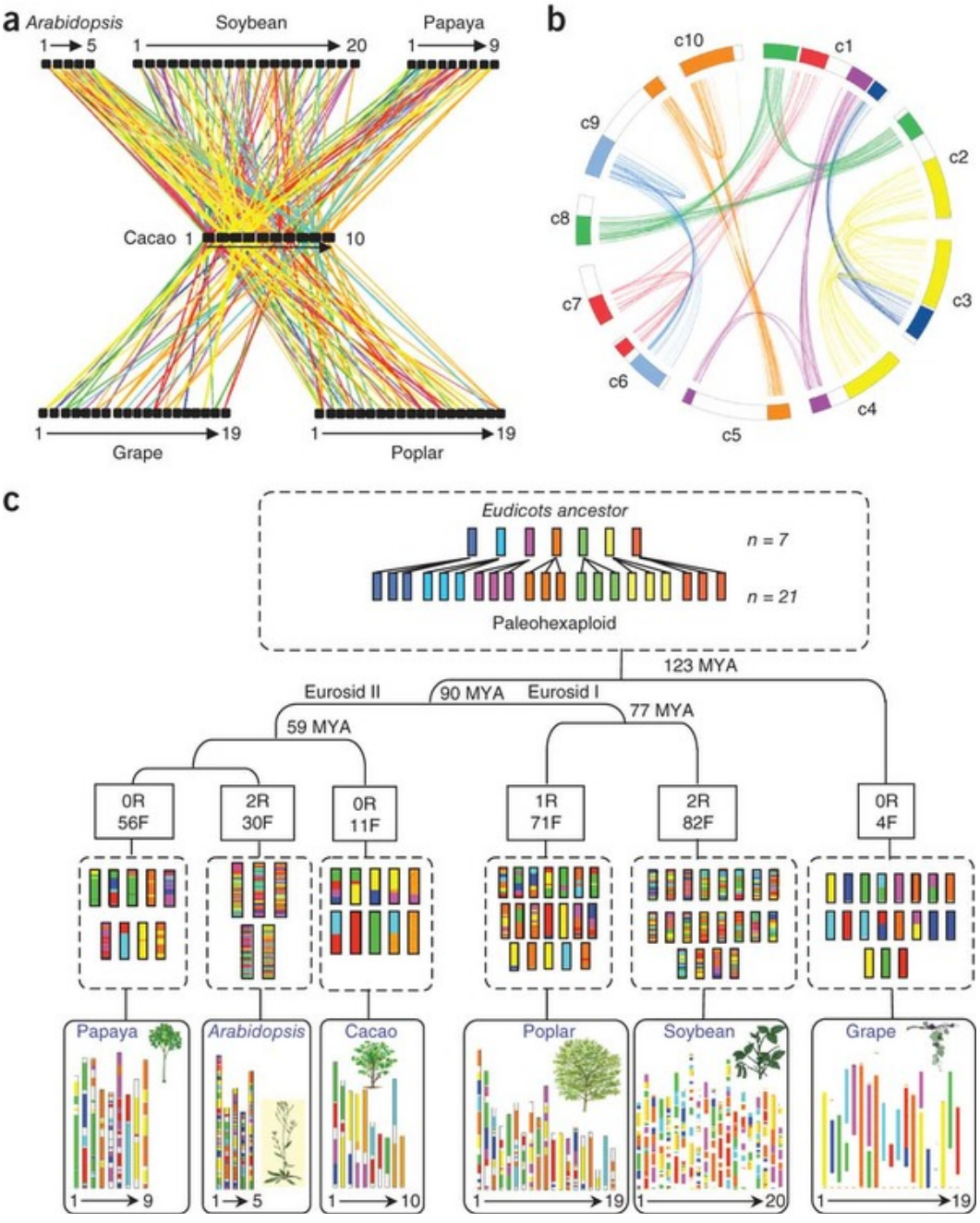


The ancestral dicot contained seven chromosomes

- The ancestor underwent a duplication to produce a paleohexaploid ancestor of modern dicots

Subsequent events

- Whole genome duplications within some lineages
- Breakage, fusion to generate new chromosomes in all lineages



The Gene-based Evolution of Duplicated Genes

If duplications are a major signature of plant genomes

- Copy number of genes should equal the number of rounds of duplication.

Table 1

- Number of genes found within plant species
    - Complete genome sequence
        - If the hexoploidy concept is true for dicots, and
        - Grape only contains this hexaploid event
            - Estimate
                - Ancestral dicot contains ~10,000 genes (=30,000/30).

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

Similarily

- Poplar underwent an additional duplication,
    - Theoretically # of genes = 60,000 genes
- *A.thaliana* underwent two duplications
    - Theoretically # of genes = 120,000 genes
- ***Not observed***

Monocot calculations

- Rice, Brachypodium, and sorghum only contain a duplication event
    - Number of ancestral monocot genes
        - 15,000 (=30,000/2).
    - Maize
        - Additional duplication event
            - But has undergone a reduction to ~30,000 genes
- Conclusion
    - Necessary to reduce the number of genes to ensure the success of the species.

Diploidization.

The polploid past history of plants

- Surprising result for Arabidopsis and rice genomes
    - Why??
        - Selected for sequencing because of their small genome sizes

Consequences of polyploidy?

- Doubling or tripling of the number of chromosomes
    - Evident for monocots.

Fate of the additional gene set from the WGD

- Concept
    - Species cannot maintain the entire set of duplicate chromosomes
    - New genes a problem
        - Generate deleterious mutations
        - Compromises the fitness of a genome
    - Genome must transition back to its original state.
        - Process is called
            - ***Diploidization***.

To revert back to the diploid state

- Many duplicate genes must be eliminated from the gene set
    - But a recently duplicated genome
        - Soybean
            - Withstands the extra copies
            - Genome about 2X the basic set of 30,000 genes of hexoploid ancestral eudioct

Events associated with diploidization

- Duplicate genome must change its chromosome pairing pattern
    - After the duplications,
        - Four chromosomes pair
        - Form quadravalents
    - Chromosomal structure must be changed so
        - Bivalents must be formed
            - Result
                - Doubling of the chromosome number
                    - Seen for the monocot lineage
            - Once bivalents are formed
                - Gene sets can evolve
                    - Processes
                        - Deletions and chromosomal rearrangements

Duplicate genes can undergo specific changes

- Common fate
    - Gene death of new copies
        - Loses associated with
            - Chromosomal breakage
            - Rearrangements.
        - Result
            - New basic set of chromosomes and genes will have appeared

Duplicate genes fate differs

- Some are retained as multicopy
    - Up to the ploidy level for that species
- Other reduced to only a single copy

“Deletion resistant” genes

- Not reduced to single copy
    - Dosage dependent
        - Mainly encode
            - Transcription factors
        - May lead to
            - Complex morphologies

“Duplication resistant” genes

- Must be maintained as single copy
    - Mainly encode
        - Enzymes or genes of unknown function

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

Developing new functions

Duplicate set of genes cannot be maintained

- Deleterious mutations can arise
- Duplicate genes are modified
    - Changes will provide
        - New functions
        - Altered altered functions
    - New functions may lead to the evolution of the species
        - Higher level of fitness
        - Evolutionary modifications of duplicate genes

Neofunctionalization.

- One duplicate gene maintains its original function
- Second gene evolves a function
    - May increase the adaptability of an individual

Subfunctionalization

- Modifies the duplicates
- Basic structure of both copies altered
    - Expression pattern of the gene changes
        - Results in a higher level of the protein production
- Alternately, the function of the original gene is maintained
    - Structure of both copies is significantly changed.
        - New copies retains
            - Part of the original function
        - Two genes work together
            - Function of the original gene maintained

Synteny: The Result of WGD and Reconstructing Plant Genomes

Synteny among plant species.

- Major result of the duplication history
    - Synteny
        - Maintenance of gene order between two species
    - Classic approach to synteny
        - Based on shared markers mapped onto two different species.
            - Macrosynteny is detected by
                - Large scale chromosomal blocks shared by two species.

Fig. 9

- Example of macrosynteny
    - Tomato and eggplant
        - Eggplant linkage group 4
            - Evolutionarily related to tomato
        - Linkage groups 10S and 4L.
            - Highly conserved marker order over many centimorgans of the two genomes

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

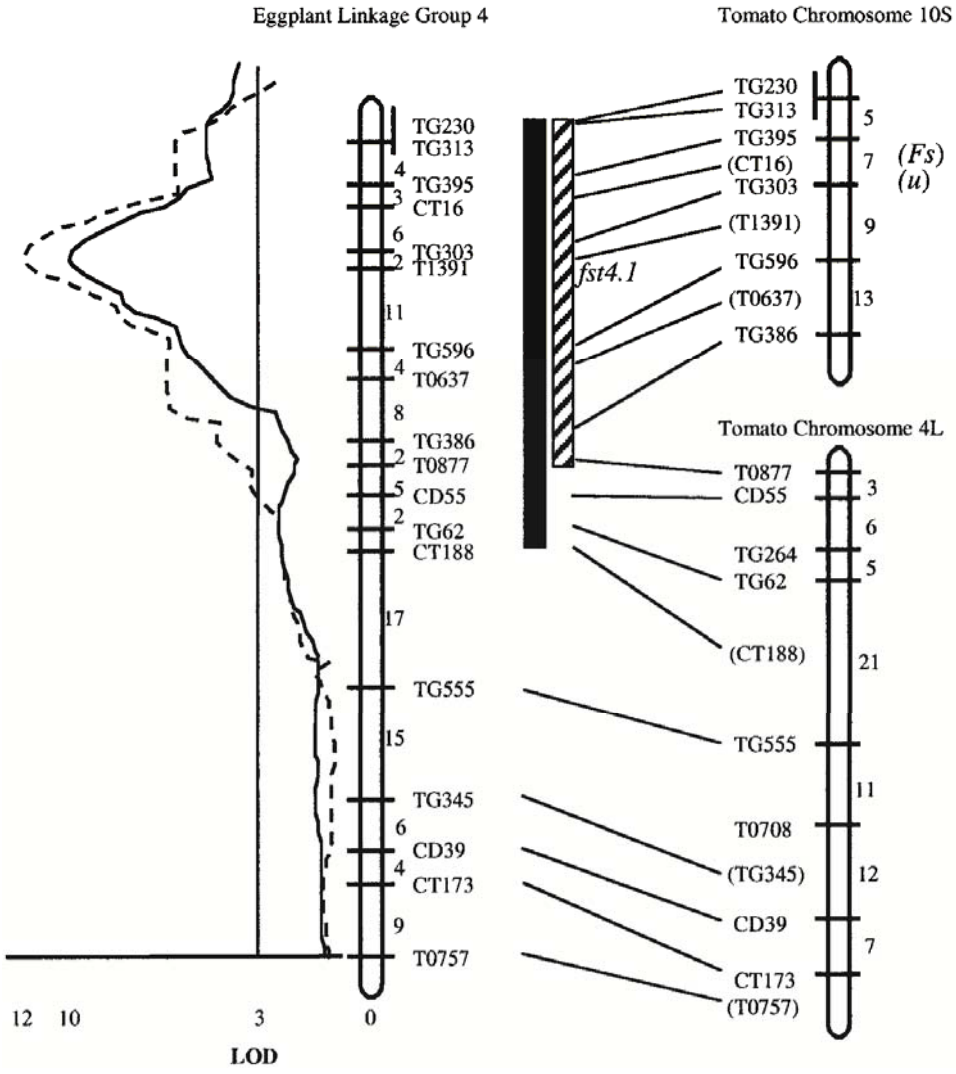


Figure 3.—Comparative mapping off ruit stripe locus on eggplant linkage group 4. Simple interval analysis for fst4.1 is shown to the left of the molecular map of eggplant linkage group 4 (solid line for NY data, dashed line for FR data). Bars to the right of the linkage group represent the position of 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).

Genetic mapping of shared genes

- First method of comparing species
- Only way to compare species that have not been sequenced
- Many examples of synteny mapping in plants.
- The power of synteny mapping
    - Discovery of shared loci from two species
        - Control the same phenotype
            - Map to the same genetic location.

Fig. 9 again

- Major QTL for fruit striping
    - Eggplant linkage 4.
    - Previous work with tomato
        - Major QTL
            - Linkage group 10 of tomato
    - Syntenic marker and QTL observed here
- Hypothesis
    - Multiple loci are shared in the same macrosyntenic order
        - Same ancestral gene is controlling this trait in these two species.

Leveraging knowledge in one species for gene discovery in a second species

- Phenotypic traits mapped extensively in one species
    - Points a researcher working on a second species
    - Likely location of a similar gene in second species.
    - Leverage is
        - Great aid for genetic discovery
            - For species in where the discovery of important genetic factors are limited by a lack of funding