Title: Epistatic interactions drive biased gene retention in the face of massive nuclear introgression

Authors: Evan S. Forsythe[1], Andrew D. L. Nelson[1], Mark A. Beilstein[1*]

Affiliations:

[1]School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA.

Abstract: Phylogenomic analyses are recovering previously hidden histories of hybridization, revealing the genomic consequences of these events on the architecture of extant genomes. We exploit a suite of genomic resources to show that introgressive hybridization occurred between close relatives of Arabidopsis, impacting our understanding of species relationships in the group. We show that cytonuclear discordance arose via massive nuclear, rather than cytoplasmic, introgression. We develop a divergence-based test to distinguish donor from recipient lineages and find that selection against epistatic incompatibilities acted to preserve alleles of the recipient lineage, while neutral processes also contributed to genome composition through the retention of ancient haplotype blocks.

One Sentence Summary: Introgression of donor and retention of recipient alleles is governed in part by selection for cytonuclear compatability.

Main Text:

Hybridization is a driving force in plant evolution (*1*), occurring naturally in ~10% of all plants, including 22 of the world's 25 most important crops (*2*). Botanists have long realized that through backcrossing to parents, hybrids can serve as bridges for the transfer of genes between species, a process known as introgression (IG). As more genome sequences become available, comparative analyses have revealed the watermarks of historical IG events in plant and animal genomes (*3–5*). Cytonuclear discordance is a hallmark of many IG events, occurring, in part, because nuclear and cytoplasmic DNA differ in their mode of inheritance. In plants, this discord is often referred to as "chloroplast capture," which has been observed in cases where IG of the chloroplast genome occurs in the near absence of nuclear IG or via nuclear IG to a maternal recipient (*6*). Disentangling IG from speciation is particularly important because IG may facilitate the transfer of adaptive traits that impact the evolution of the recipient lineage (*4*, *6*). Moreover, unlinked nuclear and cytoplasmic IG creates an interaction interface for independently evolving nuclear and cytoplasmic alleles, either of which may have accumulated mutations that result in incompatibilities with deleterious effects when they are united in hybrids. Such incompatibilities could exert a selective pressure that influences which hybrid genotypes are permissible thereby favoring the co-introgression of alleles for interacting genes (*7*). Here, we exploit a suite of genomic resources to explore a chloroplast capture event involving Arabidopsis and its closest relatives. We document cytonuclear discordance and ask if it arose through IG of organelles or extensive IG of nuclear genes. Further, we develop a method to ask which lineage was the recipient of introgressed alleles. Finally, we explore the extent to which neutral processes, such as physical linkage as well as non-neutral processes, such as selection against incompatible alleles at interacting loci, shaped the recipient genome.

The wealth of genomic and functional data in Arabidopsis (*8*), combined with publicly available genome sequence for 26 species make the plant family Brassicaceae an ideal group for comparative genomics. Phylogeny of the group has been the focus of numerous studies (*9–15*), providing a robust estimate of its evolutionary history. While the genus *Arabidopsis* is well circumscribed (*12*, *16*), the identity of its closest relatives remains an open question. Phylogenetic studies to date recover three monophyletic groups: clade A, including the sequenced genomes of *A. thaliana* (*8*) and *A. lyrata* (*17*); clade B, including the *B. stricta* genome (*18*); and clade C, comprising the genomes of *Capsella rubella*, *C. grandiflora* (*19*), and *Camelina sativa* (*20*, *21*). Analyses using nuclear markers strongly support A(BC), which is most often cited as the species tree (*9*, *11*, *13–15*). Organellar markers strongly support B(AC) (*10*, *11*, *22*, *23*) (Fig. 1A and table S1). We explored the processes underlying this incongruence by inferring gene trees for markers in all three cellular genomes from six available whole genome sequences.

We searched for incongruent histories present within and among nuclear and organellar genomes in representative species from each clade. We included *Cardamine hirsuta* (*24*) and *Eutrema salsugineum* (*25*) as outgroups. We considered three processes capable of producing incongruent histories: duplication and loss, incomplete lineage sorting (ILS), and IG. To factor out the effects of gene duplication followed by differential paralog loss, we focused our analyses on single copy genes (figs. S1 and S2)(*21*). In the chloroplast, we found 32 single copy genes, while in mitochondria we identified eight. Maximum likelihood (ML) analyses of these yielded well-supported B(AC) trees (Fig. 1A and fig. S3). We identified 10,193 single copy nuclear genes spanning the eight chromosomes of *C. rubella* (fig. S2), whose karyotype serves as an estimate of the common ancestor (*26*). ML analyses of nuclear genes yielded 8,490 (87.6%) A(BC), 774 (8.0%) B(AC), and 429 (4.4%) C(AB) trees (Fig. 1, B-F, fig. S3, and table S2). Our analyses confirm the incongruent histories present in the organellar and nuclear genomes. To distinguish between IG and ILS, we applied the *D*-statistic (*5*, *27*) and show that ILS is sufficient to explain the frequency of C(AB) but not for the observed frequencies of A(BC) and B(AC) in the nuclear genome (table S3).

IG produces the observed patterns of unbalanced incongruent trees when it occurs between non-sister species subsequent to speciation (*4*, *5*, *27*). Thus, IG nodes are expected to be younger than speciation nodes (*28*, *29*), therefore we determined the relative timing of the B(AC) and A(BC) branching events (Fig. 2A) (*28*). More specifically, we calculated the depth of T_2, the node uniting clade A with clade C in B(AC) trees from the nucleus, and compared it with the depth of T_2, uniting the B and C clades in A(BC) trees (Fig. 2B). We first smoothed the rates of evolution across trees using a penalized likelihood approach (*30*) before calculating median T_2 node depths. T_2 for A(BC) was significantly shallower than T_2 for B(AC) (Fig. 2C, figs. S4 and S5, and table S4; $p<2.2e-16$, Wilcoxon), indicating that A(BC) trees are the product of IG rather than speciation. Hence, A and C diverged from each other prior to the exchange of genes between clade B and C via IG. This scenario stands in opposition to trees inferred from single or concatenated nuclear genes, which strongly favor A(BC) (*12*, *14–16*). However, it bolsters the argument that B(AC) best represents the species branching order despite the low frequency of these genes in the nucleus, and further suggests that the vast majority of nuclear genes in either B or C arrived there via IG.

To determine which of the two clade ancestors was the donor and which was the recipient of introgressed alleles, we developed a divergence based approach to infer the directionality of IG. First, we calculated rate of pairwise synonymous divergence (*dS*) for all pairs of species. We

used these to determine the average *dS* between pairs of clades (B vs. C = *S1*; A vs. C = *S2*; A vs. B = *S3*) (fig. S6). *dS* values are indicated as $S1\text{-}3_{SP}$ for B(AC) (the species branching order), and $S1\text{-}3_{IG}$ for A(BC) (IG branching order) (Fig. 2, D and E). We compared $S1_{IG}$, $S2_{IG}$, and $S3_{IG}$ to $S1_{SP}$, $S2_{SP}$, and $S3_{SP}$, respectively, to ask if divergence is consistent with IG from B to C (Fig. 2D) or from C to B (Fig. 2E). We found that $S1_{SP} > S1_{IG}$ (p<2.2e-16, Wilcoxon), $S2_{SP} < S2_{IG}$ (p=2.365e-12), and $S3_{SP} = S3_{IG}$ (p=0.1056), indicating IG from clade B to clade C. Since cytoplasmic inheritance is matrilineal in Brassicaceae, we conclude that clade C was the maternal recipient of paternal clade B nuclear alleles.

The IG that occurred during the evolution of clade C resulted in a genome in which the majority of maternal nuclear alleles were displaced by paternal alleles from clade B, while native organellar genomes were maintained. We asked whether we could detect patterns within the set of nuclear genes that were also maintained alongside organelles during IG. We hypothesized that during the period of exchange, selection would favor the retention of alleles that maintain cytonuclear interactions, especially when replacement with the paternal allele is deleterious (*7*). Using Arabidopsis Gene Ontology (GO) data (*31*), we asked if B(AC) nuclear genes were significantly enriched for chloroplast and mitochondrial-localized GO terms, indicating that these genes are more likely to be retained than are other maternal genes. We calculated enrichment (*E*) for each GO category by comparing the percentage of B(AC) nuclear genes with a given GO term to the percentage of A(BC) genes with that term (*21*). Positive *E* indicates enrichment among B(AC) genes; negative *E* indicates enrichment among A(BC) genes. B(AC) nuclear genes are significantly enriched for chloroplast (E=0.10, p=0.00443, 1-tail Fisher's) and mitochondrial localized (E=0.13, p=00250) GO terms (Fig. 3A and table S5). Enrichment was also detected at the level of organelle-localized processes such as photosynthesis (E=0.29, p=0.01184), including the light (E=0.44, p=0.00533) and dark (E=0.65, p=0.04469) reactions. Interestingly, pentatricopeptide repeat-containing (PPR) genes, a large family of nuclear-encoded genes known to form critical interactions in organelles that mediate cytoplasmic male sterility (*32*), are significantly enriched among B(AC) topology genes (E=0.21, p=0.01682). The opposite enrichment pattern exists for nuclear localized genes (E=-0.06, p=0.00936) (Fig. 3A). Thus, we find evidence that selection acted during IG, resulting in resistance of organelle interacting nuclear genes to replacement by paternal alleles. Maternal nuclear alleles that function in chloroplasts or mitochondria in fundamental processes were not replaced at the same rate as maternal alleles localized to other areas of the cell or for other functions. These genes may constitute a core set whose replacement by paternal alleles is deleterious.

We also asked if epistatic interactions between nuclear genes influenced the likelihood of replacement by paternal alleles. Using Arabidopsis protein-protein interaction data (*33*), we constructed an interaction network of our single copy nuclear genes (Fig 3B). To assess whether genes with shared history are clustered in the network, we calculated its assortativity coefficient (*A*) (*21*). We assessed significance by generating a null distribution for *A* using 10,000 networks with randomized topology assignments. In our empirical network, *A* was significantly positive (A=0.0885, p=0.00189, *Z*-test), and hence topologies are clustered (Fig. 3C), indicating that selection acted against genotypes containing interactions between maternal and paternal alleles. Thus, it appears that selection against epistatic conflicts also contributed to the composition of maternal alleles retained in the nucleus during IG.

While gene function and epistatic interactions exerted influence on nuclear IG, we also wondered whether blocks of genes with similar histories were physically clustered on chromosomes. We looked for evidence of haplotype blocks using the *C. rubella* genome map

(Fig. 3D). Previous studies in this group estimate linkage disequilibrium to decay within 10kb (*34*, *35*), creating blocks of paternal or maternal genes around that size. We assessed the physical clustering of genes with shared history by two measures: 1) number of instances in which genes with the same topology are located within 10kb of each other (fig. S7A), and 2) number of instances in which neighboring genes share topology, regardless of distance (fig. S7B). We compared both measures to a null distribution generated from 10,000 randomized chromosome maps. By both measures, we found significant clustering of A(BC) (measure 1: p=3.022e-8; measure 2: p=1.41364e-10, Z-test) and B(AC) (measure 1: p=0.003645; measure 2: p=1.7169e-11) genes (fig. S7, C-H). The observed clustering indicates that haplotype blocks of co-transferred and un-transferred genes are detectable in extant genomes, pointing to physical linkage as a factor influencing whether genes are transferred or retained.

In summary, our comparative genomic analyses revealed massive unidirectional nuclear IG driven by selection and influenced by linkage, thereby refining our understanding of the processes that can lead to an observation of "chloroplast capture." The species branching order in this group is more accurately reflected by B(AC), and thus similar to the findings of (*28*), nuclear IG obscured speciation such that the latter was only recoverable from extensive genomic data. What makes IG here particularly interesting is that its impact on the genome is evident despite the fact that it must have occurred prior to the radiation of clade A 13 – 9 million years ago (*12*, *14*). Hence, it's likely that, as additional high-quality genomes become available, comparative analyses will reveal histories that include nuclear IG, even when the genomes considered are more distantly related. We show that once identified these cases permit the development of analytical tools to infer the details of IG, such as its directionality. Our results from this test are consistent with massive unidirectional IG from clade B to C. During this onslaught, a core set of nuclear genes resisted displacement by exogenous alleles; purifying selection removed genotypes with chimeric epistatic combinations that were deleterious, just as Bateson-Dobzhansky-Muller first described (*7*, *36*). Will other IG events reveal similar selective constraints as those we detail? If so, it could point us toward key interactions between cytoplasmic and nuclear genomes that lead to successful IG, thereby refining our understanding of the factors governing the movement of genes among species.

References and Notes:

1. G. L. Stebbins, The Significance of Hybridization for Plant Taxonomy and Evolution. **18**, 26–35 (1968).
2. S. B. Yakimowski, L. H. Rieseberg, The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. *Am. J. Bot.* **101**, 1247–1258 (2014).
3. L. H. Rieseberg, J. Whitton, C. R. Linder, Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Bot. Neerl.* **45**, 243–262 (1996).
4. K. K. Dasmahapatra *et al.*, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*. **487**, 94–98 (2012).
5. R. E. Green *et al.*, A Draft Sequence of the Neandertal Genome. *Science (80-.).* **328**, 710–722 (2010).
6. L. H. Rieseberg, D. E. Soltis, Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. trends Plants*. **5**, 65–84 (1991).
7. D. B. Sloan, J. C. Havird, J. Sharbrough, The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Mol. Ecol.* **26**, 2212–2236 (2017).

8. P. Lamesch *et al.*, The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **40**, 1202–1210 (2012).
9. C. D. Bailey *et al.*, Toward a global phylogeny of the Brassicaceae. *Mol. Biol. Evol.* **23**, 2142–2160 (2006).
10. M. A. Beilstein, I. A. Al-Shehbaz, E. A. Kellogg, Brassicaceae phylogeny and trichome evolution. *Am. J. Bot.* **93**, 607–619 (2006).
11. M. A. Beilstein, I. A. Al-Shehbaz, S. Mathews, E. A. Kellogg, Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *Am. J. Bot.* **95**, 1307–27 (2008).
12. M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, S. Mathews, Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Pnas*. **107**, 18724–18728 (2010).
13. T. L. P. Couvreur *et al.*, Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* **27**, 55–71 (2010).
14. C.-H. Huang *et al.*, *Mol. Biol. Evol.*, in press, doi:10.1093/molbev/msv226.
15. R. K. Oyama *et al.*, The shrunken genome of Arabidopsis thaliana. *Plant Syst. Evol.* **273**, 257–271 (2008).
16. I. a. Al-Shehbaz, S. L. O'Kane, Taxonomy and Phylogeny of Arabidopsis (Brassicaceae). *Arab. B.* **6**, 1–22 (2002).
17. T. T. Hu *et al.*, The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. **43**, 476–481 (2011).
18. C.-R. Lee *et al.*, Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nat. Ecol. Evol.* **1**, 119 (2017).
19. T. Slotte *et al.*, The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–5 (2013).
20. S. Kagale *et al.*, The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* **5**, 3706 (2014).
21. Materials and Methods are available as supplementary materials at the Science website.
22. A. Franzke, D. German, I. A. Al-Shehbaz, K. Mummenhoff, Arabidopsis family ties: molecular phylogeny and age estimates in Brassicaceae. *Taxon*. **58**, 425–427 (2009).
23. M. Koch, B. Haubold, T. Mitchell-Olds, Molecular systematics of the brassicaceae: Evidence from coding plastidic matK and nuclear Chs sequences. *Am. J. Bot.* **88**, 534–544 (2001).
24. X. Gan *et al.*, The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nat. Plants*. **2**, 16167 (2016).
25. R. Yang *et al.*, The Reference Genome of the Halophytic Plant Eutrema salsugineum. *Front. Plant Sci.* **4**, 1–14 (2013).
26. M. E. Schranz, A. J. Windsor, B.-H. Song, A. Lawton-Rauh, T. Mitchell-Olds, Comparative genetic mapping in Boechera stricta, a close relative of Arabidopsis. *Plant Physiol.* **144**, 286–98 (2007).
27. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
28. M. C. Fontaine *et al.*, Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science (80-.).* **347**, 1258522–1258522 (2015).
29. B. K. Rosenzweig, J. B. Pease, N. J. Besansky, M. W. Hahn, Powerful methods for detecting introgressed regions from population genomic data. *Mol. Ecol.*, 2387–2397

(2016).
30. M. J. Sanderson, Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).
31. J. A. Blake *et al.*, Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
32. L. Gaborieau, G. G. Brown, H. Mireau, The Propensity of Pentatricopeptide Repeat Genes to Evolve into Restorers of Cytoplasmic Male Sterility. *Front. Plant Sci.* **7** (2016), doi:10.3389/fpls.2016.01816.
33. M. M. Brandão, L. L. Dantas, M. C. Silva-Filho, AtPIN: Arabidopsis thaliana protein interaction network. *BMC Bioinformatics*. **10**, 454 (2009).
34. B. H. Song *et al.*, Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in Boechera stricta, a wild relative of Arabidopsis. *Genetics*. **181**, 1021–1033 (2009).
35. S. Kim *et al.*, Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.* **39**, 1151–1155 (2007).
36. H. A. Orr, Dobzhansky, Bateson, and the genetics of speciation. *Genetics*. **144**, 1331–1335 (1996).
37. N. Merchant *et al.*, The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **14**, 1–9 (2016).
38. D. M. Goodstein *et al.*, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, 1178–1186 (2012).
39. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
40. J. L. Bowers, B. A. Chapman, J. Rong, A. H. Paterson, Unraveling angiosperms genome evolution by phylogenitc analysis of chromosomal duplications events. *Nature*. **422**, 433–438 (2003).
41. R. De Smet *et al.*, Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2898–903 (2013).
42. J. M. Duarte *et al.*, Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61 (2010).
43. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
44. Z. Zhang *et al.*, ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
45. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–3 (2014).
46. D. Kim *et al.*, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
47. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
48. M. Kearse *et al.*, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**, 1647–1649 (2012).
49. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals

unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
50. G. Vaidya, D. J. Lohman, R. Meier, SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*. **27**, 171–180 (2011).
51. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*. **20**, 289–290 (2004).
52. K. P. Schliep, phangorn: Phylogenetic analysis in R. *Bioinformatics*. **27**, 592–593 (2011).
53. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
54. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–91 (2007).
55. C. Lurin *et al.*, Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. *Plant Cell*. **16**, 2089–2103 (2004).
56. M. E. J. Newman, Mixing patterns in networks. *Phys. Rev. E*. **67**, 26126 (2003).
57. D. H. Phanstiel, A. P. Boyle, C. L. Araya, M. P. Snyder, Sushi.R: Flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics*. **30**, 2808–2810 (2014).
58. R. Gentleman *et al.*, Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
59. P. J. Alexander, M. D. Windham, R. Govindarajulu, I. a. Al-Shehbaz, C. D. Bailey, Molecular Phylogenetics and Taxonomy of the Genus Boechera and Related Genera (Brassicaceae: Boechereae). *Syst. Bot.* **35**, 559–577 (2010).
60. C. D. Bailey, I. A. Al-shehbaz, G. Rajanikanth, Generic Limits in Tribe Halimolobeae and Description of the New Genus Exhalimolobos (Brassicaceae). **32**, 140–156 (2017).
61. T. Slotte, A. Ceplitis, B. Neuffer, H. Hurka, M. Lascoux, Intrageneric phylogeny of Capsella (Brassicaceae) and the origin of the tetraploid C. bursa-pastoris based on chloroplast and nuclear DNA sequences. *Am. J. Bot.* **93**, 1714–1724 (2006).
62. I. Galasso, A. Manca, L. Braglia, E. Ponzoni, D. Breviario, Genomic fingerprinting of *Camelina* species using cTBP as molecular marker. *Am. J. Plant Sci.* **6**, 1184–1200 (2015).

Acknowledgments: The data reported in this paper are provided in the Supplementary Materials. Scripts used to perform analyses are available at: https://github.com/esforsythe/Brassicaceae_phylogenomics. This work was funded by NSF grants 1409251, 1444490, and 1546825 to MAB. We thank M. J. Sanderson, M. M. McMahon, E. Lyons, R. N. Gutenkunst, A. E. Baniaga, and S. M. Lambert for helpful discussions and M. T. Torabi and M. C. Borgstrom for statistical consultation. Finally, this work benefited greatly from input of the PaBeBaMo research group in the School of Plant Sciences, University of Arizona.

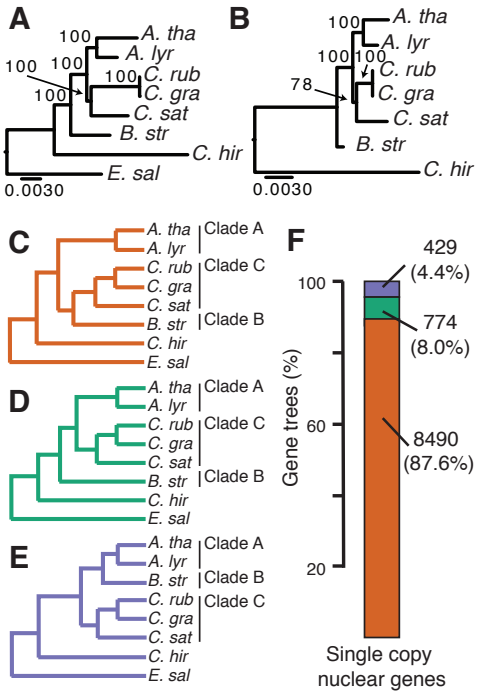

Fig 1. Incongruent gene tree topologies are observed within and between nuclear and organellar genomes. (A) Chloroplast and **(B)** mitochondria ML trees. Branch support from 100 bootstrap replicates. Scale bars represent mean substitutions/site. **(C-F)** ML gene trees inferred from nuclear single copy genes rooted by *E. salsugineum.* **(C)** A(BC), **(D)** B(AC) and **(E)** C(AB) topologies. **(F)** Numbers and frequencies of gene trees displaying A(BC)(orange), B(AC) (green), and C(AB) (purple).

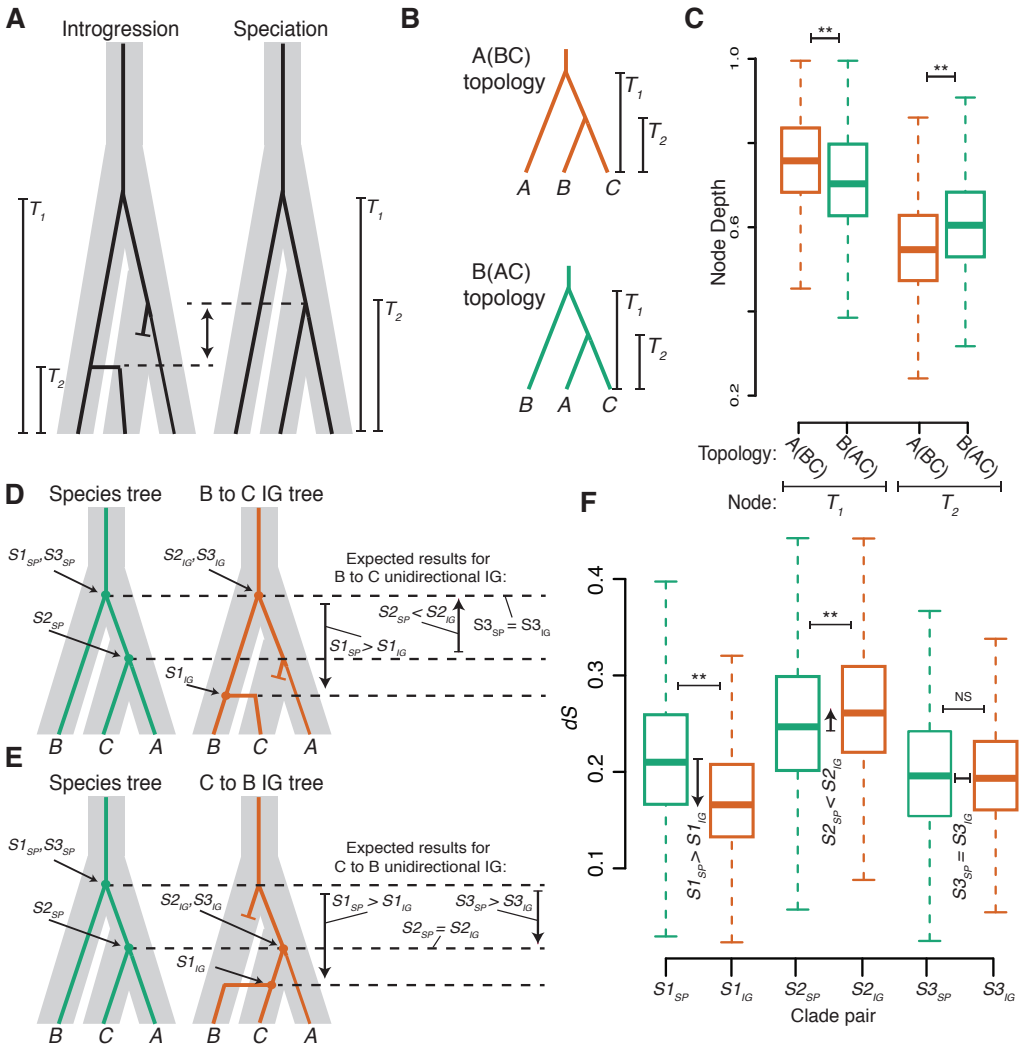

Fig. 2. Extensive introgression led to transfer of nuclear genes from clade B to clade C. (A) Model depicting expected T_1 and T_2 node depths for genes undergoing IG (left) or speciation (right). Speciation history is represented by thick grey bars. Individual gene histories are represented by black branches. Blunt ended branches represents native allele that was replaced by IG allele. Vertical arrow indicates expected difference in T_2 node depth. **(B)** T_1 and T_2 node depths on A(BC)and B(AC) trees. **(C)** Observed T_1 and T_2 *node depths*. **(D-E)** Model depicting pairwise *dS* distances between clades A, B, and C. Arrows indicate distances defined as *S1*, *S2*, and *S3* on the species tree (B(AC)) and the IG tree (A(BC)) indicated with *SP* and *IG* subscripts, respectively. Expected node depths under IG from clade B to clade C **(D)** or from clade C to B **(E)**. Vertical arrows depict expected differences between gene trees representing speciation and IG. **(F)** Observed *dS* distances on speciation gene trees (orange boxes; $S1_{SP}$, $S2_{SP}$, *and* $S3_{SP}$) and IG gene trees (green boxes; $S1_{IG}$, $S2_{IG}$, and $S3_{IG}$). Arrows indicate observed differences between *SP* and *IG* for *S1, S2,* and *S3* comparisons. Horizontal bars above boxes in **C** and **F** represent distribution comparisons. $^{**}p<0.01$, *NS* $p>0.05$.

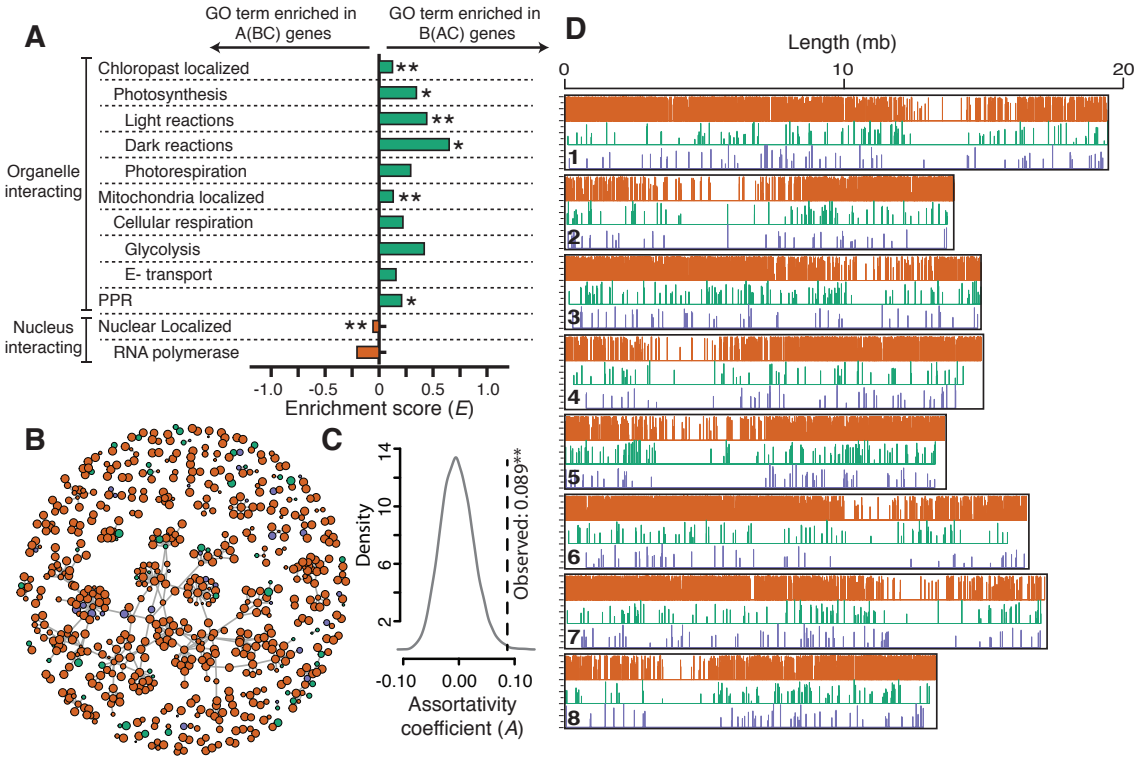

Fig. 3. The genomic consequences of epistasis and genetic linkage during IG.
(A) Enrichment (*E*) for GO terms = (% B(AC) genes – % A(BC) genes) / (% B(AC) + A(BC) genes). **(B)** Protein-protein interaction network for Arabidopsis protein complexes. Node fill, gene tree topology; node diameters proportional to bootstrap support (Fig. S3A-C). **(C)** Assortativity coefficient (*A*) of the network. Null distribution of *A* (grey curve); dotted line, observed *A*. **(D)** Nuclear genes mapped to *C. rubella*. Vertical lines, genes (colored by topology). Line heights proportional to bootstrap support (Fig. S3A-C). $^{**}p<0.01$, $^{*}p<0.05$, *NS* $p>0.05$.

Supplementary Materials:

Materials and Methods

Figures S1-S7

Tables S1-S5

Topology_results.xlsx

References (*1-62*)