The evolution of plant genomes — scaling up from a population perspective
Jonathan M Flowers and Michael D Purugganan

Plant genomes exhibit tremendous diversity in both their size and structure, with genome sizes across land plants ranging over two to three orders of magnitude and significant variation in structural organization was observed across species (EA Kellogg, JL Bennetzen, The evolution of nuclear genome structure in seed plants, Am J Bot 2004, 91:1709–1725). Five plant genomes are now either completely sequenced or in the draft stage; the grape (O Jaillon et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 2007, 449:463–467) and papaya (R Ming et al., The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus), Nature 2008, 452:991–997) whole genome sequences were reported most recently. Moreover, sequencing of 41 additional genomes is in progress. There is now an emerging consensus that understanding genome evolution requires consideration of the population genetics of genome diversification, and that description of evolutionary forces at the level of populations and within species can help identify the features that led to plant genome diversity (M Lynch, JS Conery, The origins of genome complexity, Science 2003, 302:1401–1404). In this review we focus on advances in our understanding of the mechanisms that drive the diversification of genomes. In particular, we look at the extent to which demographic features such as effective population size changes within species can drive genome evolution, discuss population genetic models of genome diversification associated with transposable element (TE) mobilization, and describe recent studies on the evolution of gene families.

Address
Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003-6688, United States

Corresponding author: Purugganan, Michael D (mp132@nyu.edu)

Population genetic models of genome diversification
Population genetic processes are governed to a large extent by a species’ effective population size (Ne), and there has been recent theoretical work that suggests that changes in Ne play a key role in the evolution of genome composition and size [1,2,3**]. Population size changes affect the dynamics of genome evolution by altering the efficacy of natural selection. Under this nearly neutral model, reduction in Ne increases the probability of fixation of deleterious mutations [4], and genome-wide diversification thus proceeds largely via the accumulation of nonadaptive molecular changes [1]. Smaller population sizes, for example, are predicted to facilitate the passive accumulation of deleterious genome features, such as TE insertions, or nonsynonymous mutations in gene-coding sequences.

The extent to which genome evolution is driven by processes that are dependent on effective population sizes is unknown. Certainly, several features of genome structure have been shown to be consistent with this model over broad evolutionary timescales, including genome size, intron lengths, and transposon copy numbers [1,2,3**]. Reduced single nucleotide polymorphism (SNP) levels and large excesses of slightly deleterious variation are also observed in selfing species [5*,6,7**], consistent with the lower effective population sizes of inbreeding species compared to outcrossing taxa [8].

Despite these observations, however, there have been questions raised about the significance of the nearly neutral model for genome evolution both on theoretical [9] and empirical grounds [2,10,11]. Population and comparative genomic studies of closely related plant species offer an opportunity to evaluate population genetic theories of genome diversification [2,12], by examining the evolutionary consequences of these effects by comparing plant species with different life histories or comparing patterns of evolution across recombination gradients within genomes [5*,13,14].

There have been several comparative genomic studies contrasting A. thaliana, an outcrossing species, and the self-compatible A. lyrata that possesses a lower Ne [2]. There is little evidence, however, that nearly neutral mutations have accumulated at different rates in the genomes of these two species. An increased rate of protein evolution and a decrease in codon usage bias is not observed in A. thaliana, despite predictions of the nearly neutral model for these patterns in inbred species [12]. Moreover, the patterns of amino acid polymorphism and between-species divergence are indistinguishable between A. thaliana and A. lyrata; this is again inconsistent with the nearly neutral theory [15*]. Similar results were
obtained for a wider array of species comparisons [5*], although in this larger study there is some evidence for greater excesses of replacement polymorphism in selfing lineages.

The failure of the Arabidopsis comparison to yield results consistent with population genetic predictions has been attributed to a recent origin of selfing, as has also been similarly suggested for self-fertilizing species C. elegans [16]. Alternatively, unless there is a large class of mutations that are additive in their fitness effects, expectations based on Ne may be obscured by the masking of deleterious recessives in heterozygous, out-crossing taxa [17]. Clearly, there is a need to extend the analysis to a much broader array of taxonomic comparisons. These initial studies do suggest, however, that the patterns of genome structure and variation observed at large evolutionary timescales may not necessarily be evident in microevolutionary comparisons within and between closely related species. Bridging these two levels remains an important challenge in comparative genome studies.

Transposable elements and genome size

TEs comprise the major fraction of repetitive DNA in eukaryotes and appear to be responsible, in large part, for differences in genome sizes among species [18]. Retrotransposon insertions, for example, have been shown to account for the phenomenon of hybrid genome expansion in interspecific Helianthus hybrids [19]. These hybrid species in this sunflower genus, which are not polyploid taxa, have up to 50% greater genome sizes than their progenitor species genomes, and Ty3/gypsy retro-element proliferation is responsible for between 60 and 80% of this size differential [19]. Genome size evolution in diploid and allopolyploid Gossypium species also appear to be governed by differences in genomic insertion/deletion rates, including those attributable to TEs [20].

In plants, many TE families have existed over long evolutionary periods and experience eruptions of activity followed by phases of dormancy. Sequencing of 74 randomly selected BAC clones in Zea mays, for example, revealed a preponderance of LTR retrotransposons, and two major expansions of copy number in the last two million years that are partly responsible for the large genome size of maize [21]. In Oryza australiensis, a relative of Asian cultivated rice, bursts of transposition activity in the last three million years among three retrotransposon families (RIRE1, Wallabi, and Kangourou) resulted in the accumulation of ~90,000 element copies and a rapid twofold increase in genome size [22]. One of these families, RIRE1, also shows evidence for interspecific transfer to O. australiensis from other reproductively isolated Oryza species [23], although greater sampling of elements among these species is necessary to confirm the possibility of horizontal gene transfer.

TEs also appear to affect rates of gene fission in plants. Although the rates of gene fusion are similar in O. sativa versus A. thaliana, the rice genome has higher rates of gene splitting, correlated with the longer gene sequences in the former and the greater abundance of TEs [24]. While TEs represent key components, however, other genome features in plants also appear to be factors in long-term trends in genome size changes. In broad comparisons between A. thaliana and O. sativa, there does appear to be a net loss of ~10–13 introns in plant genomes for every 1 intron gains, suggesting erosion in genome size (and gene complexity) over large evolutionary timescales [25].

Population genetic models, transposons, and genome diversity

Much of the effort aimed at understanding the effects of TE activity on plant genome size has focused on mechanisms of transposition rate changes and recombination. Significantly less attention has been paid to how natural selection and insertion biases may limit the proliferation of TEs in plant genomes [26]. In metazoans, population genetic evidence suggests purifying selection against slightly deleterious TE insertions may be important in limiting the spread of TEs [27–29], which has been interpreted as evidence of a population size effect on the efficacy of purifying selection [30]. Evidence of lower population frequencies of Ac-III elements in A. lyrata compared with a neutral distribution observed in A. thaliana is also consistent with this view [31]. Further support is provided by the observation that MITE elements in the nonrecombining Y chromosome of Silene latifolia also segregate at considerably higher frequencies than on recombining chromosomes [32*].

While these observations reveal broad comparative patterns, determining the mechanisms by which TE accumulation (accompanied by the associated genome diversification) proceeds requires explicit tests of alternative evolutionary models. Two leading models describe, in part, how purifying selection can determine TE copy numbers and distribution [33]. The gene disruption model states that transposons are preferentially excluded from regions with high gene density because of disruptions of gene structure and regulation. The ectopic exchange model, in contrast, proposes that purifying selection eliminates TEs because of deleterious rearrangements caused by homologous recombination between identical, dispersed TE copies.

It has been argued based on theoretical grounds that recombinational ectopic exchange may not be an important force in limiting the accumulation of TEs in selfing species such as A. thaliana [2,34], and indeed
the correlation of Arabidopsis TE abundance with gene density in noncoding DNA appears to support the alternative deleterious insertion model [35]. There is some evidence, however, for the ectopic recombination model in A. thaliana. Under this model, the frequency of ectopic recombination is expected to operate more strongly to remove TE insertions in regions of high recombination, on larger TE families, and longer polymorphic element copies [36]. These predictions appear to fit a recent population survey of Basho elements in A. thaliana. Hollister and Gaut observed that longer TE copies segregate at lower frequency and persist for shorter periods than smaller elements [37]. Basho copy number may be especially likely to be subject to ectopic exchange because of their abundance (2% of the A. thaliana genome), large size (~2 kb), and high sequence similarity because of recent expansion of the family. There have also been observations in O. sativa that short elements tend to accumulate in regions of high recombination and long elements are over-represented in regions of low recombination that may also be attributed to the operation of ectopic exchange [38,39].

These results illustrate the advantage of a population genomic approach in testing evolutionary hypotheses on the specific evolutionary forces that drive genome diversification. Further study to test explicit predictions of these and other models may allow us to determine the precise factors that contribute to TE patterns that are so crucial to the evolution of genome size and structure (Box 1).

**Gene duplication and polyploidy**

Gene duplication is generally regarded as a major force in the origin of new genes and genetic functions, and gene families account for more than half of the genes in plant genomes. Population genetic models have provided a framework to study the fates of duplicated genes, including the partitioning of ancestral functions (subfunctionalization) or the evolution of new functions (neofunctionalization) [40], and data continue to accumulate on the extent to which these can explain the rate, patterns, and fates of duplicate gene evolution. In a recent study of 280 paralogous gene pairs, 85% of duplicate pairs in A. thaliana showed a pattern of expression change consistent with both subfunctionalization and neofunctionalization, with 45% of gene duplicates evolving complementary expression patterns [41]. Neofunctionalization of duplicate genes also appears to result in novel gene organizations associated with metabolic function [42]. In both Arabidopsis and Avena sativa, it was shown that duplication followed by neofunctionalization and genome reorganization leads to the formation of operon-like gene clusters that contain coexpressed loci involved in triterpene biosynthesis [42].

There are now opportunities to examine the patterns of gene family evolution in the context of physical or genetic interaction networks associated with duplicated loci. This is clearly illustrated by a study of the MADS-box regulatory gene family, where phylogenetic analysis and interaction data from different plant taxa were investigated [43]. Although this gene family may have >110 members in extant angiosperms, these appear to have arisen from a set of 9–11 ancestral genes, and that rounds of whole genome duplications or polyploidization events have preserved the basic ancestral patterns of protein–protein interactions. In general, interactions between MADS-box proteins appear to be generally conserved across plant species, and the elaboration of this gene family through duplications is accompanied by increased interaction connectivity between subfamilies among MADS-box genes and shared interactors between gene duplicates [43].

This study of the MADS-box gene family and the role of whole genome duplication highlight continued interest in the fates of genes in polyploid genomes. One trend that is now emerging is that asymmetric gene retentions or losses between homologous chromosomes following allopolyploidization appear to occur, as has been documented in grass species [44]. This asymmetry has been hypothesized to arise from transcriptional dominance of one homologous genes over the others, a phenomenon that has been shown in a genome-wide scale in A. arenosa × A. thaliana hybrids [45]. There is also evidence for subfunctionalization of duplicated genes in polyploid Gossypium species [46]. Cotton arises as an allopolyploid between two species with divergent genomes (the A and D genomes). In an assay of nearly 1500 homoeologous genes in the domesticated polyploid taxa, 25–37% of expressed genes during seed trichome development shows biased gene

---

**Box 1 Population genetic concepts and definitions**

**Effective population size (Ne)** — the size of an idealized population that experiences genetic drift at the same rate as the population in question. Ne is affected by census size, level of inbreeding, population substructure, and other population and demographic attributes. Genomic regions can also differ in effective size because of the effects of background selection against deleterious mutations and selective sweeps of beneficial mutations localized at specific regions.

**Nearly neutral theory** — a model developed largely by Tomoko Ohta that proposes that a large fraction of mutations in the genome have a mildly deleterious or mildly advantageous effects on individual fitness. Slightly deleterious mutations are expected to segregate at low frequencies in populations, and to drift to fixation with higher probability in small populations owing to the effects of random genetic drift.

**Ectopic recombination** — crossing-over between dispersed regions of the genome that share sequence homology. The outcome of ectopic recombination may lead to chromosomal rearrangements.
expression of either the A or D homeolog locus, and there is evidence for subfunctionalization in these polyploid duplicate genes [46].

**Conclusion**

Genomic studies at various comparative levels — within populations, between closely related species and at higher taxonomic levels — have provided unprecedented insights into patterns of genome structure, and there are now concerted efforts to understand the evolutionary mechanisms that underlie this genome diversity (see Figure 1). Recent models to explain genome architecture appear to explain large-scale trends in plant genome evolution, but work in the last few years when comparing closely related species suggests that these models need to be extended or alternative models considered. Predictions on the effects of Ne on genome evolution, for example, are not met in comparative studies of closely related species with different life histories, both in plants and even animals [5]. Understanding the causal basis of genome evolution will require continued explicit tests of models of DNA sequence evolution, TE proliferation and loss, and gene duplication. Analyses of large evolutionary trends and at microevolutionary timescales provide promising avenues to understand how genomes evolve and the evolutionary forces that drive their diversification.

**Acknowledgements**

This work was funded in part by grants from the US National Science Foundation Plant Genome Research Program (DBI-0701382) and Advances in Bioinformatics Program (DBI-0820757).

**References and recommended reading**

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

- of special interest
- of outstanding interest


The first comprehensive survey of nucleotide variation across the entire genome in a plant population.


First study to compare polymorphism and divergence data for a large number of genes from two closely related plant species.


Evidence of how retrotransposition activity increases after interspecific hybridization, leading to enlarged genomes in hybrid species.


This study demonstrates that Y-linked MITEs segregate at substantially higher frequency than those found on autosomes consistent with a relaxation of purifying selection in nonrecombining regions.


A systematic analysis of the fates of duplicated genes in a plant genome, demonstrating the evolution of subfunctionalization and neofunctionalization.


A large-scale survey of the expression patterns of duplicate genes following polyploidy genome formation.