Genome size diversity in angiosperms and its influence on gene space

Steven Dodsworth¹, Andrew R Leitch¹ and Ilia J Leitch²

Genome size varies c. 2400-fold in angiosperms (flowering plants), although the range of genome size is skewed towards small genomes, with a mean genome size of 1C = 5.7 Gb. One of the most crucial factors governing genome size in angiosperms is the relative amount and activity of repetitive elements. Recently, there have been new insights into how these repeats, previously discarded as ‘junk’ DNA, can have a significant impact on gene space (i.e. the part of the genome comprising all the genes and gene-related DNA). Here we review these new findings and explore in what ways genome size itself plays a role in influencing how repeats impact genome dynamics and gene space, including gene expression.

Addresses
¹ School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK
² Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK

Corresponding author: Leitch, Ilia J (i.leitch@kew.org)

Introduction

Large-scale comparative analyses of plant genome sizes (GS) available in the Plant DNA C-values database (www.data.kew.org/evalues) have shown that angiosperms (flowering plants) are remarkable in their GS diversity. Not only do they have the largest range for any comparable eukaryotic group, varying c. 2400-fold (1C = 0.063–148.8 Gb), but they also include the largest eukaryotic genome so far recorded¹ [i.e. Paris japonica, 1] which is c. 950× larger than the genome of Arabidopsis thaliana (1C = 0.157 Gb). Nevertheless, the distribution of GS is strongly skewed towards small genomes, with the modal and mean values being just 1C = 0.6 Gb and 5.7 Gb, respectively [Figure 1]. There are two major drivers of this astonishing GS diversity: (i) polyploidy, or whole-genome duplication [3,4], causing, at least initially, step-wise increases in GS, and (ii) deviation in repeat copy numbers, that can either result in subtle or more dramatic GS changes [2].

Repetitive DNA sequences account for the majority of the genomic DNA in most plant species, occurring in a few to millions of copies [5]. As GS increases, so does the proportion of repetitive DNA, up to a certain point, after which degraded repeats that are difficult to classify represent a significant portion of what has been termed the genomic ‘dark matter’ [6]. By comparison, the size of the gene space probably remains relatively constant. In angiosperms the repeat types (i.e. (retro)transposable elements, (micro)satellite DNA, and truncated derivatives — see Glossary) can be fast evolving in absolute copy numbers and sequence, such that in species from many plant families there are reports of repeat element half-lives of 3–4 million years [species in Poaceae, Brassicaceae, Fabaceae, and Vitaceae, 7], near complete repeat turnover in the genome over timeframes of 5–10 million years [Solanaceae, 8], and repeat copy numbers changing GS two or three fold over just a few million years [Poaceae and Malvaceae, 9].

Changes in the number, location, and diversity of repeat sequences have a significant impact on gene space evolution [9]. Here we focus on recent insights into this dynamic interplay. We propose that an understanding of gene evolution and gene regulation requires a deep understanding of the genomic context of gene space, that is, the repeat landscape and genome architecture within which a gene is embedded. In addition, we explore the extent to which these processes operate at the upper end of the scale in terms of GS.

Influence of repeats on gene expression and function

It has been widely documented that the mobility and amplification of repeats, both satellite and transposable elements (TEs), can influence gene expression and function [reviewed in, e.g. 9–13], and, if left unchecked, will lead to increasing GS with potentially detrimental consequences on the phenotype [14]. To reduce the frequency of these processes, eukaryote genomes have evolved a variety of mechanisms to epigenetically silence repeat activity, including RNA-directed DNA methylation (RdDM; involving small interfering RNAs, siRNAs),

¹ Although larger genome sizes have been reported in some unicellular eukaryotes, their estimates are considered unreliable (see http://www.genomesize.com/statistics.php) as they have never been confirmed using appropriate techniques.
maintenance methylation, and histone modifications [15,16]. Yet such silencing of repeats can have repercussions on adjacent gene domains as RdDM has been shown to ‘seed’ the spread of methylation into regions that were not originally targets of siRNAs.

Nevertheless, recent studies of different TE families across the whole genome in maize and Arabidopsis have shown that spreading of methylation is not a characteristic of all TEs [17,18**], hence not all TEs impact adjacent genes (within 1 kb) in this way. West et al. [18**] have also shown that there are differences in the amount of TE methylation spreading between species, with more spreading into flanking regions in maize than Arabidopsis. However, in maize, the boundaries between genes and TEs are marked by elevated cytosine methylation at CHH motifs (forming CHH methylation islands, triggered by the activities of RdDM), resulting in altered chromatin conformation. This may act to preferentially inhibit TE amplification whilst enabling gene expression [19].

Recently, it has become clear that as well as these cis-effects, siRNAs produced following activation of TEs can also regulate the expression of Arabidopsis genes in trans [20]. Given the relatively low number of genes that are targeted by siRNAs in Arabidopsis (30%) compared with rice (80%), which has a larger genome [21], this raises the question as to whether the impact of such trans-effects of siRNAs on gene expression may become increasingly complex as GS increases.

In addition to these repeat-silencing effects, specific structural features of LTR retrotransposons make them particularly likely to influence the expression of nearby genes. Promoter/regulatory sequences at both LTRs allow 3’ LTRs to drive bleed-through transcription, which can extend into neighbouring sequences. Indeed, it is likely that the average plant genome has hundreds or thousands of genes controlled by regulators originally derived from TEs [10]. Depending on both the position of insertion in gene regions, and the orientation of the LTR, this type of transcription can lead to multiple and antagonistic effects on gene expression [22]. Indeed, it is now becoming apparent that TEs enable fine-tuning of gene expression and can have an important regulatory role. This echoes original work on TEs in maize by Barbara McClintock, with her original name ‘controlling elements’ — ‘The real point is control. The real secret of all of this is control. It is not transposition’ [22,23].

It is known that exaptation of TEs is important in enabling phenotypic plasticity in plants, producing variability in some important agricultural traits, and in responses to stress. The classic examples of red and white grape phenotypes and the blood-orange phenotype are each associated with insertion of TEs into upstream promoters [11]. More recently, a genome-wide association study (GWAS) of 368 maize inbred lines showed that a CACTA-like TE within the promoter of the gene ZmCCT was associated with reduced photoperiod sensitivity in maize and reduced flowering time, and hence likely to be significant in its post-domestication expansion to temperate regions [24]. Evidence is also growing that shows TEs have been exapted in plant defence responses. For example, knock-out of a Ty1/copia retrotransposon, ATCO-PIA4, adjacent to RPP5 genes was shown to result in increased susceptibility to downy mildew in Arabidopsis [25]. Furthermore, in Arabidopsis the study of cytosine methylation mutants with increased susceptibility to Fusarium oxysporum showed that a significant fraction of genes that are differentially expressed are also associated with reduced CHH methylation in upstream promoters carrying TEs [26**]. Such results suggest that the TEs and their regulation are involved in disease resistance. There is also likely to be a wider role of TEs in regulating the nature of the transcriptome response to abiotic stress. For example, an analysis of the maize inbred line B73 under temperature and UV stress has shown that genes which are up-regulated are significantly more likely to be closely associated with 20 TE families, whilst genes that are down-regulated are frequently associated with a further three TE families [27**].

Together these studies highlight the increasingly diverse ways in which repeats can impact how genes are regulated and expressed in response to the environment.

**Influence of repeats on the evolution of gene space**

Most flowering plant lineages have undergone multiple rounds of polyploidy in their ancestry, a process that is
ongoing in many lineages [28]. Consequently, supposedly diploid species are in fact palaeopolyploids. For example, the ‘diploid’ *A. thaliana* may be as high as 48-ploid, and ‘tetraploid’ cytotypes (96-ploid) exist in nature [28]. Recurring polyploidy has resulted in the evolution of large gene families, with genes frequently occurring in multiple syntenic blocks arranged in a co-linear order, the distribution of these blocks reflecting the lineage’s history of chromosomal rearrangements. However, the fate of the duplicated gene copies themselves following polyploidy can differ. Some genes are resistant to losses post-polyploidy [i.e. gene balance hypothesis, 29], whilst others are free to drift back to lower ‘diploid’ copy numbers.

It is clear that for many genes, copy numbers can diverge quickly, with a long-term tendency to reduce copy number for all gene duplicates that do not have selection pressures maintaining them. One mechanistic driver of that reduction in copy number is likely to be the proximity of LTRs and the frequency of unequal recombination, which leads to the deletion of sequences between adjacent LTRs. Recombination between TEs and indeed any adjacent repeats can have multiple effects. First, recombination-based removal from the genome limits the impact of repeats on gene expression [e.g. methylated LTR retrotransposons in rice are preferentially removed from regions surrounding genes, 13,30]. Second, recombination between adjacent repeats can also involve the deletion or duplication of intervening genes, giving rise to copy number and presence/absence variants.

These structural variants (SVs) generate genomic complexity that differentiates species and populations/lines within species [31]. SVs can occur at an astonishingly high frequency. In maize it has been shown that SVs influence thousands of genes [e.g. ~83% of 8681 transcripts were only expressed in subsets of 503 diverse inbred lines, 32]. Indeed, the maize reference genome, B73 [33], carries only c. 70% of all the low-copy sequences identified in 27 diverse maize accessions [34].

Thus, removal or amplification of repeats and genes generates considerable structural variation upon which selection can act. Over time, in *Arabidopsis*, Gaut *et al.* [35] suggest that the dynamics of gene duplication via ancestral polyploidy, and losses and gains of genes through recombination has resulted in a genome whereby surviving duplicates derived from each mechanism occur in similar numbers. Furthermore, differential selection pressures on duplicates lead to genome fractionation,
whereby regions of the genome become enriched for genes resistant to post-duplication losses [36].

**Influence of repeats on gene space across the range of plant GSs**

Comparative mapping in a range of grass species differing in GS has revealed that the evolution of gene space in terms of organization, duplication rates, and number of genes reflects the GS of the species. An accelerated rate of evolution was found in the larger genomes of species in the Triticeae tribe of the grass family compared with the smaller genomes of *Oryza sativa*, *Brachypodium distachyon*, and *Sorghum bicolor* [37–39].

Given these observations, and extrapolating these data to the biggest angiosperm genomes which are several-fold larger than the grasses studied above, one might expect that gene regulation and expression networks in species with giant genomes would be in utter chaos. Clearly this is not the case. One reason could be that repeats do not accumulate randomly in the genome and/or their removal is not random. Consequently, with increasing genome size, repeats can accumulate in ever increasing blocks, pushing genes into islands in an ever more partitioned genome [40]. It is also possible that genomic and epigenetic processes, influencing chromatin conformation, gene expression, and recombination, are not operating in the same way across the range of GSs encountered in angiosperms.

The GS of an individual represents the balance between processes that amplify and delete sequences, for example, polyploidy, (retro)transposition, illegitimate and unequal recombination, and non-homologous end joining in DNA repair. The epigenetic silencing of repeats described above, whilst it may indeed reduce the frequency of, for example, retrotransposition, will also influence recombination and DNA repair pathways because the chromatin may be heterochromatinised and hence rendered less accessible. In particular it will influence the balance between homologous and non-homologous DNA repair and the frequency of DNA deletion through unequal recombination and illegitimate recombination — both of which have been shown to contribute to genome downsizing. Indeed Fedoroff [41] stated, ‘I contend that it was precisely the evolution of prokaryotic mechanisms to regulate homologous recombination within the eukaryotic genome that made it possible for genomes to grow’. Thus, it can be argued that large, repeat-rich genomes become locked down by epigenetic silencing, reducing the frequency of repeat removal [42**].

In support of this hypothesis, Kelly *et al.* [42**] showed that in *Fritillaria*, the genus with the largest known GSs amongst diploid plants, the repeat profile is not dominated by a few, rather homogeneous repeats that make up a substantial proportion of the genome, as is typical of species with small genomes. This phenomenon is also seen in some species of amphibians and lung fish which also have very large genomes [43]. Instead, in *Fritillaria*, at least, there is a plethora of diverse repeats, many in large copy number, but each accounting for only a small proportion of the whole genome. The data also indicate that the repeats are ancient, suggesting that they are not being deleted and turned over, but rather are slowly accumulating over time.

Furthermore, the dynamic means that they are free to diverge through accumulation of mutations, becoming low-abundance unique and inactive DNA that represents substantial proportions of the genome [perhaps up to 40–50% in very large genomes, 42**]. A consequence of this erosion of repeats is that as GS increases, the genes may be found in an accumulating sea of non-repetitive DNA, despite the overall huge GS. Thus, paradoxically gene space may be less vulnerable to the effects of repeats than in species with small genomes. However, that stability may itself come at an evolutionary cost, as it is the variation generated by repeat dynamics that makes up a significant amount of genetic variation upon which selection can act.

**Acknowledgements**

We thank NERC for funding (NE/G01724/1) and a PhD studentship to SD.

**References and recommended reading**

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


17. Eichten SR, Ellis NA, Makarevitch I, Yeh C-T, Gent JI, Guo L, Mcginnis KM, Zhang X, Schnable PS, Vaughn MW et al.: Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLOS Genetics 2012, 8:e1003127.


This paper investigated TE associations with nearby genes in a genome-wide manner, showing a significant association of 20 TE families with up-regulated genes during abiotic stress in maize, thereby providing stress-responsive enhancer-like functions. They also provide evidence that TE insertions create allelic diversity influencing stress responses.


The authors investigate 503 inbred maize lines to characterise the maize pan-genome, showing that 16.4% of 8681 representative transcript assemblies are expressed in all lines, and that 82.7% are only expressed in subsets of the lines.


lack of DNA removal characterizes extreme expansions in genome size. New Phytol 2015, 208:596-607. These authors investigate the nature of repeats in the angiosperm genus *Fritillaria*, which includes diploids ranging in GS by 30 Gb. They show that the genome is composed of many diverse repeats, none of which constitutes a substantial proportion of the genome. The data are consistent with a hypothesis that failure to remove DNA rather than runaway repeat expansion causes GS enlargement.

43. Metcalfe CJ, Casane D: Accommodating the load — the transposable element content of very large genomes. Mobile Genetic Elements 2013, 3:e24775.