

Correlated evolution of genome size and seed mass

Jeremy M. Beaulieu¹, Angela T. Moles², Ilia J. Leitch³, Michael D. Bennett³, John B. Dickie⁴ and Charles A. Knight¹

Summary

- Previous investigators have identified strong positive relationships between genome size and seed mass within species, and across species from the same genus and family.
- Here, we make the first broad-scale quantification of this relationship, using data for 1222 species, from 139 families and 48 orders. We analyzed the relationship between genome size and seed mass using a statistical framework that included four different tests.
- A quadratic relationship between genome size and seed mass appeared to be driven by the large genome/seed mass gymnosperms and the many small genome size/large seed mass angiosperms. Very small seeds were never associated with very large genomes, possibly indicating a developmental constraint. Independent contrast results showed that divergences in genome size were positively correlated with divergences in seed mass.
- Divergences in seed mass have been more closely correlated with divergences in genome size than with divergences in other morphological and ecological variables. Plant growth form is the only variable examined thus far that explains a greater proportion of variation in seed mass than does genome size.

Key words: contribution index, genome size, independent contrasts, polyploidy, quantile regression, seed mass.

Introduction

Nuclear DNA amount varies over four orders of magnitude in plants (Bennett & Leitch, 2005). Recent phylogenetic studies have revealed the dynamic nature of genome size evolution, where both increases and decreases have taken place within lineages (Leitch *et al.*, 1998, 2005; Soltis *et al.*, 2003; Johnston *et al.*, 2005; Price *et al.*, 2005). The amplification of transposable elements (Bennetzen, 2002; Kidwell, 2002; Bennetzen *et al.*, 2005) and polyploidy are both thought to be pervasive mechanisms for increasing bulk nuclear DNA amount (2C DNA amount). The mechanistic basis for genome reduction is still poorly understood (Petrov *et al.*, 2000; Bennetzen *et al.*, 2005) but in plants is partly associated with the re-diploidization of the polyploid genome with accompanied

downsizing of the monoploid genome (1Cx DNA amount; Leitch & Bennett, 2004). Genome size reduction may require strong selection pressures (Petrov *et al.*, 2000), which implies that there may be some cost associated with large genomes, or benefits associated with small genome size. Therefore, there has been a growing interest in the phenotypic consequences of variation in genome size (Knight *et al.*, 2005). There is a strong positive relationship between cell size and 2C DNA amount (Bennett, 1972, 1973; Edwards & Endrizzi, 1975; Sugiyama, 2005). In addition to cell size, 2C DNA amount is positively correlated with cell cycle duration (Rees *et al.*, 1966; Baetcke *et al.*, 1967; Bennett *et al.*, 1983; Lawrence, 1985). Based on these cellular correlations, it is conceivable that many other morphological and physiological traits may scale with DNA content.

Previous studies have shown consistent, positive associations between genome size and seed mass in comparisons between populations of the same species and across groups of species within the same genus or family (Table 1). Seed mass varies over nearly 12 orders of magnitude, from the dust-like seeds of Orchidaceae to the 20 kg double coconut. Seed mass variation carries significant agronomic and ecological consequences, and therefore understanding the genetic basis of seed mass variation is of great interest. In our view, genome size may be related to seed mass through cell size effects within seed organs (i.e. endosperm and embryo). Step increases in genome size may lead to larger endosperm cells, resulting in increased seed volume and mass. Increased cell size in any other seed organ (cotyledons or hypocotyls, for example) could also lead to increased seed mass. Here we report results of a large analysis involving 1222 species (by far the largest study to date) where we test the hypothesis that variation in genome size is positively correlated with seed mass variation.

To analyze the relationship between genome size and seed mass we used a statistical framework that included four different tests, each of which asked a different question of the data:

- Is there a predictable statistical association between genome size and seed mass? To answer this question we use simple regression statistics.
- Is the relationship between genome size and seed mass polygonal, and can we detect boundaries or limits to the bivariate relationship? We use quantile regression to provide a more complete view of the relationship than what is captured by the median regression statistics alone.
- Are divergences in genome size associated with divergences in seed mass? Has there been correlated evolution of these traits? From our dataset we construct a consensus ‘mega-tree’ phylogeny with the most current molecular diversification times (Wikström *et al.*, 2001) and use independent contrast analyses (Felsenstein, 1985) to answer this question.
- Which evolutionary divergences contribute the most to present-day variation in genome size and seed mass, and do

nodes that contribute significant variation in genome size also contribute significantly to extant variation in seed mass? To answer this we use contribution index scores, which estimate the amount of present-day variation explained by divergences at each node in the phylogeny (Moles *et al.*, 2005a). We use a rank correlation statistic between contribution index scores for each trait to test this question.

By performing all four of these analyses we provide a comprehensive view of the relationship between genome size and seed mass for 1222 species.

Materials and Methods

Genome size and seed mass

The term genome size has been widely used in the literature to refer either to the total DNA amount in the nucleus or, in a more restricted sense, to the DNA content of the monoploid genome. To avoid confusion Greilhuber *et al.* (2005) proposed that ‘genome size’ should continue to be used in the broad sense as a covering term but proposed the terms ‘holoploid genome size’ or ‘2C value’ to refer to the DNA content of the unreplicated nucleus, and ‘monoploid genome size’ or ‘1Cx value’ to refer to the DNA content of the monoploid genome with chromosome number x. The 1Cx value (calculated by dividing the 2C value by the ploidy level) is predicted to be similar between a diploid and autopolyploid race of the same species, while the 2C DNA content should show step increases. However, it appears that polyploid formation may be accompanied by genome downsizing, which results in a smaller 1Cx amount compared with the diploid progenitor species (Leitch & Bennett, 2004). Here we test for associations between both 2C DNA and 1Cx DNA content with seed mass.

Estimates of 2C DNA content were taken from the Plant DNA C-values database (prime estimates; Bennett & Leitch, 2005) and were combined with seed mass estimates from the Seed Information Database (Flynn *et al.*, 2004); both databases

Table 1 Previous studies on the relationship between genome size and seed mass

Correlation ^a	Level ^b	Description	Authors
+	Pop.	15 <i>Dasyprum villosum</i>	Caceres <i>et al.</i> (1998)
+	Pop.	12 Soybean strains	Chung <i>et al.</i> (1998)
+	Sp.	131 British angiosperms	Thompson (1990)
+	Sp.	43 British plants	Grime <i>et al.</i> (1997)
+	Sp.	22 <i>Crepis</i> spp.	Jones and Brown (1976)
+	Sp.	12 <i>Allium</i> spp. and 6 <i>Vicia</i> spp.	Bennett (1972)
+	Sp.	85 <i>Pinus</i> spp. worldwide	Grotkopp <i>et al.</i> (2004)
+	Sp.	19 Mediterranean annuals	Maranon & Grubb (1993)
+	Sp.	148 species in California flora	Knight & Ackerly (2002)
+	Sp.	Several Poaceae and Fabaceae	Mowforth (1985)
NS	Sp.	16 grassland species, UK	Leishman (1999)

^aCorrelations are either positive (+) or not significant (NS).

^bStudies were classified into different levels – those dealing with different populations of the same species (Pop.) or multiple species (Sp.).

are maintained at the Royal Botanic Gardens at Kew. There were 1222 species with known 2C DNA content that also had a seed mass in the seed mass dataset. Where there were multiple estimates for seed mass, the geometric mean was used as the species value. We calculated 1Cx DNA content for species with known ploidy by dividing the 2C value by the ploidy level (i.e. 2x, 4x, etc.). Because many species have a range of ploidy, which can confound the calculation of the monoploid genome size, we only used species where one ploidy level was reported. Therefore 1Cx values were calculated for only 999 species.

For 2C DNA amount, our dataset comprised 1087 angiosperms and 135 gymnosperms. 2C DNA amounts ranged 525-fold from 0.284 to 148.95 Gbp. The angiosperm sample was well distributed phylogenetically; it included 139 families out of 443 families currently known (Stevens, 2005), including representatives of the basal angiosperms as well as species from each of the major angiosperm clades (i.e. monocot and eudicot). Our dataset provided an adequate representation of the full range of 2C values currently known for angiosperms (2C = 0.134–254.8 Gbp). The gymnosperm sample included nine families from all four gymnosperm orders (i.e. Cycadales, Ginkgoales, Gnetales, Pinales) and the 2C DNA values ranged nearly sixfold from 12.8 to 70.6 Gbp (the range for 207 gymnosperms currently in the Plant DNA C-values database is 15-fold from 4.6 to 70.6 Gbp).

Constructing a 'mega-tree'

We constructed a 'mega-tree' hypothesis using Phylomatic (tree version: R20040402; Webb & Donoghue, 2005). This online software is a compilation of previously published phylogenies and its ordinal 'backbone' and family resolutions are based on the Angiosperm Phylogeny Website (APweb; Stevens, 2005), the best current estimate of relationships among higher plants. Currently, Phylomatic has complete familial representation. The program first matches a species by genus, and then by family. If one genus is missing within a family, the entire set of genera for that family is returned as a polytomy within the 'mega-tree'. Because our dataset spans many genera within many families, most relationships were placed as polytomies. However, many of these polytomies could be resolved by consulting the current literature (supplementary material, Table S1). As a rule for resolving these polytomies, when there were conflicting branching patterns in the literature, a polytomy at the most ancestral node of the family was maintained.

Phylomatic assumes extant gymnosperms are monophyletic. While this view is controversial (morphological data support Gnetales being sister to angiosperms; Donoghue & Doyle, 2000; Friedman, 2006), molecular data generally support it (Chaw *et al.*, 2000; Soltis *et al.*, 2002; Burleigh & Mathews, 2004). However, placement of the four orders of gymnosperms (Cycadales, Ginkgoales, Gnetales, Pinales) is inconsistent. For this reason, we conservatively maintain the phylogeny output

by Phylomatic that contains a basal polytomy across the four orders of gymnosperms. For comparison when analyzing gymnosperms alone, we also tested three alternative phylogenies of gymnosperms based on current molecular data (Chaw *et al.*, 2000; Soltis *et al.*, 2002; Burleigh & Mathews, 2004).

Branch length information for our 'mega-tree' phylogeny was taken from age estimates published by Wikström *et al.* (2001). These authors applied a nonparametric rate-smoothing algorithm (allows for different clades to evolve at different rates) to a three-gene dataset that spanned nearly 75% of all angiosperm families. Estimates were then calibrated at a single point within the fossil record (the Fagales–Curcubitales divergence, 84 million yr ago (Mya)), to obtain the first comprehensive hypothesis of angiosperm diversification times. Recently, Bell *et al.* (2005) incorporated multiple fossil calibrations using Bayesian relaxed clock (BRC) and penalized likelihood (PL) methods to derive divergence times of several major basal groups, which included the origin of angiosperms. The age estimates of Bell *et al.* (2005) were not significantly different from estimates by Wikström *et al.* (2001). Because the Wikström *et al.* (2001) ages were more comprehensive, we used those age estimates here.

Dated nodes from Wikström *et al.* (2001) matched 49 divergences in our phylogeny. We then used the branch length adjustment algorithm in Phylocom (BLADJ; Webb *et al.*, 2006) to estimate the age for undated nodes. BLADJ sets a root node at a specified age and fixes all other known aged nodes. Branch lengths for undated nodes are then interpolated by evenly distributing them between nodes with known ages, which minimizes variance in branch length (Webb *et al.*, 2006). The ages within our phylogeny should be treated as approximations.

Analytical approach

We analyzed the relationship between genome size (which includes both 2C and 1Cx DNA content) and seed mass using four methods: regression, quantile regression, independent contrasts, and contribution index scores (described later in this section). Since our dataset includes two distinct groups of seed plants, gymnosperms and angiosperms, we also determined the influence that each group had on the overall relationships independently. We also analyzed the relationship within well-represented (in terms of sample size) families and genera of both angiosperms and gymnosperms (Table S2). All variables violated the assumptions of normality; therefore we log-transformed the data before analysis.

Regression. We used least-squares regression to test for an association between genome size and seed mass across all species. While we advocate using independent contrasts to infer adaptive or correlated evolutionary hypotheses (see later), regression analyses can provide predictive power, even in the absence of

significant independent contrast results, and therefore should be performed in conjunction with independent contrast analyses. In addition, the results of the regression test can be directly compared to quantile regression analyses (see following section).

Quantile regression. We used quantile regression to identify limits, boundaries, and shifting relationships within our bivariate distributions. Quantile regression extends classical least-squares regression by estimating slopes not only through the mean or median, but also through each quantile (or percentile) of the bivariate relationship. Significant changes in slope, or quadratic coefficients, through the quantiles of a bivariate distribution imply that the distribution is polygonal, or filled unimodal, rather than linear. Such bivariate distributional shapes are common in ecological and evolutionary analyses, yet only recently have statistical methods been available to quantify them. Quantile regression thus provides a more complete view of the relationship between x and y than what is captured by median least-squares regression alone.

We estimated the quantile regression coefficients for the 5th through the 95th quantiles. For example, the 65th quantile regression is calculated by minimizing residual errors around a line where 65% of the observations fall below the line and 35% fall above the line. Residuals for points that fall above the line are weighted by the quantile (0.65 for the 65th quantile), while the points falling below the line are weighted by one minus the quantile (corresponding to 0.35 for the same 65th quantile). The 50th quantile corresponds to the traditional median least-squares regression estimate, where an equal amount of points fall above and below the line. Koenker & Hallock (2001), Knight & Ackerly (2002), and Cade & Noon (2003) all provide a detailed discussion of quantile regression methods. We used the ‘quantreg’ package (Koenker *et al.* (2005) to perform our quantile regression analyses.

Independent contrasts. We used the analysis of traits (AOT) module (developed by Ackerly, 2006) of Phylocom (Webb *et al.*, 2006) to perform independent contrasts on our ‘mega-tree’ phylogeny. The AOT algorithm calculates standardized divergences of extant species and estimates internal node averages and divergences incorporating branch lengths (Felsenstein, 1985). A unique feature of AOT is that it can handle polytomies; our ‘mega-tree’ phylogeny contained many. AOT uses the method developed by Pagel (1992) to calculate independent contrasts with phylogenies that contain polytomies. AOT takes a particular polytomy and ranks species based on the value of the independent variable (in this case 2C DNA content or 1Cx DNA content), where the median value is then used to create two groups. Mean values are calculated for each trait between the two groups and the difference between these means is treated as a single independent contrast.

The consistency in the direction of subtraction when performing independent contrasts is important. AOT is useful

in that it sets the sign of the contrast for X (here we set genome size as X) to always be positive, and all other traits (seed mass) are then compared in the same direction across the node (Ackerly, 2006). Since the direction of subtraction is clearly arbitrary, reversing the direction of subtraction will result in a contrast of the opposite sign. Thus, all contrasts inherently have a mean value of zero and regression analysis of independent contrasts must be forced through the origin to account for this property (Garland *et al.*, 1992). We utilized the output of our standardized contrasts from AOT and used R (R Development Core Team, 2005) to obtain slope estimates and r^2 from a regression analysis forced through the origin.

AOT also calculates an absolute measure of trait radiation (divergence width) at each node, which is analogous to a standard deviation. We used the divergence width instead of trait differences (independent contrasts) because the standard deviation can be used when polytomies are present in the phylogeny (Moles *et al.*, 2005a; Ackerly, 2006). To examine the pattern trait evolution through time, we plotted the divergence width in genome size and seed mass with age estimates of Wikström *et al.* (2001) and age interpolations from BLADJ (see earlier section on Constructing a ‘mega-tree’). We then fit a loess curve to the data, with 5% of the points influencing the smoothness of the line, to uncover any particular geologic times that may have been more divergent than others.

Contribution index. The contribution index is a measure of how much a divergence at a particular node in the ‘mega-tree’ explains present-day variation within a trait (Moles *et al.*, 2005a). The contribution index is the product of two variance components: (i) the amount of variation within a focal clade resulting from a focal divergence; and (ii) the amount of the total variation within that focal clade compared to the whole tree (Moles *et al.*, 2005a). Each component is calculated from different decompositions of the sum of square deviations from internal node averages estimated by Phylocom. The decomposition of the sum of squares for trait divergences at each node was obtained from AOT to calculate each component, and subsequently the contribution index. The contribution index was calculated for genome size (2C and 1Cx) and seed mass separately. A Spearman rank correlation was used to determine if nodes with high contribution to genome size variation were also nodes with a high contribution to seed mass variation.

Results

Regression and quantile regression

Across all species, the relationship between 2C DNA content and seed mass appears curvilinear and concave. Small 2C DNA content species have a wide range of seed masses, while species with large 2C DNA content tend to have larger seeds

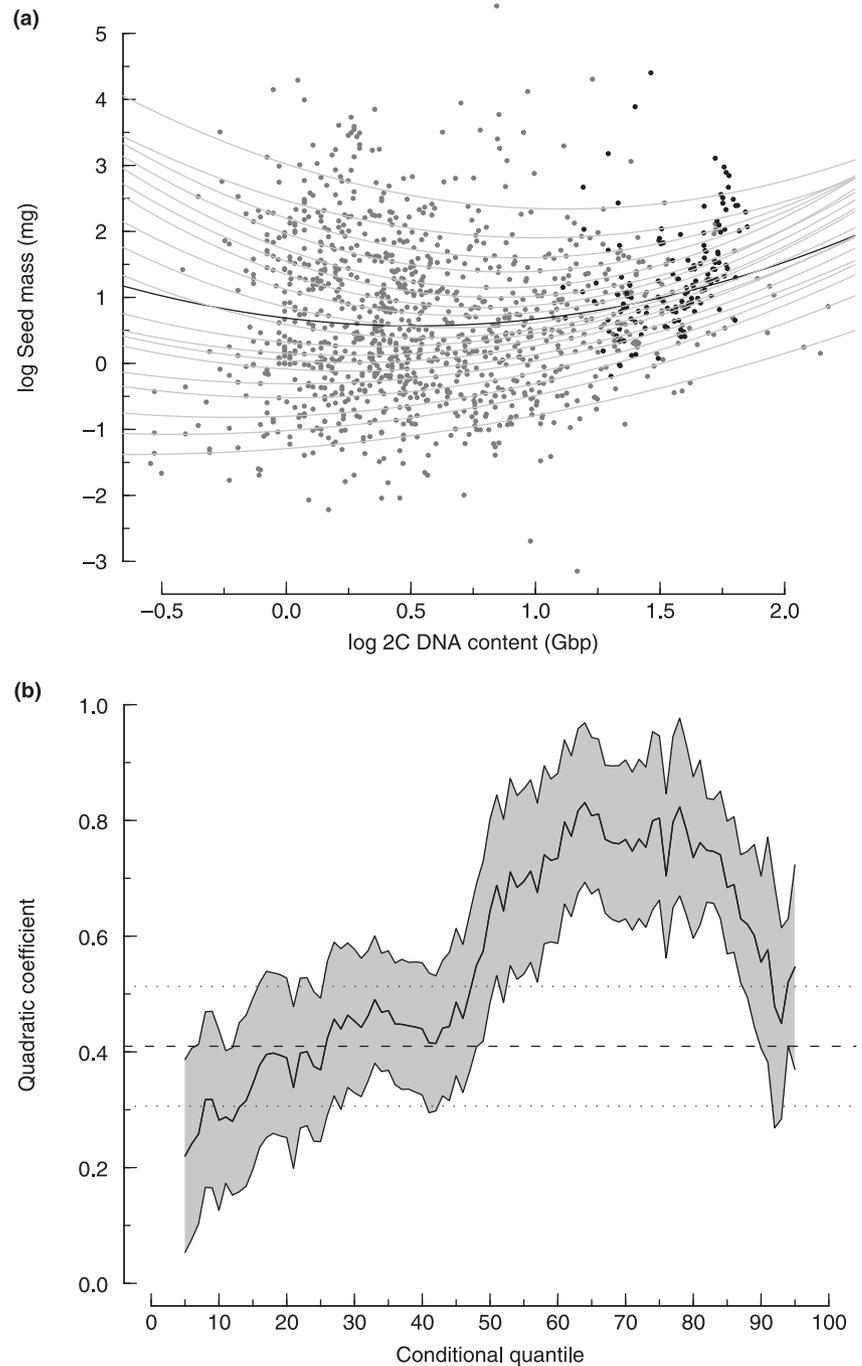


Fig. 1 (a) The relationship between 2C DNA content and seed mass. Data are split into gymnosperms (black circles) and angiosperms (gray circles). Angiosperms alone showed no trend, while gymnosperms are significantly positive. However, when correcting for influence of phylogeny, the overall relationship is significant, angiosperms are significantly positive, and gymnosperms show no trend (see Table 2). Each gray line in (a) corresponds to a different quantile of the quadratic coefficient, with the 50th quantile (least-squares estimate) highlighted (solid black line). (b) Quantile regression analysis of the quadratic coefficient (a) showing an increasing quadratic coefficient (solid line) with increasing quantiles. Gray is the standard error of each quadratic coefficient estimate. The horizontal lines in (b) represent the least-squares line (dashed line) with standard error (dotted lines).

(Fig. 1a). Interestingly, the space occupied by mid-range 2C DNA content is depopulated for large seed mass species; the relationship looks quadratic (Fig. 1a). Therefore, we performed a partial F -test, *a posteriori*, to determine if quadratic regression is more appropriate than a 'straight line' linear regression. Across all species, the quadratic term reduced the squared errors (from 1582.2 to 1562.6; $F=15.66$, $P<0.001$). However, with the data separated into the different groups, gymnosperms appear to drive the quadratic pattern. The addition of a quadratic

term to the gymnosperms alone also significantly reduced the squared errors from 72.9 to 66.2; $F=13.56$, $P<0.001$. For angiosperms alone the quadratic term did not add any more explanatory power ($F=0.392$, $P=0.532$) and the slope of a 'straight line' linear regression was not significantly different from zero ($r^2 \leq 0.001$, slope = -0.015 , $P=0.845$; but see quantile regression results later in this section).

Considering the results above, we fitted a second-order polynomial regression model for data involving all species,

which was highly significant ($r^2 = 0.026$, $F_{2,1219} = 16.29$, $P < 0.001$). The least-squares quadratic coefficient was positive (slope = 0.409, $t_{2,1219} = 3.96$, $P < 0.001$). Quantile regression also showed that, from the 5th through the 95th quantiles, the quadratic coefficient was also significantly different from zero and positive. In general, the magnitude of the quadratic coefficient steadily increased throughout the quantiles (Fig. 1b), indicating that the quadratic relationship became more significant for species with the largest seed mass.

Between the 50th and 85th quantiles the quadratic coefficient was significantly different from the least-squares quadratic regression, highlighting the utility of quantile regression.

When analyzing angiosperms alone, the concavity of the relationship was no longer apparent (see F -test earlier, Fig. 2a). Therefore a first-order regression model was fitted. The model was not significant ($r^2 < 0.001$, $F_{1,1085} = 0.038$, $P = 0.845$), however, quantile regression analyses indicated that the linear coefficient steadily declined as the quantiles increased (Fig. 2b).

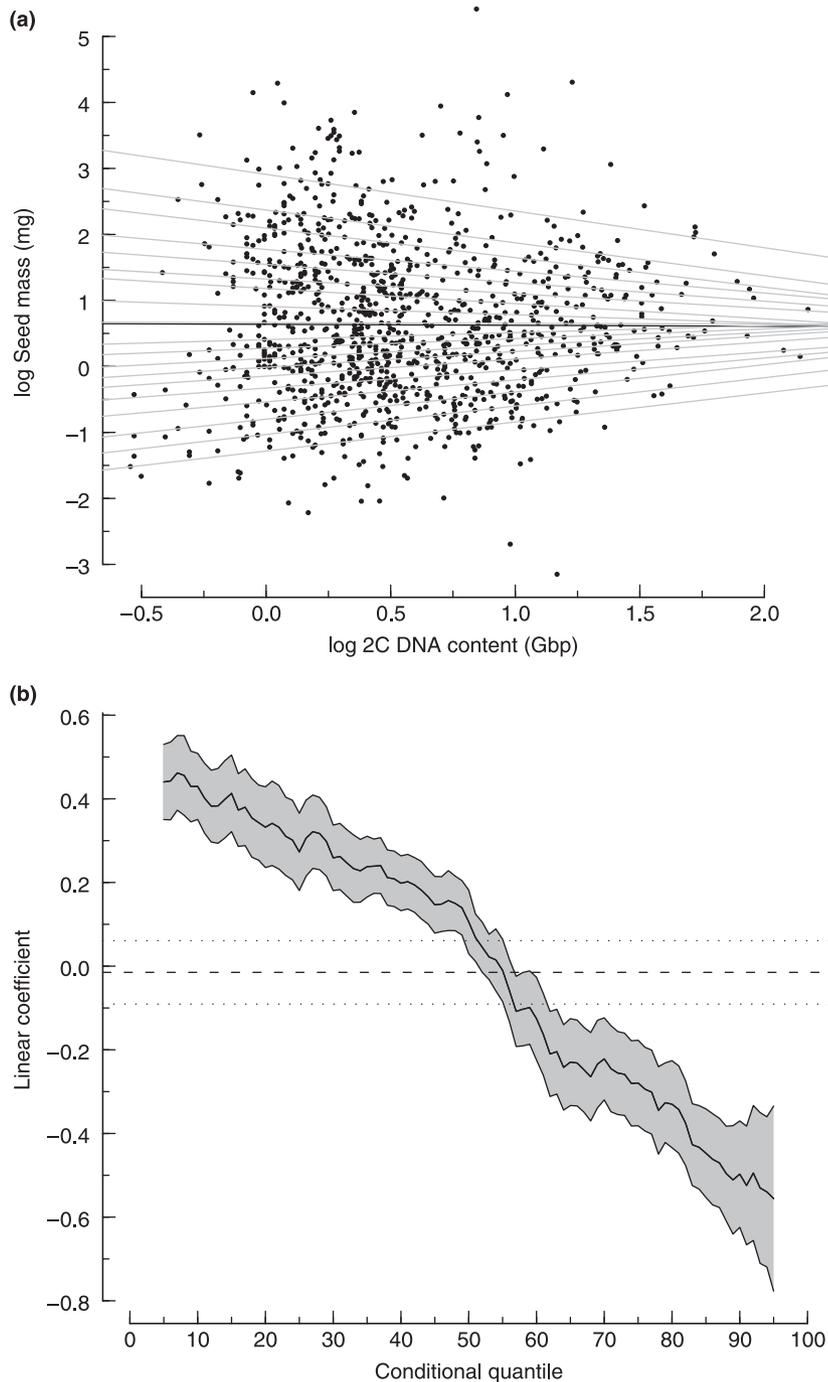


Fig. 2 (a) Scatter plot of 2C DNA content and seed mass in angiosperms. Each gray line in (a) corresponds to a different quantile regression result, with the 50th quantile (least-squares estimate) highlighted (solid black line). (b) Quantile regression analysis of (a) showing a decreasing slope with increasing quantiles. Gray is the standard error of each quantile regression estimate. The horizontal lines in (b) represent the least-squares line (dashed line) with standard error (dotted lines) and slope = 0 (solid line).

Only the coefficients corresponding to the 50th through the 61st quantiles were not significantly different from zero. Therefore, there was a shift from a significantly positive slope in the lower quantiles to a significantly negative slope in the upper quantiles.

The gymnosperm data alone did exhibit some concavity. A second-order linear regression model was significant ($r^2 = 0.214$, $F_{2,132} = 17.98$, $P < 0.001$; Fig. 3a). The quadratic coefficient was positive (slope = 7.14, $t_{2,132} = 3.68$, $P < 0.001$) and from the 16th through the 95th quantiles the quadratic coefficient

was significantly different from zero, but never significantly different from the least-squares quadratic regression. Nevertheless, these results indicated increasing concavity with increasing seed mass (Fig. 3b).

Regression coefficients for the 1Cx analyses had a greater magnitude than the coefficients of the 2C DNA analyses (except across all species, see Table 2). Quantile regression results for 1Cx DNA content and seed mass paralleled results found when using 2C DNA content.

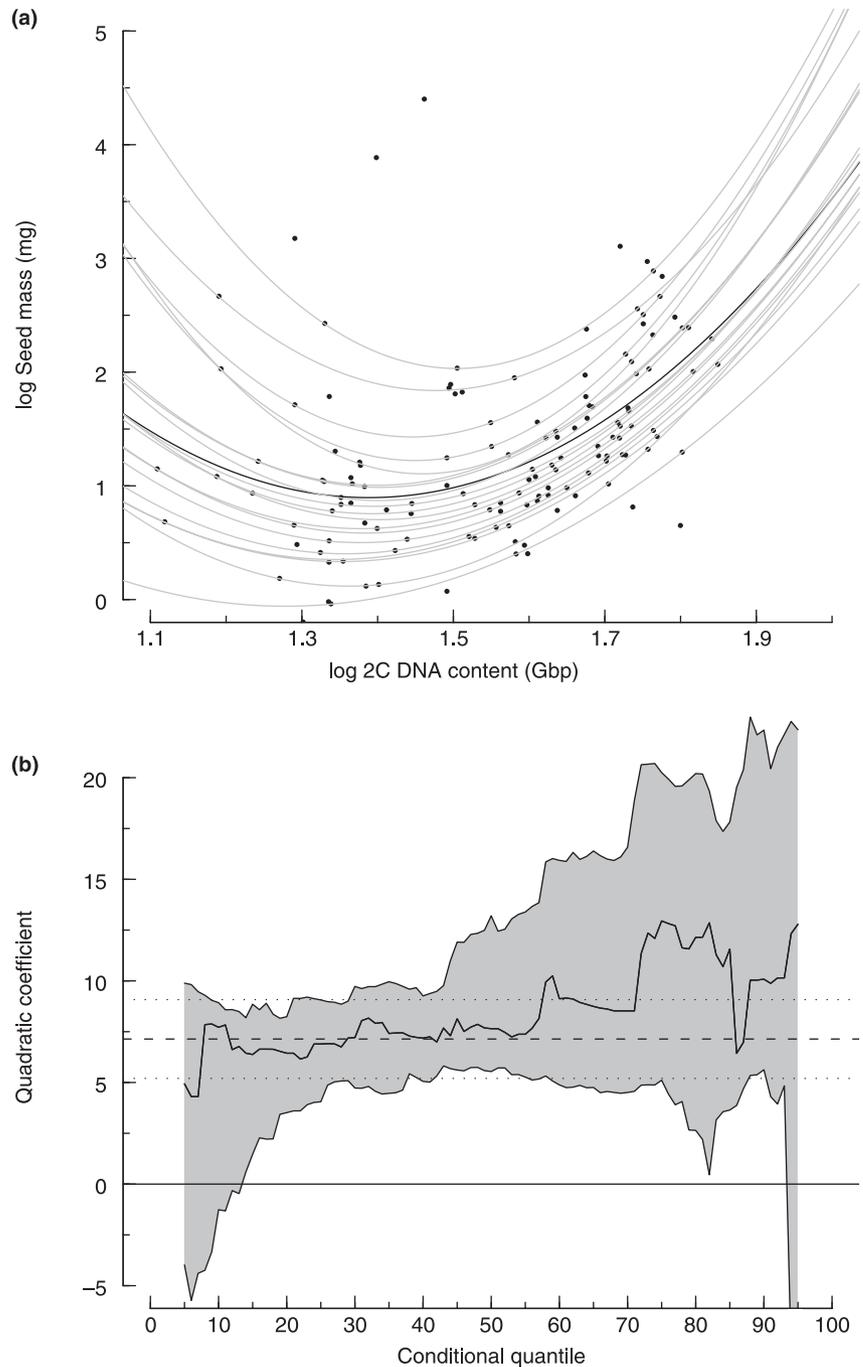


Fig. 3 (a) Scatter plot of 2C DNA content and seed mass in gymnosperms. Each gray line in (a) corresponds to a different quantile dependent result using the quadratic coefficient, with the 50th quantile (least-squares estimate) highlighted (solid black line). (b) Quantile regression analysis of (a), the quadratic coefficient showing an increase in slope with increasing quantiles. Gray is the standard error of each quadratic coefficient estimate. The horizontal lines in (b) represent the least-squares line (dashed line) with standard error (dotted lines) and slope = 0 (solid line).

Table 2 Results for the regression and independent contrast analyses across all species, and for angiosperms and gymnosperms analyzed separately, for 2C and 1Cx DNA content with seed mass

	Regression			Independent contrasts		
	<i>n</i>	<i>r</i> ²	Slope	<i>N</i> _{cont}	<i>r</i> ²	Slope
2C						
All species	1222	0.026**	0.409†**	686	0.033**	0.382**
Angiosperms alone	1087	< 0.001	-0.015	590	0.033**	0.381**
Gymnosperms alone	135	0.214**	7.14†**	95	0.023	0.620
1Cx						
All species	999	0.041**	0.304†**	550	0.062**	0.594**
Angiosperms alone	886	0.004	0.163	467	0.062**	0.592**
Gymnosperms alone	113	0.209**	7.47†**	82	0.047*	1.01*

Regressions for the independent contrasts were forced through the origin.

*N*_{cont}: number of contrasts in the independent contrasts analyses.

†, quadratic slope; *, *P* < 0.05; **, *P* < 0.01.

Independent contrasts

There was a significant positive relationship between 2C DNA content and seed mass across all species when using independent contrasts ($r^2 = 0.033$, slope = 0.382, $P < 0.001$; Table 2; Fig. 4a). We found a significant positive relationship when analyzing angiosperms alone ($r^2 = 0.033$, slope = 0.381, $P < 0.001$), but not for gymnosperms alone ($r^2 = 0.023$, slope = 0.620, $P = 0.137$). When testing 1Cx DNA content and seed mass across all species, the relationship was highly significant and explained nearly twice the variation as 2C DNA ($r^2 = 0.062$, slope = 0.594, $P < 0.001$; Fig. 4b). Angiosperms alone also had a highly significant positive independent contrast result when testing 1Cx (Table 2). Gymnosperms alone did show a marginally significant trend; however, we found no significant correlated evolution for gymnosperms, for both 2C and 1Cx DNA content, when using three differing resolutions of the gymnosperm phylogeny (Table 3).

Across the history of seed plants, divergences in both 2C DNA and seed mass have remained relatively constant (Fig. 5). However, the largest divergences in both 2C DNA content and seed mass have occurred more recently in geologic time (Fig. 5). The pattern when examining the divergence width in 1Cx DNA content was identical to results when using 2C DNA content.

Contribution index

Nodes that made large contributions to present-day genome size variation also made large contributions to present-day seed mass variation (Fig. 6). Spearman rank correlation showed that contribution scores for 2C DNA content were positively associated with contribution scores for seed mass ($n = 686$, Spearman's $r = 0.422$, $P < 0.001$). This was also true for 1Cx DNA content and seed mass ($n = 550$, Spearman's $r = 0.400$, $P < 0.001$).

For our dataset, the single most important contribution to present-day 2C DNA content was the divergence between angiosperms (1087 spp., mean 2C value = 3.53 Gbp) and gymnosperms (estimated at 325 Mya; Judd *et al.*, 2002; 135 spp., mean 2C value = 35.6 Gbp; Table 4). This divergence was also the most important for present-day seed mass variation. Angiosperms have smaller seeds (mean seed mass = 4.26 mg) compared with the large-seeded gymnosperms (mean seed mass = 21.6 mg). The second most important contribution to 2C DNA content variation was a three-way polytomy at the base of Poales (estimated at 72 Mya; Wikström *et al.*, 2001), which included Typhaceae (two spp., mean 2C value = 0.810 Gbp), Cyperaceae/Juncaceae (24 spp., mean 2C value = 1.26 Gbp), and Xyridaceae/Poaceae (161 spp., mean 2C value = 7.63 Gbp). However, this node did not show a large contribution to seed mass variation; all nodes

Table 3 Independent contrasts results for alternative resolutions of the gymnosperm phylogeny

Differing topologies within gymnosperms	2C DNA			1Cx DNA		
	<i>N</i> _{cont}	<i>r</i> ²	Slope	<i>N</i> _{cont}	<i>r</i> ²	Slope
Ginkgoales sister to all other gymnosperms; Gnetales within conifers	97	0.016	0.558	84	0.040	0.716
Ginkgoales sister to all other gymnosperms; Gnetales sister to conifers	97	0.015	0.718	84	0.041	0.541
Ginkgoales sister to Cycadales; Gnetales sister to conifers	97	0.016	0.581	84	0.045	0.738

*N*_{cont}: number of contrasts.

Regressions for the independent contrasts were forced through the origin. None of the analyses was significant.

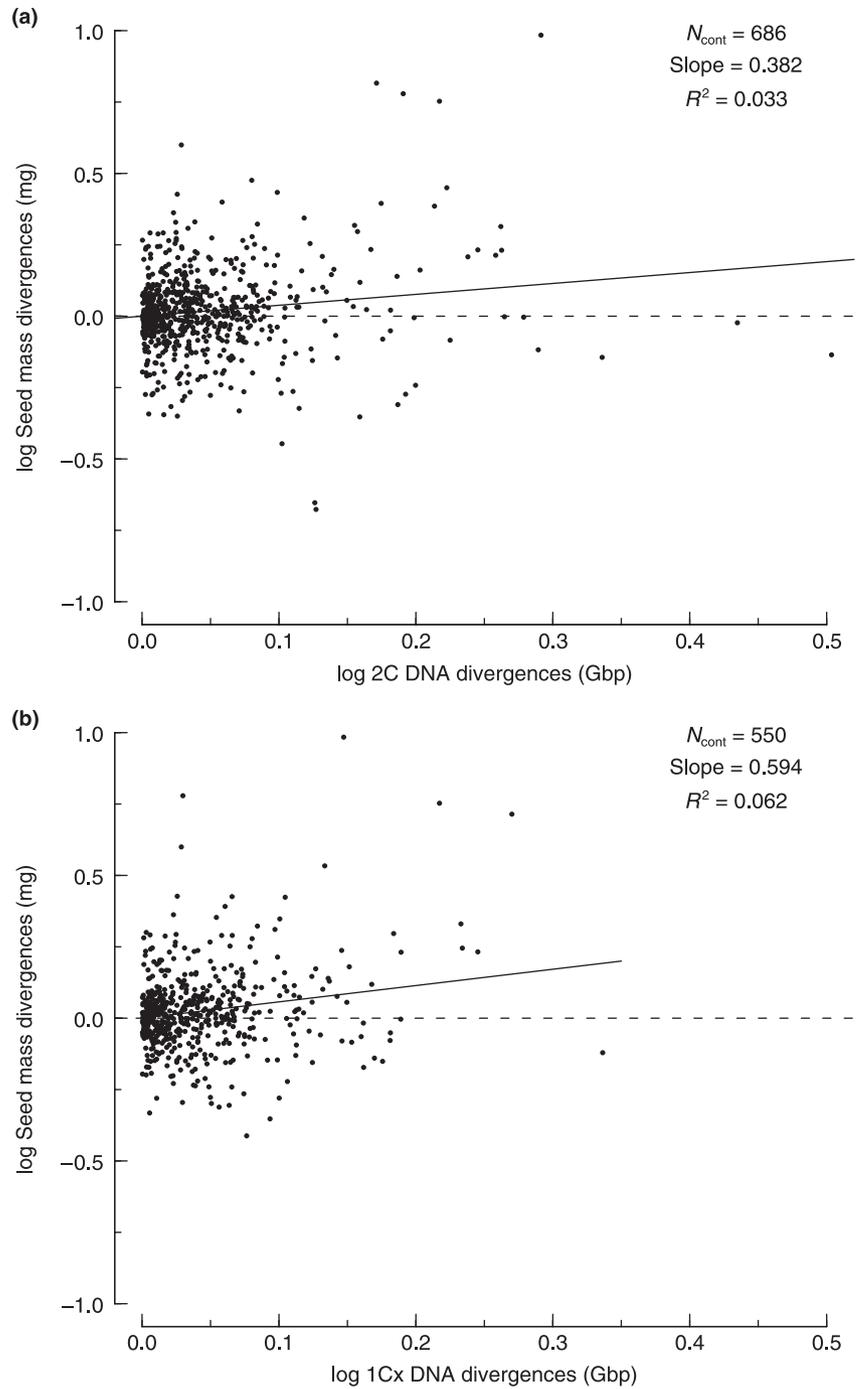


Fig. 4 The relationship between divergences in genome size and divergences in seed mass. For both the 2C (a) and 1Cx DNA datasets (b), there was a significant and positive relationship with divergences in seed mass, which were primarily driven by divergences within angiosperms (divergences in gymnosperms were relatively small). 1Cx divergences explained more variation (6.2%) in seed mass evolution than did divergences in 2C DNA content (3.3%; see Table 2). Regression lines (solid lines) are forced through the origin and the horizontal slope (dashed line separating the two quadrants) is also shown.

within this polytomy led to clades that produced relatively small seeds (Typhaceae, mean seed mass = 0.901 mg; Cyperaceae/Juncaceae, mean seed mass = 0.314 mg; Xyridaceae/Poaceae, mean seed mass = 2.16 mg).

The most significant contribution to present-day 1Cx DNA variation was also the divergence between angiosperms (886 spp., mean 1Cx = 1.47 Gbp; mean seed mass = 4.21 mg) and gymnosperms (113 spp., mean 1Cx = 16.6 Gbp; mean seed mass = 20.8 mg; Table 5). The second most important

contribution to present-day 1Cx DNA content was a divergence (estimated at 161 Mya; Wikström *et al.*, 2001) at the node that led to the monocots (215 spp., mean 1Cx = 2.50 Gbp) and the rest of the angiosperms (magnoliids and the eudicots; 670 spp., mean 1Cx = 1.24 Gbp). There was an opposite trend in the contribution to present-day seed mass at this node; the monocots produced slightly smaller seeds (mean seed mass = 3.59 mg) than the combined mean of the magnoliids and eudicots (mean seed mass = 4.43 mg).

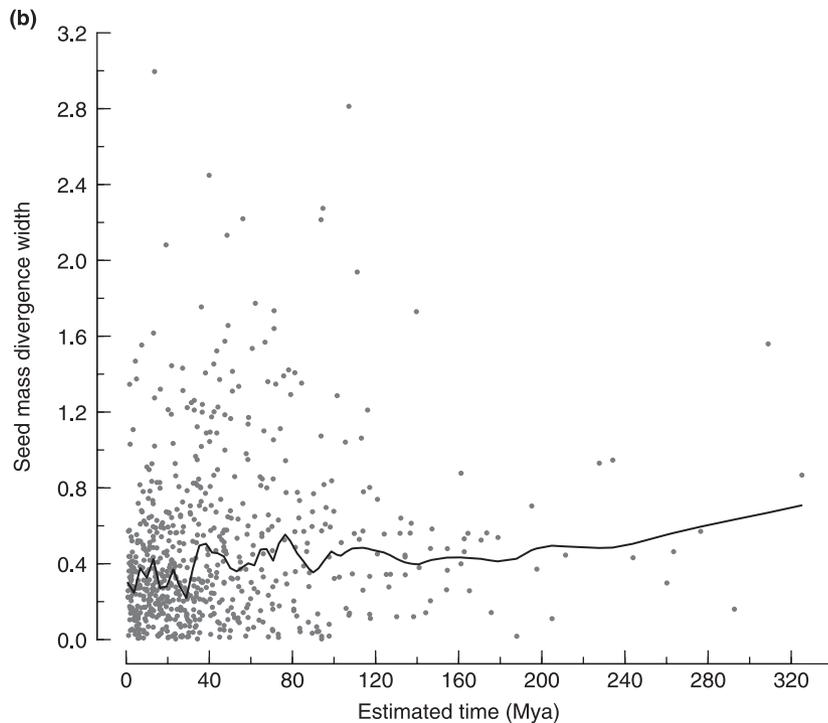
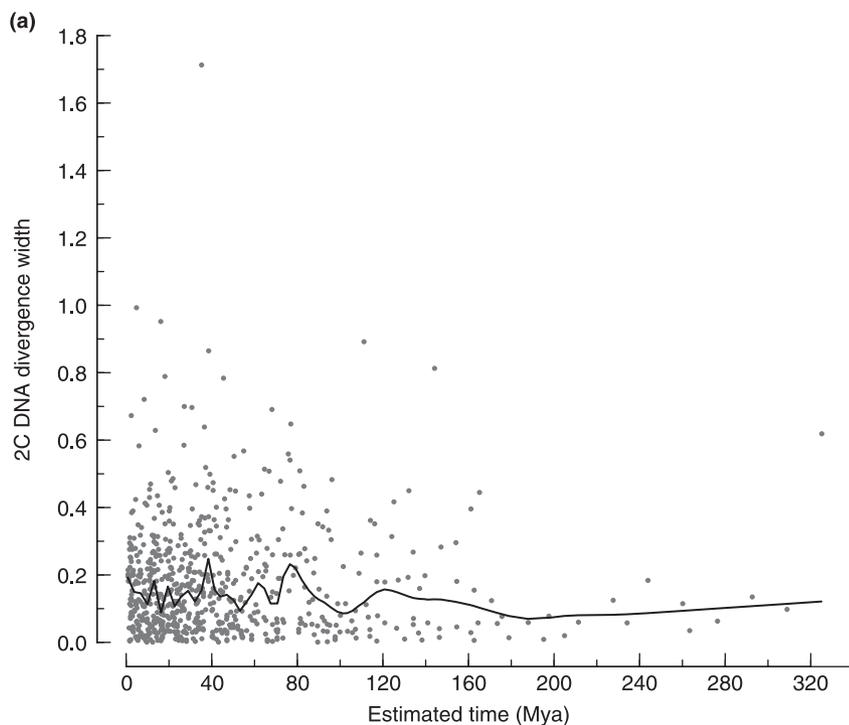


Fig. 5 The history of 2C DNA content and seed mass divergences through time. The divergence width is the absolute measure of trait radiation at each node. Estimated time is taken from age estimates of Wikström *et al.* (2001) and interpolations using a branch length adjustment algorithm (BLADJ; Webb *et al.*, 2006). The regression line is a loess curve with 5% of points influencing the smoothness. For both 2C DNA (a) and seed mass (b), the average divergence width has remained constant; however, the largest divergences have been relatively recent. Mya, million yr ago.

Discussion

Because the relationship between genome size and seed mass is complex (not straight-line linear) and exists within a phylogenetic framework, we used a variety of statistical techniques to describe the relationship between these traits.

We found that across extant species there was a significant positive relationship between genome size and seed mass. However, regression analysis only explained a small percentage of the error variation (Table 2). The r^2 values of our regression slopes were much weaker than we expected based on the uniformly positive and significant regression and correlation

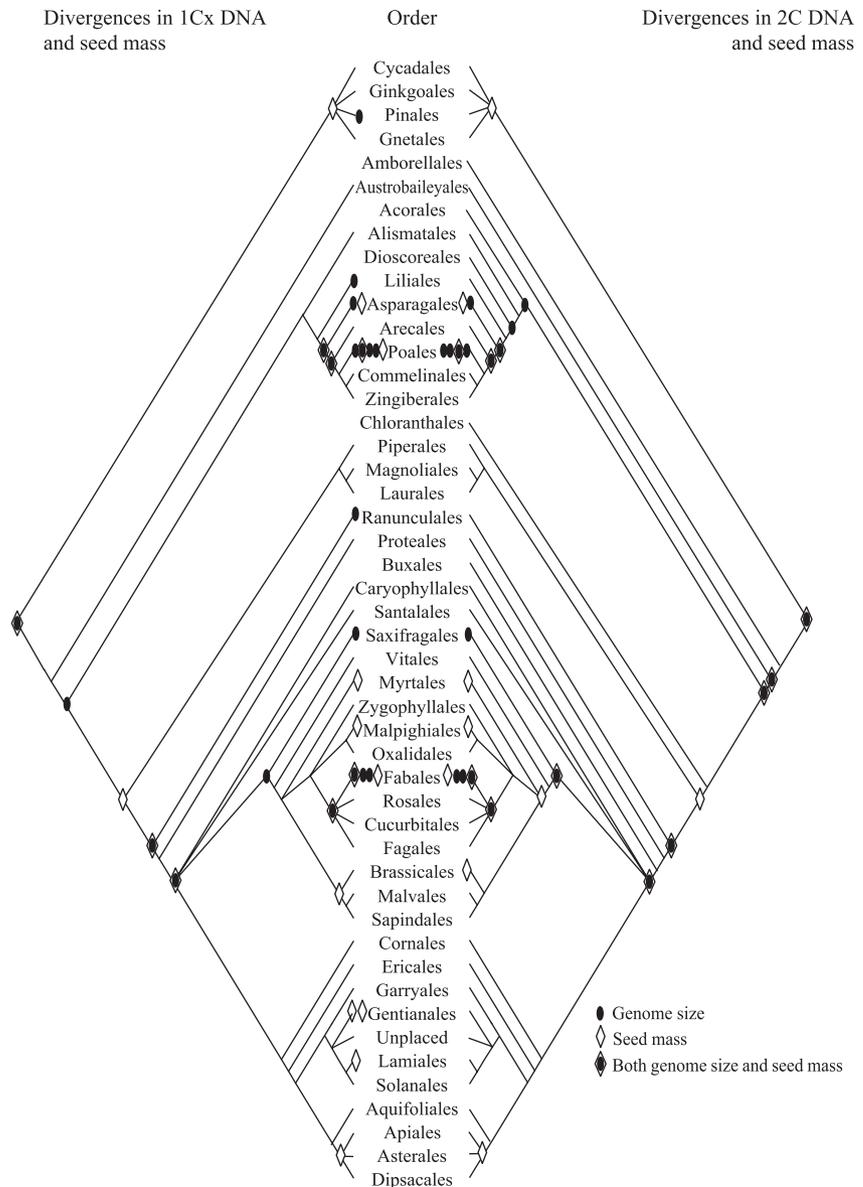


Fig. 6 Our mega-tree phylogeny to the order level, displaying the 20 largest contributions to present-day variation for both DNA content (1Cx and 2C DNA content) and seed mass. Black ovals represent the 20 divergences with the highest contribution score for genome size. White diamonds represent the 20 divergences with the highest contribution score for seed mass. Black ovals within with white diamonds represent divergences that were in both of the above sets. Diamonds and ovals at the tips of this tree represent divergences within orders.

results from 10 previous studies (Table 1). Furthermore, the slope for angiosperms alone was nearly zero. This may be because there are correlated divergences within the groups studied by other investigators (which were often congeners) but there are large leaps between groups that do not necessarily follow the same evolutionary trend. Across all species, independent contrast analyses showed that divergences in genome size are positively correlated with divergences in seed mass (Table 2; Fig. 4).

Genome size may set a minimum seed mass, if there is a developmental relationship between genome size and seed mass. Large seed masses have evolved in species with both small and large genomes, but large genome species rarely have small seed sizes (Figs 1–3). Interestingly, there is also an absence of species with mid-range genome sizes and

large seed masses (Fig. 1a). The reason for this absence is intriguing.

While the ‘straight line’ linear slope within angiosperms was nearly zero, the slope changed across quantiles, shifting from a significantly positive slope in the lower quantiles to a significantly negative slope in the upper quantiles. This suggests that the bivariate distribution forms a filled triangle (Fig. 2). The edges of this triangle represent limits to the distribution. For example, large genome species are unlikely to have small seeds (note the lack of species in the lower right-hand quadrant in Fig. 1a). But, large genome species in angiosperms do not have the largest seeds. Small genome angiosperms like *Cocos nucifera* (Arecaceae; 2C = 6.96 Gbp), *Castanospermum australe* (Fabaceae; 2C = 1.11 Gbp), and *Mangifera indica* (Anacardiaceae; 2C = 0.88 Gbp) hold that

Table 4 The 20 divergences making the largest contribution to present-day 2C DNA content variation (ranked 1–20) with accompanied contribution to seed mass variation explained by these nodes (seed mass rank)

Rank	2C DNA contribution	Divergences making the largest contribution	Seed mass Rank	Seed mass contribution
1	0.601	Angiosperms vs gymnosperms (c. 325 Mya)	1	0.321
2	0.072	Polytomy at the base of Poales (c. 72 Mya)	92	0.001
3	0.043	Monocots vs the rest of the angiosperms (c. 154 Mya)	20	0.010
4	0.014	Asparagales vs commelinids (c. 116 Mya)	3	0.033
5	0.011	Austrobaileyales vs rest of angiosperms (c. 161 Mya)	12	0.016
6	0.011	Saxifragales vs Vitales and the rosids (c. 114 Mya)	7	0.020
7	0.011	Dioscoreales vs Liliales, Asparagales and commelinids (c. 125 Mya)	79	0.002
8	0.011	<i>Brachypodium</i> spp. vs rest of the poidids; Poaceae (c. 8 Mya)	210	< 0.001
9	0.010	Iridaceae vs the rest of Asparagales (c. 76.4 Mya)	110	0.001
10	0.008	Polytomy across Fabales, Rosales, Cucurbitales, Fagales (c. 87 Mya)	16	0.013
11	0.008	<i>Trifolium</i> spp. vs rest of Fabaceae (c. 36 Mya)	31	0.007
12	0.007	Acorales vs rest of the monocots (c. 144 Mya)	408	< 0.001
13	0.006	Polytomy at the base of core eudicots (c. 117 Mya)	8	0.019
14	0.006	Xyridaceae vs rest of Poales (c. 42 Mya)	2	0.019
15	0.006	Arecaceae vs Poales, Commelinales, Zingiberales (c. 95 Mya)	9	0.111
16	0.006	Polytomy at base of Saxifragales (c. 111 Mya)	32	0.007
17	0.005	Ranunculales vs other eudicots (c. 137 Mya)	6	0.021
18	0.005	<i>Oryza</i> spp. vs rest of Poaceae (c. 11 Mya)	59	0.003
19	0.004	Divergence at the base of the robinoids; Fabaceae (c. 48 Mya)	69	0.002
20	0.004	Basal divergence in Fabaceae (c. 56 Mya)	19	0.011

Table 5 The 20 divergences making the largest contribution to present-day 1Cx DNA content variation (ranked 1–20) with accompanying contribution to seed mass variation explained by these nodes (seed mass rank)

Rank	1Cx DNA contribution	Divergences making the largest contribution	Seed mass Rank	Seed mass contribution
1	0.670	Angiosperm vs gymnosperm (c. 325 Mya)	1	0.280
2	0.060	Monocots vs the rest of the angiosperms (c. 154 Mya)	161	< 0.001
3	0.039	Polytomy at the base of Poales (c. 72 Mya)	58	0.002
4	0.024	Iridaceae vs the rest of Asparagales (c. 76 Mya)	29	0.006
5	0.021	Alstroemeriaceae vs rest of Liliales (c. 125 Mya)	238	< 0.001
6	0.017	Polytomy at base of core eudicots (c. 117 Mya)	3	0.052
7	0.013	<i>Brachypodium</i> spp. vs rest of the poidids; Poaceae (c. 8.3 Mya)	212	< 0.001
8	0.012	<i>Trifolium</i> spp. vs rest of Fabaceae (c. 36 Mya)	17	0.009
9	0.008	Xyridaceae vs rest of Poales (c. 42 Mya)	8	0.024
10	0.008	Divergence near the base of <i>Pinus</i> ; Pinaceae (c. 141 Mya)	40	0.004
11	0.008	Divergence near the base of Poaceae (c. 11 Mya)	175	< 0.001
12	0.007	Saxifragales vs Vitales and the rosids (c. 114 Mya)	36	0.004
13	0.007	Asparagales vs commelinids (c. 116 Mya)	5	0.029
14	0.007	Arecaceae vs Poales, Commelinales (c. 95 Mya)	2	0.100
15	0.006	Ranunculales vs rest of eudicots (c. 137 Mya)	7	0.026
16	0.006	Polytomy at the base of Saxifragales (c. 111 Mya)	26	0.006
17	0.006	Polytomy across Fabales, Rosales, Cucurbitales, Fagales (c. 87 Mya)	13	0.012
18	0.004	Divergence at the base of the robinoids; Fabaceae (c. 48 Mya)	57	0.002
19	0.003	Polytomy at the base of Ranunculaceae (c. 55 Mya)	144	< 0.001
20	0.003	<i>Medicago</i> spp. vs the rest of vicoids; Fabaceae (c. 39 Mya)	87	0.001

distinction. However, small genome angiosperms also have the smallest seeds. Again, it appears that genome size may set a minimum seed mass, that increases with increasing genome size, but the maximum seed mass for any given genome size may be determined by other factors. Plant height and seed mass are coordinated life-history traits (Moles *et al.*, 2005b,c), and this coordinated life history variation may work in opposition

to the developmental constraint imposed by genome size on seed mass. Also, large-seeded angiosperms may do best in environments that are less favorable to large genome species (Knight & Ackerly, 2002).

Analysis using independent contrasts showed that divergences in genome size are positively correlated with divergences in seed mass across all species. This result was driven primarily by

significant independent contrast results within angiosperms; independent contrast results were not significant within gymnosperms alone. The discrepancy in these results could be explained if seed mass scales with polyploidy, as several investigations have shown (Stebbins, 1971; Halloran & Pennell, 1982; Van Dijk & Van Delden, 1990; Bretagnolle *et al.*, 1995). Polyploidy is common in angiosperms (Stebbins, 1950, 1971; Wendel, 2000), whereas in gymnosperms it is uncommon (Delevoryas, 1980; Otto & Whitton, 2000). Within monocots, for example, we found strong correlated evolution between genome size and seed mass (Table S2). It has been suggested that most, if not all, monocots are either current polyploids or re-diploidized ancient polyploids (Goldblatt, 1980).

Not only have there been correlated divergences, but divergences that contribute significantly to extant seed mass variation also contributed significantly to extant genome size variation – further strengthening the association between these traits. The split between angiosperms and gymnosperms held the most explanatory power for present-day variation in genome size and seed mass. Interestingly, across both the 2C and 1Cx datasets, there were more nodes within monocots that held high contribution index scores (i.e. were important for present-day genome size and seed mass variation; Fig. 6). Again, this may be explained by frequent polyploidy within the evolutionary history of monocots (Goldblatt, 1980). The divergence of the large-seeded palms from the mostly small-to-medium seeded commelinids (a divergence previously shown to contribute significantly to seed mass evolution; Moles *et al.*, 2005a) ranked high in explaining present-day genome size variation (Tables 4, 5). In eudicots, the earliest divergence, the split between Ranunculales and the rest of the eudicots, also ranked high in explaining both present-day 2C DNA content and seed mass variation (Tables 4, 5).

Our regression results for angiosperms contradict previous studies (Table 1). However, when analyzing the data using independent contrasts, divergences between the two traits were significantly correlated (Table 2; Fig. 4a,b). In the history of genome size and seed mass evolution, the average divergence width has remained relatively constant, but the largest divergences in both traits have occurred more recently in geologic time (Fig. 5). Therefore, within angiosperms, deep nodes seldom show correlated evolution and the relationship between genome size and seed mass within presumably more recent diversifications at the order and family levels drove our independent contrast results (Tables 4, 5). This is consistent with knowledge of the fossil record, where the first half of angiosperm existence was marked by a relatively long period of stasis (*c.* 85 Mya) in seed mass, followed by a gradual diversification before the Cretaceous–Tertiary boundary (Tiffney, 1984; Eriksson *et al.*, 2000). These results may also reflect the propensity of polyploidy in speciation, if polyploidy leads to larger seed mass (Bretagnolle *et al.*, 1995; Wendel, 2000).

Gymnosperms with smaller genomes, relative to the rest of the group, have a range of seed mass; however, species with

larger genomes have increasingly larger seeds (Fig. 3). Our independent contrast results indicated that gymnosperm divergences in 1Cx DNA were weakly correlated with divergences in seed mass; however, there was not a significant correlation between divergences in 2C DNA and seed mass. These results are based on a gymnosperm phylogeny with uncertainty at the base, and no significant result was obtained for either the 1Cx or 2C DNA datasets when using any of the three published alternate basal resolutions (Table 3). Overall, our gymnosperm results suggest there is a high degree of phylogenetic signal within this group. In contrast to angiosperms, divergences deep in the gymnosperm phylogeny drove the regression analysis and subsequent divergences between the two traits are comparatively small (Table 5; Fig. 6). However, when examining the relationship within Pinales, the most representative group among our gymnosperm sample, 1Cx DNA content and seed mass did show significant correlated evolution. Further examination within Pinales showed this result was driven by divergences within Pinaceae, which is largely determined by the significant relationship within *Pinus* (Table S2). The significant correlated evolution found within *Pinus* is consistent with results of Grotkopp *et al.* (2004), but our results are slightly weaker. Although we used the same phylogenetic tree, we had a smaller sample size (51 vs 83) and included age-estimated branch lengths (Grotkopp *et al.* (2004) set all branches equal to 1). Despite significance across all three taxonomic levels (order, family and genus), there is strong phylogenetic signal within Pinales. The clear discrepancy in the strength of the regression and independent contrast results (Table S2) can be traced to the inclusion of large influential divergences deep within Pinales, but also to a more recent divergence at the base of *Pinus* (Table 5, Fig. 6).

Our results show that 1Cx DNA content holds greater explanatory power than 2C DNA content. The monoploid genome (1Cx DNA content) explained 6.2% of the variation in seed mass across all seed plants. This was more than the variation explained by seed dispersal syndrome, which was reported at 2.7% by Moles *et al.* (2005b) and also ranked as the second most important factor for seed mass evolution (changes in growth form held the greatest explanatory power; Table 6). Therefore 1Cx DNA content may be a driver of genome size/seed mass correlated evolution. 1Cx DNA content has been shown to have greater explanatory power than 2C DNA content in a number of studies, including both meiosis (Bennett, 1971) and cell cycle duration (Shuter *et al.*, 1983). The basis for this 1Cx effect is puzzling because it challenges our mechanistic hypothesis for bulk DNA content effects. Perhaps, it is not the quantity of DNA that matters but the basal monoploid genome size.

However, this result may have a statistical explanation. Central to independent contrast statistics is the estimation of ancestral states from descendent values. The use of 2C DNA content, which is irrespective of the level ploidy, to reconstruct ancestral trait values may lead to misinterpretations

Table 6 Rank of predictor variables for seed mass variation within seed plants (ranked by r^2)

Rank	Variable	Variation explained (r^2)
1	Growth form	9.7%
2	1Cx DNA content	6.2%
3	2C DNA content	3.3%
4	Dispersal syndrome	2.7%
5	Leaf area index	1.6%
6	Net primary production	1.2%
7	Precipitation	0.8%
8	Latitude	0.7%
9	Temperature	0.3%

because 2C values include both diploid and polyploid species. If reconstruction of ancestral nodes were based on diploid species, ancestral nodes would never be greater (polyploid) than the average of the descendant species. Conversely, estimating 2C DNA content from polyploid descendants would result in the inflation of a diploid progenitor. The cumulative effects of many overestimations and underestimations could have had an influence on the decrease in the variation explained by 2C DNA content. Nevertheless, this does not eliminate any possible biological implications for the differing results for 1Cx and 2C DNA content. Therefore, we advocate using both 1Cx and 2C DNA content when testing for genome-size dependent relationships, and a continued search for a mechanistic model to explain why 1Cx is important.

We used a suite of comparative methods to uncover a significant evolutionary association between genome size and seed mass. Because seed mass has well described ecological effects, the relationship between genome size and seed mass perhaps represents a genotype/phenotype/selection relationship that does not involve genes *per se* (however, specific genes may influence seed mass in concert). Further analysis should focus on testing for a more direct role genome size plays in seed mass variation. This should involve examining cell sizes in specific seed organs (i.e. embryo, endosperm, cotyledons) that may be affected by genome size. In addition, perhaps genome size is related to other life-history traits (i.e. seed dispersal syndrome, growth form, plant height; Moles *et al.*, 2005b) that are also shown to be important in the evolution of seed mass. Answering these questions requires a continued effort to join plant functional trait databases (Seed Information Database, Flynn *et al.*, 2004; Glopnet, Wright *et al.*, 2004) with the Plant DNA C-values database (Bennett & Leitch, 2005). This effort is important for uncovering other physiological, ecological, and evolutionary associations with the profound variation in plant genome size.

Acknowledgements

We thank J. Connolly, D. Ackerly, and two anonymous reviewers for comments that greatly improved the manuscript. JBD

would like to thank R. Smith, J. Tweddle, and S. Flynn for their contributions to the Seed Information Database. The Millennium Seed Bank Project is funded by the UK Millennium Commission, The Wellcome Trust and Orange plc. Royal Botanic Gardens, Kew is partially funded by the UK Department for Environment, Food and Rural Affairs.

References

- Ackerly DD. 2006. *Analysis of traits (AOT): a module of Phylocom*, Version 3.1. <http://www.phylodiversity.net/phylocom>.
- Baetcke KP, Sparrow AH, Naumann CH, Schwemme SS. 1967. The relationship of DNA content to nuclear and chromosome volumes and to radiosensitivity (LD50). *Proceedings of the National Academy of Sciences, USA* 58: 533–540.
- Bell CD, Soltis DE, Soltis PS. 2005. The age of the angiosperms: a molecular timescale without a clock. *Evolution* 59: 1245–1258.
- Bennett MD. 1971. The duration of meiosis. *Proceedings of the Royal Society of London, Series B* 178: 277–299.
- Bennett MD. 1972. Nuclear DNA content and minimum generation time in herbaceous plants. *Proceedings of the Royal Society of London, Series B* 181: 109–135.
- Bennett MD. 1973. Nuclear characters in plants. *Brookhaven Symposia in Biology* 25: 344–366.
- Bennett MD, Heslop-Harrison JS, Smith JB, Ward JP. 1983. DNA density in mitotic and meiotic metaphase chromosomes of plants and animals. *Journal of Cell Science* 63: 173–179.
- Bennett MD, Leitch IJ. 2005. Plant DNA C-values database (release 4.0, October 2005). [<http://www.kew.org/genomesize/homepage>.]
- Bennetzen JL. 2002. Mechanisms and rates of genome expression and contraction in flowering plants. *Genetica* 115: 29–36.
- Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Annals of Botany (London)* 95: 127–132.
- Bretagnolle F, Thompson JD, Lumaret R. 1995. The influence of seed size variation on seed germination and seedling vigour in diploid and tetraploid *Dactylis glomerata* L. *Annals of Botany (London)* 76: 607–615.
- Burleigh JG, Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany* 91: 1599–1613.
- Caceres ME, De Pace C, Scarascia Mugnozza GT, Kotsonis P, Ceccarelli M, Cionini PG. 1998. Genome size variation within *Dasypyrum villosum*: correlation with chromosomal traits, environmental factors and plant phenotypic characteristics and behavior in reproduction. *Theoretical and Applied Genetics* 96: 559–567.
- Cade BS, Noon BR. 2003. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* 1: 412–420.
- Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proceedings of the National Academy of Sciences, USA* 97: 4086–4091.
- Chung J, Lee JH, Arumuganathan K, Graef GL, Specht JE. 1998. Relationships between nuclear DNA content and seed and leaf size in Soybeans. *Theoretical and Applied Genetics* 96: 1064–1068.
- Delevoryas T. 1980. Polyploidy in gymnosperms. In: Lewis WH, ed. *Polyploidy: biological relevance*. New York, NY, USA: Plenum Press, 215–218.
- R Development Core Team. 2005. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. [<http://www.Royal-project.org>.]
- Donoghue MJ, Doyle JA. 2000. Seed plant phylogeny: demise of the anthophyte hypothesis? *Current Biology* 10: R106–R109.
- Edwards GA, Endrizzi JL. 1975. Cell size nuclear size and DNA content relationships in *Gossypium*. *Canadian Journal of Genetics Cytology* 17: 181–186.

- Eriksson O, Friis EM, Lofgren P. 2000. Seed size, fruit size, and dispersal systems in angiosperms from the early Cretaceous to the late Tertiary. *American Naturalist* 156: 47–58.
- Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* 125: 1–15.
- Flynn S, Turner RM, Dickie JB. 2004. Seed Information Database (release 6.0, October 2004), [<http://www.rbgkew.org.uk/data/sid/>].
- Friedman WE. 2006. Embryological evidence for developmental lability during early angiosperm evolution. *Nature* 441: 337–340.
- Garland T Jr, Harvey PH, Ives AR. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* 42: 265–292.
- Goldblatt P. 1980. Polyploidy in angiosperms: monocotyledons. In: Lewis WH, ed. *Polyploidy: biological relevance*. New York, NY, USA: Plenum Press, 219–236.
- Greilhuber J, Doležel J, Lysák MA, Bennett MD. 2005. The origin, evolution, and proposed stabilization of terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Annals of Botany (London)* 95: 255–260.
- Grime JP, Thompson K, Hunt R, Hodgson JG, Cornelissen JHC, Rorison IH, Hendry GAF, Ashenden TW, Askew AP, Band SR, Booth RE, Bossard CC, Campbell BD, Cooper JEL, Davison AW, Gupta PL, Hall W, Hand DW, Hannah MA, Hillier SH *et al.* 1997. Integrated screening validates primary axes of specialization in plants. *Oikos* 79: 259–281.
- Grotkopp E, Rejmanek M, Sanderson MJ, Rost TL. 2004. Evolution of genome size in Pines (*Pinus*) and its life-history correlates: supertree analyses. *Evolution* 58: 1705–1729.
- Halloran GM, Pennell AL. 1982. Grain size and seedling growth of wheat at different ploidy levels. *Annals of Botany (London)* 49: 103–113.
- Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ. 2005. Evolution of genome size in Brassicaceae. *Annals of Botany (London)* 95: 229–235.
- Jones RN, Brown LM. 1976. Chromosome evolution and DNA variation in *Crepis*. *Heredity* 36: 91–104.
- Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ. 2002. *Plant systematics: a phylogenetic approach*, 2nd edn. Sunderland, MA, USA: Sinauer Associates.
- Kidwell MJ. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115: 49–63.
- Knight CA, Ackerly DD. 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters* 5: 66–76.
- Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology, and phenotype. *Annals of Botany (London)* 95: 177–190.
- Koener R. 2005. *Quantreg: quantile regression. R package version 3.84*. [<http://www.econ.uiuc.edu/~roger/research/rq/rq.html>].
- Koener R, Hallock KF. 2001. Quantile regression: an introduction. *Journal of Economic Perspectives* 15: 143–156.
- Lawrence ME. 1985. *Senecio* L. (Asteraceae) in Australia: nuclear DNA amounts. *Australian Journal of Botany* 33: 221–232.
- Leishman MR. 1999. How well do plant traits correlate with establishment ability? Evidence from a study of 16 calcareous grassland species. *New Phytologist* 141: 487–496.
- Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82: 651–663.
- Leitch IJ, Chase MW, Bennett MD. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals of Botany (London)* 82: 85–94.
- Leitch IJDE, Soltis DEPS, Soltis PS, Bennett MD. 2005. Evolution of DNA amounts across land plants (Embryophyta). *Annals of Botany (London)* 95: 207–217.
- Maranon T, Grubb PJ. 1993. Physiological basis and ecological significance of the seed size and relative growth rate relationship in Mediterranean annuals. *Functional Ecology* 7: 591–599.
- Moles AT, Ackerly DD, Webb CO, Tweddle JC, Dickie JB, Pitman AJ, Westoby M. 2005b. Factors that shape seed mass evolution. *Proceedings of the National Academy of Sciences, USA* 102: 10540–10544.
- Moles AT, Ackerly DD, Webb CO, Tweddle JC, Dickie JB, Westoby M. 2005a. A brief history of seed size. *Science* 307: 576–580.
- Moles AT, Ackerly DD, Webb CO, Tweddle JC, Dickie JB, Westoby M. 2005c. Response to comment on 'A brief history of seed size'. *Science* 310: 783.
- Mowforth MA. 1985. Variation in nuclear DNA amounts in flowering plants: an ecological analysis. *PhD thesis*. Sheffield, UK: University of Sheffield.
- Otto SJ, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34: 401–437.
- Pagel MD. 1992. A method for the analysis of comparative data. *Journal of Theoretical Biology* 156: 431–442.
- Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* 287: 1060–1062.
- Price HJ, Dillon SL, Hodnett G, Rooney WL, Ross L, Johnston JS. 2005. Genome size evolution in the genus *Sorghum* (Poaceae). *Annals of Botany (London)* 95: 219–227.
- Rees H, Cameron FM, Hararika MH, Jones GH. 1966. Nuclear variation between diploid angiosperms. *Nature* 211: 828–830.
- Shuter BJ, Thomas JE, Taylor WD, Zimmerman AM. 1983. Phenotypic correlated of genomic DNA content in unicellular eukaryotes and other cells. *American Naturalist* 122: 26–44.
- Soltis DE, Soltis PS, Bennett MD, Leitch IJ. 2003. Evolution of genome size in the angiosperms. *American Journal of Botany* 89: 1670–1681.
- Soltis DE, Soltis PS, Zanis MJ. 2002. Phylogeny of seed plants based on evidence from eight genes. *American Journal of Botany* 89: 1670–1681.
- Stebbins GL. 1950. *Variation and evolution in plants*. New York, NY, USA: Columbia University Press.
- Stebbins GL. 1971. *Chromosomal evolution in higher plants*. Reading, MA, USA: Addison-Wesley.
- Stevens PF. 2005. *Angiosperm phylogeny website, Version 6. May 2005*. [<http://www.mobot.org/MOBOT/research/APweb/>].
- Sugiyama S. 2005. Developmental basis of interspecific differences in leaf size and specific leaf area among C3 grass species. *Functional Ecology* 19: 916–924.
- Thompson K. 1990. Genome size, seed size and germination temperature in herbaceous angiosperms. *Evolutionary Trends in Plants* 4: 113–116.
- Tiffney BH. 1984. Seed size, dispersal syndromes, and the rise of angiosperms: evidence and hypothesis. *Annals of the Missouri Botanical Gardens* 71: 551–576.
- Van Dijk P, Van Delden W. 1990. Evidence for autotetraploidy in *Plantago media* and comparisons between natural and artificial cytotypes concerning cell size and fertility. *Heredity* 65: 349–357.
- Webb CO, Ackerly DD, Kembel SW. 2006. *Phylocom: software for the analysis of community phylogenetic structure and character evolution, Version 3.40*. [<http://www.phylodiversity.net/phylocom/>].
- Webb CO, Donoghue M. 2005. *Phylocomatic, a database for applied phylogenetics*. [<http://www.phylodiversity.net/phylocomatic.Treeversion:R20040402/>].
- Wendel J. 2000. Genome evolution in polyploids. *Plant Molecular Biology* 42: 225–249.
- Wikström N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London, Series B* 268: 2211–2220.
- Wright IJ, Reich PB, Westoby M, Ackerly DD, Baruch Z, Bongers F,

Cavender-Bares J, Chapin FS, Cornelissen JHC, Diemer M, Flexas J, Garnier E, Groom PK, Gulias J, Hikosaka K, Lamont BB, Lee T, Lee W, Lusk C, Midgley JJ *et al.* 2004. The world-wide leaf economics spectrum. *Nature* 428: 821–827.

Table S2 Results for the regression and independent contrast analyses of well-represented families and genera for both 2C DNA and 1Cx DNA content with seed mass

Supplementary Material

The following supplementary material is available for this article online:

Table S1 Sources of systematic data used to resolve relationships within families, following the construction of the family-level tree by Phylomatic (Webb & Donoghue, 2005; see the Materials and Methods section)