the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **74,** 829–839 doi: 10.1111/tpj.12168

Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences

Simon Renny-Byfield[1], Ales Kovarik[2], Laura J. Kelly[1,3], Jiri Macas[4], Petr Novak[4], Mark W. Chase[3], Richard A. Nichols[1], Mahesh R. Pancholi[1], Marie-Angele Grandbastien[5] and Andrew R. Leitch[1,]*

[1]*Queen Mary University of London, School of Biological and Chemical Sciences, Mile End Road, London, E1 4NS, UK,*
[2]*Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i. Kralovopolska 135, CZ-61265, Brno, Czech Republic,*
[3]*Jodrell Laboratory, Royal Botanic Gardens, Richmond, Surrey, TW9 3DS, UK,*
[4]*Biology Centre ASCR, Institute of Plant Molecular Biology, Branišovská 31, České Budějovice, CZ–37005, Czech Republic, and*
[5]*Institute Jean-Pierre Bourgin, Institut National de la Recherche Agronomique, Versailles, 78026, France*

Received 19 January 2013; revised 23 February 2013; accepted 1 March 2013; published online 21 March 2013.

*For correspondence (e-mail a.r.leitch@qmul.ac.uk).

SUMMARY

Recent advances have highlighted the ubiquity of whole-genome duplication (polyploidy) in angiosperms, although subsequent genome size change and diploidization (returning to a diploid-like condition) are poorly understood. An excellent system to assess these processes is provided by *Nicotiana* section *Repandae*, which arose via allopolyploidy (approximately 5 million years ago) involving relatives of *Nicotiana sylvestris* and *Nicotiana obtusifolia*. Subsequent speciation in *Repandae* has resulted in allotetraploids with divergent genome sizes, including *Nicotiana repanda* and *Nicotiana nudicaulis* studied here, which have an estimated 23.6% genome expansion and 19.2% genome contraction from the early polyploid, respectively. Graph-based clustering of next-generation sequence data enabled assessment of the global genome composition of these allotetraploids and their diploid progenitors. Unexpectedly, in both allotetraploids, over 85% of sequence clusters (repetitive DNA families) had a lower abundance than predicted from their diploid relatives; a trend seen particularly in low-copy repeats. The loss of high-copy sequences predominantly accounts for the genome downsizing in *N. nudicaulis*. In contrast, *N. repanda* shows expansion of clusters already inherited in high copy number (mostly chromovirus-like Ty3/*Gypsy* retroelements and some low-complexity sequences), leading to much of the genome upsizing predicted. We suggest that the differential dynamics of low- and high-copy sequences reveal two genomic processes that occur subsequent to allopolyploidy. The loss of low-copy sequences, common to both allopolyploids, may reflect genome diploidization, a process that also involves loss of duplicate copies of genes and upstream regulators. In contrast, genome size divergence between allopolyploids is manifested through differential accumulation and/or deletion of high-copy-number sequences.

Keywords: next-generation sequencing, allopolyploidy, genome downsizing, SRA045794, SRA051392, *Nicotiana repanda*, *Nicotiana nudicaulis*, *Nicotiana sylvestris*, *Nicotiana obtusifolia*.

INTRODUCTION

All angiosperms have experienced at least one, if not more, rounds of whole-genome duplication (WGD or polyploidy) in their ancestry (Vision *et al.*, 2000; Bowers *et al.*, 2003; Jaillon *et al.*, 2007; Barker *et al.*, 2009; Jiao *et al.*, 2011). Subsequent to polyploidy, genomes undergo a process of diploidization, whereby duplicate copies of genes may be lost and chromosome number may decrease, so that over time the signature of ancestral polyploidy becomes more and more obscure. Although global analyses of genome size in angiosperms reveal a trend towards DNA loss subsequent to polyploidy (genome downsizing) (Leitch and Bennett, 2004; Leitch *et al.*, 2008), increases in genome size are also known to occur. Such changes in genome size are thought to arise via accumulation of repetitive DNA (Hawkins *et al.*, 2009; Renny-Byfield *et al.*, 2011).

Evolution and change in the copy number of repetitive DNA may be analysed via 'genome skimming', in which short-read next-generation sequencing data are used to reconstruct and quantify the repetitive fraction of the genome (Hribova *et al.* 2010, Macas *et al.*, 2011, 2007; Renny-Byfield *et al.*, 2011; Swaminathan *et al.* 2007, Wicker *et al.* 2009). The approach efficiently characterizes sequences present in multiple copies, and read-depth analysis provides estimates of repeat abundance within the genome, allowing quantification and comparisons of repeats between species.

Recent studies have used these methods to better understand genome evolution following allopolyploidy. Comparisons of the young allotetraploid *Nicotiana tabacum* (<0.2 million years ago; Clarkson *et al.*, 2005) with its diploid progenitors revealed a bias towards removal of paternally derived repeats in conjunction with a reduction in the repetitive fraction of the genome (Renny-Byfield *et al.*, 2011). Patterns of sequence loss observed in *N. tabacum* are repeated in synthetic lines after only four generations (Renny-Byfield *et al.*, 2012). The loss of DNA in synthetic (Petit *et al.*, 2010) and natural (Petit *et al.*, 2007) tobacco is also revealed using sequence-specific amplified polymorphisms (SSAPs) targeting Tnt1 and Tnt2 retroelement families. Likewise, repetitive DNA loss targeted to chromosomes and genomes has been reported in wheat (*Triticum aestivum*) (Ozkan *et al.*, 2001; Salina *et al.*, 2004).

Nicotiana section *Repandae* provides an ideal model group for dissecting the two phenomena of genome size divergence and diploidization following ancient polyploidy (approximately 5 million years ago). It is thought that *Repandae* formed from a single hybridization event between relatives of extant *Nicotiana sylvestris* (2637 Mbp/1C) and *Nicotiana obtusifolia* (1511 Mbp/1C), which, following speciation, produced four allopolyploids (Chase *et al.*, 2003; Clarkson *et al.*, 2004, 2005, 2010; Kelly *et al.*, 2012).

In *Nicotiana* section *Repandae* (Parisod *et al.*, 2012), there is considerable departure from additivity in seven transposable element families, typically through deletion, especially of transposable elements derived from the *N. obtusifolia* parent. In addition, there are also large numbers of new SSAP bands, probably derived from new element insertion sites. However, SSAP data cover only a small fraction of the genome, targeting a small number of repeats. In this paper, we use next-generation sequencing approaches to generate a global overview of repetitive DNA evolution in the context of allopolyploidy. We examine two species in section *Repandae*, *Nicotiana repanda* (5320 Mbp/1C) and *Nicotiana nudicaulis* (3477 Mbp/1C) (Leitch *et al.*, 2008), to better understand the processes of diploidization and genome size change. We used genomic *in situ* hybridization (GISH), next-generation sequencing and a graph-based clustering pipeline to simultaneously identify, quantify and assess thousands of repeat families in the genomes of *N. repanda* (with genome upsizing) and *N. nudicaulis* (with genome downsizing), and compare these with repeats in close relatives of their diploid progenitors. Thus, this paper dissects the signatures of two phenomena: diploidization and genome size divergence.

RESULTS

Reconstructing ancestral genome size

We estimated genome size changes in lineages leading to *N. nudicaulis* and *N. repanda* by reconstructing ancestral genome sizes in *Nicotiana* section *Repandae* using Markov chain Monte Carlo reconstruction methods (Figure 1). Ancestral genome size estimates are in good agreement using sequence data from both allopolyploid sub-genomes (Figure 1), and indicate an increase in genome size in the lineages leading to *N. repanda* (+23.6%), *Nicotiana nesophila* (+14.7%) and *Nicotiana stocktonii* (+14.3%), but downsizing in the lineage leading to *N. nudicaulis* (−19.2%; Figure 1). These values are similar to a model that assumes simple additivity of genome sizes recorded in extant relatives of the diploid progenitors (Leitch *et al.*, 2008).

Clustering

Graph-based clustering was used to characterize, quantify and compare highly repetitive DNA sequences in the genomes of the diploid species *N. sylvestris* and *N. obtusi-*

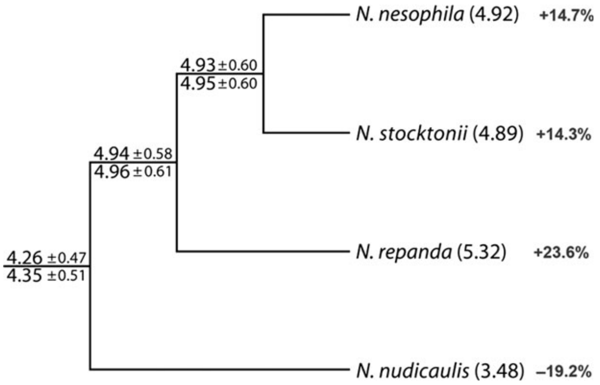

Figure 1. Reconstruction of ancestral genome size in *Nicotiana* section *Repandae*.
Tree summarizing results from Bayesian phylogenetic analysis of sequence data from separate parental sub-genomes (data from both sub-genomes yield identical highly supported topologies for section *Repandae*; see Figure S2). Genome sizes for extant species (1C values in Gb, taken from Leitch *et al.*, 2008) are given in parentheses after species names. Ancestral genome sizes reconstructed using BayesTraits are shown for internal nodes; values (means ± SD) above branches are those estimated using trees from Bayesian analysis of the *N. sylvestris*-like sub-genome, and the values below branches were estimated using trees from Bayesian analysis of the *N. obtusifolia*-like sub-genome. The percentage change in genome size is indicated in grey, and is calculated using the mean of the two estimates of ancestral genome size for the common ancestor of section *Repandae*.

folia and their derived allotetraploids *N. repanda* and *N. nudicaulis*. Clustering of 6 812 631 Illumina reads, each 95 bp long (equivalent to 5% coverage of each genome), produced 492 696 clusters, with the smallest comprising two reads and the largest comprising more than 80 000 reads. The majority (79%) of the largest clusters (defined as having more than ten reads from either or both of the progenitors) included reads from at least one allopolyploid. Furthermore, 66% of clusters contained reads from both allopolyploids. However, a small number of clusters were derived from only one of the four species included in the analysis.

Clusters correspond to families of repetitive DNA that may be assessed for similarity to known repeats as well as abundance in each genome. The largest cluster (CL1) comprised 81 004 reads, and sequence similarity searches indicate it is derived from a Ty3/*Gypsy* retroelement (examples of the resulting 3D networks produced from the clustering algorithm are shown in Data S1). Reads within the largest clusters were assembled into contiguous sequences using CAP3 (Huang and Madan, 1999) on a cluster-by-cluster basis. Resulting contigs ranged from 95 bp (the minimum possible, indicating complete overlap of more than one read) to several thousand nucleotides in length.

Genome characterization

To investigate the nature of repetitive DNA in each species, we annotated repeat clusters using similarity to known repetitive DNA [using a RepBase library (Jurka *et al*., 2005) and RepeatMasker (Smit *et al*., 2010)]. Of the clusters we could identify as a known repeat, the majority were retroelements, contributing between 29.95–38.30% of the genome depending on the species (Table 1). Most of the retroelements were Ty3/*Gypsy*-like (between 24.31 and 33.32% of the genome), with Ty1/*Copia*-like elements being less abundant (between 3.31 and 4.71% of the genome;

Table 1 Genome characterization in *Nicotiana* section *Repandae*

	N. sylvestris		*N. obtusifolia*		Expected		*N.repanda*		*N.nudicaulis*	
Description	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)
Retroelements	50 263 645	38.14	22 628 905	29.95	72 892 550	35.15	101 887 405	38.30	65 246 950	37.53
LTR/Gypsy	43 505 725	33.01	18 369 580	24.31	61 875 305	29.84	88 629 680	33.32	53 975 105	31.05
LTR/Copia	5 589 515	4.24	2 610 315	3.46	8 199 830	3.95	8 802 985	3.31	8 196 600	4.71
LTR/Caulimovirus	224 295	0.17	609 710	0.81	834 005	0.4	487 920	0.18	572 565	0.33
LINE/L1	36 860	0.03	728 365	0.96	765 225	0.37	242 535	0.09	626 620	0.36
LINE/Penelope	144 590	0.11	45 220	0.06	189 810	0.09	140 885	0.05	144 970	0.08
LINE/RTE-BovB	537 605	0.41	151 050	0.2	688 655	0.33	2 583 050	0.97	1 124 420	0.65
SINE	5700	0	1235	0	6935	0	1425	0	3135	0
SINE/tRNA	219 355	0.17	113 430	0.15	332 785	0.16	998 925	0.38	603 535	0.35
DNA transposons	2 617 725	1.99	900 030	1.19	3 517 755	1.70	3 105 455	1.17	2490 235	1.43
DNA/CMC-EnSpm	102 790	0.08	373 160	0.49	475 950	0.23	861 365	0.32	711 360	0.41
DNA/hAT-Ac	359 100	0.27	469 870	0.62	828 970	0.4	1 395 360	0.52	906 110	0.52
DNA/hAT-Tag1	570	0	665	0	1235	0	1615	0	855	0
DNA/hAT-Tip100	21 565	0.02	7600	0.01	29 165	0.01	20 520	0.01	20 900	0.01
DNA/MULE-MuDR	9310	0.01	1710	0	11 020	0.01	312 645	0.12	49 685	0.03
DNA/PIF-Harbinger	1330	0	8170	0.01	9500	0	4085	0	5510	0
DNA/TcMar-Pogo	0	0	14 820	0.02	14 820	0.01	30 305	0.01	24 035	0.01
DNA/TcMar-Stowaway	2 123 060	1.61	24 035	0.03	2 147 095	1.04	479 560	0.18	771 780	0.44
Others										
RC/Helitron	2375	0	760	0	3135	0	17 100	0.01	3325	0
rRNA	1 290 765	0.98	454 955	0.6	1 745 720	0.84	1 143 135	0.43	866 590	0.5
Satellite	945 630	0.72	156 8355	2.08	2 513 985	1.21	6 629 575	2.49	5 828 725	3.35
Simple repeat	4 909 220	3.72	2 397 230	3.17	7 306 450	3.52	6 644 110	2.5	5 230 035	3.01
Low complexity	7 060 970	5.36	2 214 165	2.93	9 275 135	4.47	26 595 060	10.00	9 432 550	5.43
Unknown[a]	23 804 055	18.06	11 411 495	15.1	35 215 550	16.98	39 628 015	14.9	23 593 915	13.57
Small clusters[b]	23 894 305	18.13	19 168 245	25.37	43 062 550	20.77	43 525 105	16.36	32 372 390	18.62
Singletons	17 011 270	12.91	14 805 750	19.6	31 817 020	15.34	36 825 040	13.84	28 785 285	16.56
Total	131 799 960	100	75 549 890	100	207 349 850	100	266 000 000	100	173 850 000	100

[a] These clusters returned no matches to RepBase.

[b] These clusters consist of fewer than ten reads from both or either of the progenitor diploids, and were not screened against RepBase.

Comparison of the major repetitive DNA component of the genome in two allopolyploids of section *Repandae* (*N. repanda* and *N. nudicaulis*) and close relatives of their diploid progenitors (*N. sylvestris* and *N. obtusifolia*). GR, genome representation; GP, genome proportion. The expected GR is the sum of GR in the progenitors, reflecting complete additivity in a nascent allotetraploid. Expectation is also given as a percentage of the genome (GP).

Table 1). The smallest genome analysed (*N. obtusifolia*) contained the smallest proportion of retroelements, whereas *N. repanda* (the largest genome) contained the most.

Although the repetitive fraction of all four *Nicotiana* genomes is dominated by retroelements, there are low levels of several other repeat types. For example, DNA transposons were estimated to contribute between 1.19 and 1.99% of the genome, with *N. obtusifolia* having the smallest genomic fraction and *N. sylvestris* the largest. We also identified a number of long interspersed elements (LINEs), short interspersed elements (SINEs), low-complexity and satellite repeat families in the dataset, but these showed low abundance in all four genomes (Table 1).

Comparing observed with expected values in allotetraploids

We compared the expected abundance of repeat clusters, assuming additivity with the parents, with the abundance observed in the two allotetraploid species (Figures 2–4). Of 3480 clusters where the expected number of reads was $\geq$ 10, 3170 and 3119 were found to have fewer reads than expected for *N. repanda* and *N. nudicaulis*, respectively. Many clusters (2997) were under-represented in both allotetraploids. For the majority of these (1605 clusters) under-representation was greatest in *N. repanda,* whereas only 515 clusters were most under-represented in *N. nudicaulis*. The remaining 877 clusters were equally under-represented in both allopolyploids. In contrast, 296 and 330 clusters were over-represented in *N. repanda* and *N. nudicaulis*, respectively.

We analysed all clusters for which we predicted fewer than ten reads in the allopolyploids (based on progenitor additivity). This analysis revealed an overall excess of 4869 reads derived from *N. repanda*. This small number represents a limited contribution to genome size change in this group of clusters. In comparison, *N. nudicaulis* exhibited a deficit of 112 528 reads among clusters predicted to contain fewer than ten reads.

A small number of clusters, with few or no reads derived from the diploids, are abundant in the allopolyploids. These are CL67, CL77, CL85, CL118, CL174 and CL237. Of these clusters, CL77 contributes the most to genome size change. For this cluster, we expected to see fewer than ten reads, but observed 13 520 and 5872 reads in *N. repanda* and *N. nudicaulis*, respectively. This cluster may not be classified into a repeat type as it returned no hits to repeat databases.

Overall, *N. repanda* had a higher than expected repeat abundance: we classified 2 412 368 reads (229 174 960 bp) into repeat clusters, rather than the predicted 1 847 714 reads (175 532 830 bp). As we have analysed 5% of the genome, these additional reads equate to a 25.87% increase in genome size over that expected (4147 Mbp/

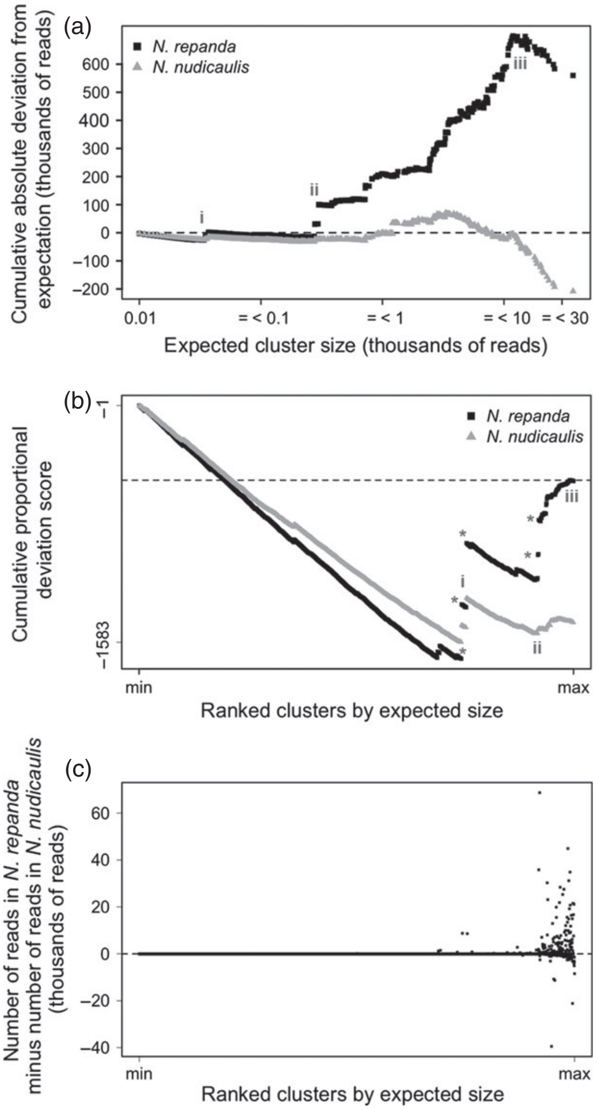

Figure 2. Deviation from expectation in two allopolyploids.
(a) Graph showing how clusters in the two allopolyploids deviate from their expected size (cumulative change over the range of cluster sizes). The marked positions (i–iii) represent apparent transitions in the profile of the graph (see Results).
(b) The cumulative proportional deviation score reflects the proportional change in cluster size from expectation [calculated as (observed/expected) – 1], and is plotted against the range of cluster sizes. The proportional deviation score accounts for any effect of cluster size. A negative slope reflects a trend towards a reduction in cluster size and a positive slope indicates the reverse. Indicators (i–iii) correspond to the same clusters as identified in (a). Note that between positions (ii) and (iii) for large clusters, the slope of the graph is strongly positive in *N. repanda*.
(c) Scatter plot indicating the difference in read numbers between *N. repanda* and *N. nudicaulis* within each cluster. Clusters are ranked by size. A data point below zero indicates higher abundance in *N. nudicaulis*, whereas a data point above zero indicate a higher abundance in *N. repanda*. Note that the *x* axis in (a) is a log scale of expected cluster size, whereas in (b) and (c), it is the rank of expected size.

1C). However, in *N. nudicaulis*, the total number of clustered reads was 1 526 997 (145 064 715 bp), which is less than expected (1 847 714 reads/175 532 830 bp),

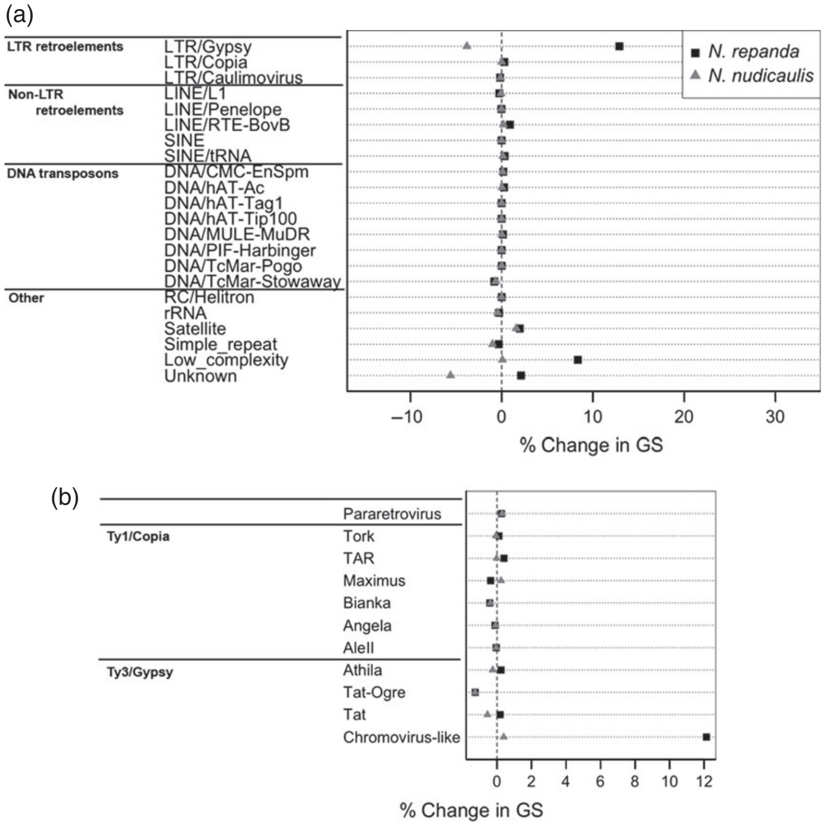

Figure 3. Repeat categories and their contribution to genome-size change in section *Repandae*.
(a) Dot chart showing the contribution of each repeat type to genome size change in the allopolyploids *N. nudicualis* and *N. repanda*. Repeat types were identified by RepeatMasker, and the corresponding abundance (number of bp) of each type compared with the sum of the progenitor diploids. Any deviation from expectation is indicated by a percentage change over the expected genome size of 4147 Mpb/1C. The vertical dashed lines indicate zero deviation.
(b) As in (a), showing sub-families of Ty1/*Copia* and Ty3/*Gypsy* retroelements.

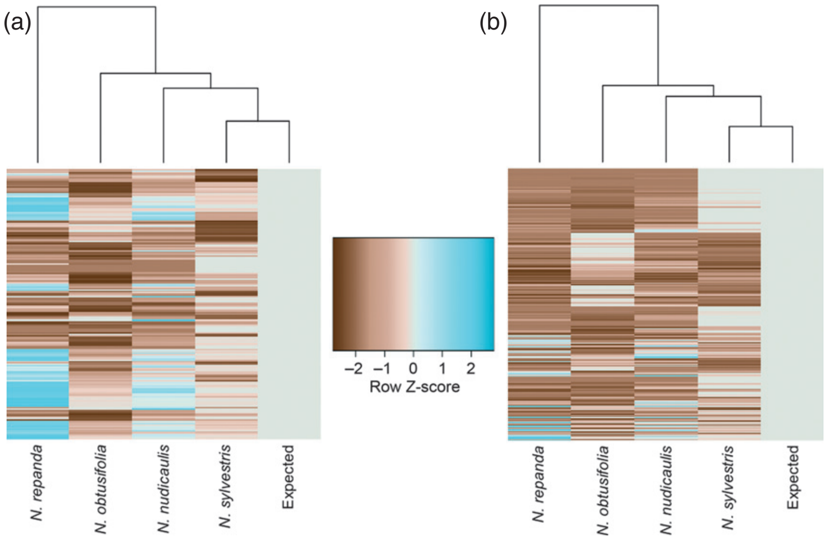

Figure 4. Comparing cluster abundance.
Heatmap analysis showing deviation from expectation for the (a) 200 and (b) 2500 clusters with the highest expected values (progenitor additivity). Deviation is normalized across clusters and represented by a Z–score (number of standard deviations away from expected). Species are grouped by dendrograms (above each panel) based on the similarity of cluster abundance. For the allopolyploids, the figure shows how the abundance of each cluster has varied from expected (brown for a cluster size decrease and blue for a cluster size increase). For the diploids, the colour shows how each parent contributes to the expected value (the deeper the brown colour, the fewer proportional reads it provides). Note that *N. nudicaulis*, *N. sylvestris* and the expected repeat abundance in the genomes of the allopolyploids form a clade of similar repeat profiles.

equivalent to a 14.69% decrease in genome size relative to expectation.

In order to determine which clusters are associated with genome upsizing in *N. repanda* and genome downsizing in *N. nudicaulis*, we ranked the clusters by size and plotted the cumulative deviation in abundance from that expected in each of the allopolyploids (Figure 2a). This revealed that clusters with low abundance are under-represented in both *N. repanda* and *N. nudicaulis* [Figure 2a, from the origin to position (i)], with both allopolyploids following a similar pattern. At position (i), there is a step change where two repeat clusters are over-represented, and thereafter the trend continues for under-representation of reads derived from low-copy-number repeats [until position (ii); Figure 2a]. For clusters with higher expected values [from position (ii); Figure 2a], the repeats are predominantly over-represented in *N. repanda*, but generally under-represented in *N. nudicaulis*. For the largest 30 clusters [from position (iii) to maximum], the number of reads in each cluster is lower than expected for both *N. repanda* and *N. nudicaulis*.

In order to remove any effect of cluster size, the data were re-analysed to obtain cumulative proportional deviation scores. This score considers the observed number of reads divided by the expected number of reads for each cluster (see Figure 2b legend). A negative score indicates that the cluster size is smaller than expected, whereas a positive score indicates a larger cluster size. This analysis reveals that there are similar proportional losses in both allopolyploids for the majority of the range in cluster size [from the smallest to position (i) in Figure 2b]. However, there are exceptions; a few clusters are over-represented in both species, resulting in a considerable change to the cumulative deviation score [e.g. positions (i) and (ii) in Figure 2b].

To assess the impact of cluster abundance on the genome size discrepancy between *N. repanda* and *N. nudicaulis*, we compared the abundance of each cluster in the allopolyploids (Figure 2c). Most clusters have minimal or no effect on genome size differentiation (as indicated by falling on or near zero on the *y* axis). However, a number of clusters, the majority of which are inherited in high copy number, have a marked effect. Most of these clusters have a higher abundance in *N. repanda* and probably account for the majority of the genome size discrepancy between the two allopolyploids. The relationship between observed and expected cluster size in the allopolyploids is visualized in Figure S1.

Sequences causing genome size divergence in *Repandae*

For both allopolyploids, we summed the abundance (number of base pairs) for each repeat type (as identified by RepeatMasker, see above and Table 1) and compared this to expectation (progenitor additivity). We show deviation as a percentage of the expected genome size (Figure 3). For the majority of repeat types, deviation from additivity is minimal, e.g. DNA transposons and rDNA sequences. For *N. nudicaulis*, repeats of unknown origin (no similarity to known repetitive sequences) are under-represented and account for approximately 6% reduction in genome size, whereas these clusters are slightly over-represented in *N. repanda*.

In *N. repanda*, there is an over-abundance of Ty3/*Gypsy*-like retroelements relative to expectation, accounting for an approximately 13% increase in genome size. All clusters containing GAG and reverse transcriptase domains were further analysed in order to ascertain which families of TY3/*Gypsy* elements are over-represented (Figure 3b). We found that most of the over-represented sequences, accounting for a 12.5% genome size increase in *N. repanda*, are chromovirus-like retroelements, while Tat-Ogre elements account for a 1% decrease in genome size. Low-complexity sequences and satellite repeats have also made a positive contribution to genome size change. For *N. nudicaulis*, there is a decrease in Ty3/*Gypsy*-like retroelements, contributing to an approximately 4% decrease in genome size.

Genome divergence in *Repandae*

To characterize the overall effect of diverging repeat abundance in the allopolyploids in comparison with the diploids, we performed a heatmap analysis, implemented in R. The heatmap reveals that most clusters are under-represented in both allopolyploids, mirroring the data displayed in Figure 3. In addition, the dendrograms in Figure 4(a,b) group species based on similarities in cluster abundance, and here the two allopolyploids differ: *N. nudicaulis* has repeat abundances that are closer to their expected values (i.e. additivity of abundance in diploids) and most similar to *N. sylvestris*, whereas *N. repanda* is a more divergent genome.

The greater similarity of *N. nudicaulis* to the diploids was further confirmed using GISH (Figure 5a), which revealed many sites of probe hybridization, including *N. obtusifolia* probe signal at sub-telomeric regions (red in Figure 5a) that may correspond to retention of a high-copy number satellite repeat called NPAL, inherited from *N. obtusifolia* (Koukalova *et al*., 2010). More uniform binding of the *N. sylvestris* probe was observed (green in Figure 5a). However, discrimination of probes is difficult, and it is not possible to resolve progenitor chromosome complements with any degree of certainty, although a few chromosomes are distinguishable as predominantly red or green.

In contrast, GISH to *N. repanda* using the same probes produced weaker binding of the *N. sylvestris* probe along most chromosomes, although there was stronger signal at sub-telomeric regions (green in Figure 5b). However, probe binding was weak compared with *N. nudicaulis*, par-

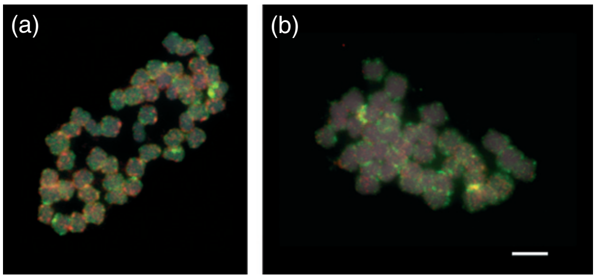

Figure 5. GISH to allopolyploids Genomic *in situ* hybridization to metaphase chromosomes of (a) *N. nudicaulis* and (b) *N. repanda* using genomic DNA probes of the progenitor species *N. obtusifolia* (red) and *N. sylvestris* (green). Chromosomes are counterstained with 4′,6–diamidino-2–phenylindole (grey). Note the overall weaker GISH signal to *N. repanda*. Scale bar = 5 μm.

ticularly for the *N. obtusifolia* probe, for which little or no signal was detected for most metaphases. As in *N. nudicaulis*, it was not possible to resolve progenitor chromosome sets.

DISCUSSION

Recent advances have revealed that WGD is ubiquitous among angiosperms (Jiao *et al.*, 2011), yet most flowering plants have a relatively small genome size (Bennett and Leitch, 2010) and appear to be functionally diploid (Soltis *et al.*, 2005). This is thought to arise through loss of sequences after polyploidy, perhaps reflecting selection against large genome sizes (Leitch and Leitch, 2012). However, in *Nicotiana* section *Repanda*, polyploids show either genome upsizing or downsizing (Figure 1). This variation in genome size is likely to have arisen post-allopolyploidy, as the section is thought to have formed from a single origin approximately 5 million years ago (Clarkson *et al.*, 2010).

Currently, we know relatively little about processes and patterns that lead to diploidization and genome downsizing in polyploid plants. To address this deficiency, we analysed differences in repetitive DNA content between the allopolyploids *N. repanda* and *N. nudicaulis*, and compared them with the extant diploids (*N. sylvestris* and *N. obtusifolia*) that are most closely related to their actual progenitors. Combining next-generation sequencing reads from these four *Nicotiana* species and subjecting the whole dataset to graph-based clustering allowed characterization, quantification and comparisons of repetitive DNA families between species (Figures 2–4, Table 1 and Data S1). Similar approaches have been used to characterize repeats in diploid (Macas *et al.*, 2007, 2011; Novak *et al.*, 2010) and allotetraploid species (Renny-Byfield *et al.*, 2011).

Using extant diploids that are most closely related to the actual progenitors of section *Repandae* is the only option available for assessing changing patterns in repeat composition in association with polyploidy. Alterations in repeat abundance reflect changes that have occurred along branches leading to both the allopolyploids and diploids. However, our reconstruction of ancestral genome size using Markov chain Monte Carlo methods, coupled with our approaches assessing patterns of repeat divergence, both point towards genome downsizing in *N. nudicaulis* and upsizing in *N. repanda*. Furthermore, if the allopolyploids arose from a single event, as is likely (Clarkson *et al.*, 2010), differences in repeat abundance between species of section *Repandae* must have occurred subsequent to allopolyploidy.

Genome diploidization through loss of low-copy number repeats

Most angiosperms fall within a narrow range of genome size (50% of species have a genome size less than 2500 Mbp/1C; Leitch and Leitch, 2012), despite multiple WGDs in their ancestry (Jiao *et al.*, 2011). Furthermore, global analyses of genome sizes in angiosperms indicate that polyploid genomes tend to decrease in size subsequent to formation (Leitch and Bennett, 2004). For example, genome downsizing has been proposed in allotetraploid *N. tabacum* (Leitch and Bennett, 2004), and mechanisms leading to genome downsizing may have acted rapidly, as some repeats are lost even in synthetic lines that are just a few generations old (Skalicka *et al.*, 2005; Renny-Byfield *et al.*, 2012).

Despite genome upsizing in *N. repanda* and downsizing in *N. nudicaulis*, our analysis shows that repeats at low abundance are predominantly under-represented in both species (Figure 2a,b). Repeat reduction in the allopolyploids is also evident in the heatmap analysis, where the majority of repeat clusters (>85%) are under-represented (Figure 4, brown).

Mechanisms resulting in genome size change are poorly understood, although various recombination-based processes have been proposed (Kejnovsky *et al.*, 2009; Grover and Wendel, 2010). For example, there is a approximately 80 Mbp difference in genome size between *A. thaliana* and *A. lyrata*, which is thought to be the result of differential rates of deletion. These deletions were often small, but numerous and common in non-coding and repetitive regions, including within transposable element (Hu *et al.*, 2011). In addition, unequal intra-strand homologous recombination and illegitimate recombination have been identified as mechanisms that remove transposable element insertions and thus contribute to genome size reduction (Devos *et al.*, 2002; Kellogg and Bennetzen, 2004). In rice (*Oryza sativa*), removal of transposable elements has resulted in the loss of approximately 190 Mbp DNA, equivalent to a 38% change in genome size over five million years (Ma *et al.*, 2004). It is possible that similar mechanisms affect the low-copy-number fraction of the genome in section *Repandae*.

A loss of low-copy sequences is potentially an integral part of the diploidization process, and is associated with loss of genes and upstream regulator regions. The loss of DNA may arise because of reduced selective constraints arising from genome duplication (Freeling *et al.*, 2012). Such diploidization processes may be ubiquitous in early polyploid divergence.

Expansion of high-copy-number repeats in *N. repanda*

Despite the general trend that most clusters are under-represented in both allotetraploids (Figure 3), there is evidence for substantial expansion of a small number of repeats (Figure 2a,b). Indeed, over-representation of these repeats is equivalent to a 26.0% increase in genome size in *N. repanda* (Figure 2a–c), close to the 24% genome size change predicted here (Figure 1). Furthermore, these repeat families are predominantly Ty3/*Gypsy* retroelements (particularly chromovirus-like retroelements) and low-complexity sequences (Figure 3), which have been inherited from the diploid progenitors in high copy number (Figure 2a–c). In contrast, there is evidence for loss of high-copy-number repeats in the genome of *N. nudicaulis* (Figure 2a,b), some of which are Ty3/*Gypsy* retreoelements (Figure 3), contributing to a reduction of approximately 14% in genome size, similar to the 19% estimated using Markov chain Monte Carlo approaches (Figure 1). These observations suggest that differential deletion and/or accumulation of high-copy-number repeats in these two allotetraploids is largely responsible for their varying genome size.

Transposable elements are often major contributors to angiosperm genomes (Kumar and Bennetzen, 1999), and Ty3/*Gypsy*-like retroelements are particularly prevalent (Macas and Neumann, 2007; Macas *et al.*, 2011), being present in all four species examined here (Table 1). The excess of Ty3/*Gypsy*-like retroelements in *N. repanda* contributes most to the increased genome size (Figure 3). Similarly, differential accumulation of Ty3/*Gyspy*-like *Gorge3* transposable elements produced a threefold variation in genome size in diploid *Gossypium* (Hawkins *et al.*, 2006). We know there is potential for genome size change to occur rapidly in allopolyploids as activation and integration of transposable elements may occur after only a few generations (Petit *et al.*, 2010). Together with the data presented here, these observations suggest that transposable element dynamics play an important role in governing genome size after allopolyploidy.

Genome turnover

The dual process of DNA loss in the low-copy-number fraction (diploidization) and large-scale changes in the high-copy-number fraction leads to 'genome turnover' (Lim *et al.*, 2007). Analyses of retroelement insertions (Ramakrishna *et al.*, 2002; Ma *et al.*, 2004; Bennetzen, 2005) and nuclear integrants from the plastid genome (Matsuo *et al.*, 2005) have suggested that genome turnover occurs at a rapid rate, with retroelement half-lives of only one to a few million years. Turnover of DNA results in the loss of GISH signal previously reported in *N. nesophila* section *Repandae* (Clarkson *et al.*, 2005; Lim *et al.*, 2007) and shown here, particularly in *N. repanda* (Figure 5). The loss of sub-genome discrimination by GISH indicates that genome turnover has acted to homogenize the genomes. The genomes of both allopolyploids are no longer compartmentalized, as in a nascent allopolyploid, and in this respect have returned to a more diploid-like state. Perhaps genome homogenization results from the loss of repeats with low abundance and the transfer of repeats between sub-genomes.

CONCLUSION

Differences in genome size between *N. repanda* and *N. nudicaulis* appear to be a consequence of differential deletion and/or accumulation of the high-copy-number fraction of the genome. It follows that evolution and amplification of *de novo* repetitive DNA sequences have had only minimal effects on genome size variation and genome divergence in these two species. On the other hand, diploidization of the genomes in both allopolyploids is associated with loss of low-copy-number nuclear sequences and blending of the two progenitor sub-genomes. As all angiosperms are paleopolyploids (Jiao *et al.*, 2011), it is likely that the processes we describe here, i.e. genome size change and diploidization, have played key roles in their evolution.

EXPERIMENTAL PROCEDURES

Phylogenetic analysis and ancestral genome size reconstruction

As all previous analyses have yielded congruent results regarding the evolutionary relationships between members of section *Repandae* (Chase *et al.*, 2003; Clarkson *et al.*, 2004, 2005, 2010; Kelly *et al.*, 2012), datasets were constructed from combined sequence data for each of the two parental sub-genomes. For the *N. sylvestris*-like sub-genome dataset, sequences from four plastid regions (Clarkson *et al.*, 2004), the nuclear ribosomal internal transcribed spacer (Chase *et al.*, 2003), and regions of the low-copy nuclear genes *NUCLEAR ENCODED PLASTID EXPRESSED GLUTAMINE SYNTHASE* (*npGS*; Clarkson et *al.*, 2010), *ALCOHOL DEHYDROGENASE* (*ADH*) and *LEAFY/FLORICAULA* (*LFY/FLO*; Kelly *et al.*, 2012) were used. For the *N. obtusifolia*-like sub-genome dataset, sequences from *GS*, *ADH*, *LFY/FLO* and the non-transcribed spacer of 5S nuclear ribosomal DNA (Clarkson *et al.*, 2005) were used. All regions were aligned separately using PRANK+F (Loytynoja and Goldman, 2008), and then combined before further optimization by eye using Mesquite version 2.74 (Maddison and Maddison, 2008). Phylogenetic reconstruction by Bayesian inference was performed using MrBayes version 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Separate partitions were used for different codon positions, introns, non-coding spacers and RNA coding regions, applying the best-fit model of evolution for each partition as selected using the Akaike information criterion in MrModelTest version 2.3 (Nylander, 2004). Methods S1 provides further detailed information.

The ancestral genome size at each node of the phylogenetic tree was reconstructed using BayesTraits version 1.1beta (http://www.evolution.reading.ac.uk/BayesTraits.html) by analysing genome sizes for the four extant species of section *Repandae* (genome sizes taken from Leitch *et al.*, 2008) as continuously varying (Pagel, 1997, 1999), together with trees from the MrBayes analysis. Values for ancestral genome size for section *Repandae* as a whole, the most recent common ancestor of *N. repanda*, *N. nesophila* and *N. stocktonii*, and the most recent common ancestor of *N. nesophila* and *N. stocktonii* were calculated by averaging all estimates for these nodes from the 90 000 post-burn-in iterations. Methods S1 provides further detailed information.

Plant material

We used *Nicotiana obtusifolia* (accession number 8947501/176) and *N. nudicaulis* (accession number 964750051) (both from the Botanical and Experimental Garden, Radboud, University of Nijmegen, The Netherlands), *N. sylvestris* (accession number ITB626) (from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France), and *N. repanda* (accession number TW18) (from the United States Department of Agriculture, North Carolina State University, NC).

DNA sequencing

DNA extractions were performed as described by Fojtova *et al.* (2003). We sequenced a random sample of DNA from the genomes of *Nicotiana sylvestris*, *N. obtusifolia*, *N. repanda* and *N. nudicaulis* using an Illumina Genome Analyzer xII (http://www.illumina.com/systems/genome_analyzer_iix.ilmn), at the Genome Centre, Queen Mary University of London, generating 108 bp reads. Raw sequence reads were deposited at the Sequence Read Archive at the National Center for Biotechnology Information under the study accession numbers SRA045794 and SRA051392. Resulting sequence reads were then screened for quality and removed if they contained more than five unidentified nucleotides or were shorter than 95 bp in length. All sequences that passed quality checks were trimmed to 95 bp and screened against plastid genomes, and reads with significant similarity were removed from further analysis.

Clustering and repeat identification

A random sample of 5% of each genome was combined into a single dataset and subjected to a graph-based clustering procedure as described by Novak *et al.* (2010). Details of the data used in this analysis are provided in Table S1. This approach identifies repetitive DNA families using a 'community' approach by grouping high-throughput sequencing reads into clusters based on shared sequence similarity. Each sequence read was compared with all other reads in a pairwise analysis using MGBLAST (Altschul *et al.*, 1990), whereby a hit required at least 90% sequence identity along 55% of the sequence read. Graph-based clustering was performed using the R programming language to create an algorithm that detects sets of reads that are more densely connected among each other than to other reads. These groups are termed 'clusters', and correspond to families of repetitive DNA that were characterized further.

Sequences within the largest clusters were analysed to produce a 3D network for each cluster, enabling visualization of similarity between individual reads. Sequence reads (nodes) were connected by edges, where edge weight is proportional to sequence similarity. Nodes were then positioned using a Fruchterman–Reingold algorithm by which reads with extensive similarity are placed close together and those that share little or none are placed further away. Subsequently, the 3D networks were viewed and inspected using the SeqGrapheR program (Novak *et al.*, 2010).

Sequence similarity between reads may be interpreted in two ways: (i) the Illumina sequencer has read the same genomic region more than once, or (ii) the sequences cover regions within repetitive DNA. As each genome was skimmed to a depth of 5%, it is most probable that reads with sequence overlap arise from similar repetitive DNA rather than coverage of the same genomic region. As all sequences are the same length, it follows that the number of reads in each cluster is a measure of abundance within the original dataset. Therefore, a count of the number of sequence reads from each species within a cluster gives a quantitative measure of abundance in the genome of each species. For each cluster we counted the number of reads and calculated the genome proportion (a percentage of the genome) for all four species. Thus genome representation (total contribution of a cluster to the dataset, in bp) and genome proportion are reflective of the total contribution of a given cluster/repeat family to genome size, and are not measures of copy number per se. For example, 10 000 copies of a LTR retroelement 5 kb long have a smaller genome proportion than the same number of elements that are 12 000 kb in size.

After graph-based clustering, reads were assembled using the TGICL (Pertea *et al.*, 2003) version of CAP3 (Huang and Madan, 1999) on a cluster by cluster basis, requiring 80% sequence similarity along a 40 bp length. Clusters consisting of at least ten reads were assessed for sequence similarity to a database of known repetitive sequences (RepBase 16.03, Jurka *et al.*, 2005) using RepeatMasker (Smit *et al.*, 2010) (with the –s option that invokes slower and more sensitive searches). To avoid spurious labelling of clusters, only those descriptions encompassing at least 10% of the reads or totalling 100 hits were considered. The number of reads in clusters with the same description was summed in order to calculate genome proportion for all repeat types.

Comparing deviation of repeat abundance in the allotetraploids

At the outset of allopolyploidy, the quantitative abundance of a given repeat is the sum of the diploid progenitors. Using this logic, we assessed each cluster for deviation from expectation in the allotetraploids. We also calculated the cumulative deviation from additvity across the range of expected repeat abundance, including only those clusters where we expected ten or more reads based on was observed in the diploids.

All analysis was performed using custom R, Perl and bash scripts, which are available at http://webspace.qmul.ac.uk/sbyfield/Simon_Renny-Byfield/Research_Projects.html and http://evolve.sbcs.qmul.ac.uk/leitch/ngs/.

Genomic in situ hybridization (GISH)

Genomic DNA was extracted from fresh leaf material of *N. obtusifolia* and *N. sylvestris* using a Qiagen (http://www.qiagen.com/) DNeasy kit according to the manufacturer's instructions. Following extraction, 1 μg genomic DNA was labelled with either biotin-14–dUTP or digoxigenin-11–dUTP using the Roche (https://www.roche-applied-science.com/sis/lad/index.jsp?id=LA050002) nick translation kit, according to the manufacturer's instructions.

Cells at metaphase were accumulated in freshly harvested root-tip meristems by pre-treatment in saturated Gammexane® (hexachlorocyclohexane, Sigma, http://www.sigmaaldrich.com/united-kingdom.html) in water for 4 h. Subsequently root tips were fixed for 24 h in 3:1 absolute ethanol/glacial acetic acid, and stored in 100% ethanol at −20°C. Root-tip material was spread onto acid-cleaned glass slides following enzyme digestion as described by Lim *et al.* (1998), and checked for quality using phase-contrast microscopy.

Genomic in situ hybridization was performed as described by Lim *et al.* (2006). Briefly, probe DNA (approximately 100 ng of each genomic probe per slide) was added to the probe hybridization mix [50% v/v formamide, 10% w/v dextran sulfate, 0.1% w/v SDS in 2 × SSC (0.3 M NaCl, 0.03 M sodium citrate, pH 7.0)]. Approximately 50 μl of the probe mixture was added to each slide, and the material was denatured using a Dyad slide heating block (MJ Research, http://www.gmi-inc.com/mj-research-dyad-dual-96-well-thermal-cycler.html) at 70°C for 2 min. After hybridization at 37°C overnight, slides were washed in 20% v/v formamide in 0.1 × SSC at 42°C for 10 min, giving an estimated hybridization stringency of 85%. Sites of probe hybridization were detected using 20 μg ml^{-1} fluorescein isothiocyanate-conjugated anti-digoxigenin IgG (Roche) and 5 μg ml^{-1} Cy3-conjugated streptavidin (Amersham Biosciences, http://www.gelifesciences.com/webapp/wcs/stores/servlet/Home/en/GELifeSciences-UK/). Chromosomes were counterstained using Vectashield with 4′,6–diamidino-2–phenylindole (DAPI, Vector Laboratories, http://www.vectorlabs.com/catalog.aspx?catID=279). Material was photographed using a Hamamatsu (http://www.hamamatsu.com/us/en/index.html) Orca ER camera and a Leica (http://www.leica-microsystems.com/) DMRA2 epifluorescence microscope. Subsequently images were processed uniformly using Improvision Openlab® (http://www.perkinelmer.co.uk/pages/020/cellularimaging/products/openlab.xhtml) and Adobe Photoshop CS2 software (http://www.adobe.com/uk/).

ACKNOWLEDGEMENTS

We thank the Natural Environment Research Council for PhD studentship funding, and Dr Richard Buggs and Dr Ilia Leitch (Royal Botanic Gardens, Kew, Richmond Surrey, UK) for constructive comments on the manuscript. The work was partially supported by the Czech Science Foundation (P501/13/10057S and P501/12/G090). We thank Robert Horton and Christopher Walker for their help with the high-performance computing cluster at Queen Mary University of London.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Data S1. Two-dimensional projections of 3D networks for the largest clusters produced in the clustering analysis.

Figure S1. Scatter plot showing observed and expected values for all clusters with expected values greater than ten in allopolyploids, *N. repanda* and *N. nudicaulis*.

Figure S2. Majority-rule consensus trees from MrBayes analyses, showing all compatible groupings and mean branch lengths.

Methods S1. Details of the analysis performed to reconstruct ancestral genome sizes using BayesTraits version 1.1beta.

Table S1. Details of the data used in the clustering algorithm

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Barker, M.S., Vogel, H. and Schranz, M.E. (2009) Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1**, 391–399.

Bennett, M.D. and Leitch, I.J. (2010) Angiosperm DNA C–values database [WWW document]. URL http://data.kew.org/cvalues/ [accessed on 12 March 2103].

Bennetzen, J.L. (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* **15**, 621–627.

Bowers, J.E., Chapman, B.A., Rong, J.K. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.

Chase, M.W., Knapp, S., Cox, A.V., Clarkson, J.J., Butsko, Y., Joseph, J., Savolainen, V. and Parokonny, A.S. (2003) Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Ann. Bot.* **92**, 107–127.

Clarkson, J.J., Knapp, S., Garcia, V.F., Olmstead, R.G., Leitch, A.R. and Chase, M.W. (2004) Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Mol. Phylogenet. Evol.* **33**, 75–90.

Clarkson, J.J., Lim, K.Y., Kovarik, A., Chase, M.W., Knapp, S. and Leitch, A.R. (2005) Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol.* **168**, 241–252.

Clarkson, J.J., Kelly, L.J., Leitch, A.R., Knapp, S. and Chase, M.W. (2010) Nuclear glutamine synthetase evolution in *Nicotiana*: phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Mol. Phylogenet. Evol.* **55**, 99–112.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079.

Fojtova, M., Van Houdt, H., Depicker, A. and Kovarik, A. (2003) Epigenetic switch from posttranscriptional to transcriptional silencing is correlated with promoter hypermethylation. *Plant Physiol.* **133**, 1240–1250.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D. and Schnable, J.C. (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr. Opin. Plant Biol.* **15**, 131–139.

Grover, C.E. and Wendel, J.F. (2010) Recent insights into mechanisms of genome size change in plants. *J. Bot.* **2010**, 382732.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* **16**, 1252–1261.

Hawkins, J.S., Proulx, S.R., Rapp, R.A. and Wendel, J.F. (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl Acad. Sci. USA*, **106**, 17811–17816.

Hribova, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J. and Dolezel, J. (2010) Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing. *BMC Plant Biol.* **10**, 204.

Hu, T.T., Pattyn, P., Bakker, E.G. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.

Huang, X.Q. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Jaillon, O., Aury, J.–M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jiao, Y., Wickett, N.J., Ayyampalayam, S. *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kejnovsky, E., Leitch, I.J. and Leitch, A.R. (2009) Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol. Evol.* **24**, 572–582.

Kellogg, E.A. and Bennetzen, J.L. (2004) The evolution of nuclear genome structure in seed plants. *Am. J. Bot.* **91**, 1709–1725.

Kelly, L.J., Leitch, A.R., Clarkson, J.J., Knapp, S. and Chase, M.W. (2012) Reconstructitng the complex evolutionary origin of wild allopolyploid tobaccos (Nicotiana section Suaveolentes). *Evolution*, **67**, 80–94.

Koukalova, B., Moraes, A.P., Renny-Byfield, S., Matyasek, R., Leitch, A.R. and Kovarik, A. (2010) Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over c. 5 million years. *New Phytol.* **186**, 148–160.

Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532.

Leitch, I.J. and Bennett, M.D. (2004) Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* **82**, 651–663.

Leitch, A.R. and Leitch, I.J. (2012) Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629–646.

Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J. and Leitch, A.R. (2008) The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot.* **101**, 805–814.

Lim, K.Y., Leitch, I.J. and Leitch, A.R. (1998) Genomic characterisation and the detection of raspberry chromatin in polyploid *Rubus*. *Theor. Appl. Genet.* **97**, 1027–1033.

Lim, K.Y., Kovarik, A., Matyasek, R., Chase, M.W., Knapp, S., McCarthy, E., Clarkson, J.J. and Leitch, A.R. (2006) Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section *Alatae*. *Plant J.* **48**, 907–919.

Lim, K.Y., Kovarik, A., Matyasek, R., Chase, M.W., Clarkson, J.J., Grandbastien, M.A. and Leitch, A.R. (2007) Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytol.* **175**, 756–763.

Loytynoja, A. and Goldman, N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.

Ma, J.X., Devos, K.M. and Bennetzen, J.L. (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869.

Macas, J. and Neumann, P. (2007) Ogre elements – a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, **390**, 108–116.

Macas, J., Neumann, P. and Navratilova, A. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427.

Macas, J., Kejnovsky, E., Neumann, P., Novak, P., Koblizkova, A. and Voyskot, B. (2011) Next generation sequencing-based analysis of repetitive DNA in the model dioecious plant *Silene latifolia*. *PLoS ONE*, **6**, e27335.

Maddison, W.P. and Maddison, D.R. (2008) Mesquite: a modular system for evolutionary analysis, version 2.75 [WWW document]. URL http://mesquiteproject.org/mesquite/mesquite.html [accessed on 12 March 2013].

Matsuo, M., Ito, Y., Yamauchi, R. and Obokata, J. (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell*, **17**, 665–675.

Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Nylander, J.A.A. (2004) MrModeltest version 2 [WWW document]. URL http://www.abc.se/~nylander/mrmodeltest2/mrmodeltest2.html [accessed on 12 March 2013].

Ozkan, H., Levy, A.A. and Feldman, M. (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops–Triticum*) group. *Plant Cell*, **13**, 1735–1747.

Pagel, M. (1997) Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26**, 331–348.

Pagel, M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Parisod, C., Mhiri, C., Lim, K.Y., Clarkson, J., Chase, M.W., Leitch, A.R. and Grandbastien, M.A. (2012) Differential dynamics of transposable elements during long-term diploidization of *Nicotiana* section *Repandae* (Solanaceae) allopolyploid genomes. *PLoS ONE* **7**, e50352.

Pertea, G., Huang, X.Q., Liang, F. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Petit, M., Lim, K.Y., Julio, E., Poncet, C., de Borne, F.D., Kovarik, A., Leitch, A.R., Grandbastien, M.A. and Mhiri, C. (2007) Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). *Mol. Genet. Genomics*, **278**, 1–15.

Petit, M., Guidat, C., Daniel, J. *et al.* (2010) Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytol.* **186**, 135–147.

Ramakrishna, W., Dubcovsky, J., Park, Y.J., Busso, C., Emberton, J., San-Miguel, P. and Bennetzen, J.L. (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics*, **162**, 1389–1400.

Renny-Byfield, S., Chester, M., Kovařík, A. *et al.* (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.* **28**, 2843–2854.

Renny-Byfield, S., Kovarik, A., Chester, M., Nichols, R.A., Macas, J., Novak, P. and Leitch, A.R. (2012) Independent, rapid and targeted loss of a highly repetitive DNA sequence derived from the paternal genome donor in natural and synthetic *Nicotiana tabacum*. *PLoS ONE*, **7**, e36963.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Salina, E.A., Numerova, O.M., Ozkan, H. and Feldman, M. (2004) Alterations in subtelomeric tandem repeats during early stages of allopolyploidy in wheat. *Genome*, **47**, 860–867.

Skalicka, K., Lim, K.Y., Matyasek, R., Matzke, M., Leitch, A.R. and Kovarik, A. (2005) Preferential elimination of repeated DNA sequences from the paternal, *Nicotiana tomentosiformis* genome donor of a synthetic, allotetraploid tobacco. *New Phytol.* **166**, 291–303.

Smit, A.F.A., Hubley, R. and Green, P. (2010) *RepeatMasker Open-3.0*. URL http://www.repeatmasker.org [accessed on 01 September 2012].

Soltis, D.E., Soltis, P.E., Endress, P.K. and Chase, M.W. (2005) *Phylogeny and Evolution of Angiosperms*. Sunderland, MA: Sinauer Associates.

Swaminathan, K., Varala, K. and Hudson, M.E. (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, **8**, 132.

Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, **290**, 2114–2117.

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. and Stein, N. (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* **59**, 712–722.