Next Generation Sequencing Reveals Genome Downsizing in Allotetraploid *Nicotiana tabacum*, Predominantly through the Elimination of Paternally Derived Repetitive DNAs

Simon Renny-Byfield,[1] Michael Chester,[1,2] Ales Kovařík,[3] Steven C. Le Comber,[1] Marie-Angèle Grandbastien,[4] Marc Deloger,[4] Richard A. Nichols,[1] Jiri Macas,[5] Petr Novák,[5] Mark W. Chase,[6] and Andrew R. Leitch*[,1]

[1]School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

[2]Laboratory of Molecular Systematics and Evolutionary Genetics, Florida Museum of Natural History, University of Florida

[3]Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i, Brno, Czech Republic

[4]Institute Jean-Pierre Bourgin, Institut National de la Recherche Agronomique-Versailles, France

[5]Biology Centre, Institute of Plant Molecular Biology, Academy of Sciences of the Czech Republic, České Budějovice, Czech Republic

[6]Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

Next generation sequence data for all species involved in the study were submitted to the Sequence Reads Archive (SRA) under the study accession number SRA023759.

***Corresponding author:** E-mail: a.r.leitch@qmul.ac.uk.

Associate editor: Naoki Takebayashi

Abstract

We used next generation sequencing to characterize and compare the genomes of the recently derived allotetraploid, *Nicotiana tabacum* (<200,000 years old), with its diploid progenitors, *Nicotiana sylvestris* (maternal, S-genome donor), and *Nicotiana tomentosiformis* (paternal, T-genome donor). Analysis of 14,634 repetitive DNA sequences in the genomes of the progenitor species and *N. tabacum* reveal all major types of retroelements found in angiosperms (genome proportions range between 17–22.5% and 2.3–3.5% for Ty3-*gypsy* elements and Ty1-*copia* elements, respectively). The diploid *N. sylvestris* genome exhibits evidence of recent bursts of sequence amplification and/or homogenization, whereas the genome of *N. tomentosiformis* lacks this signature and has considerably fewer homogenous repeats. In the derived allotetraploid *N. tabacum*, there is evidence of genome downsizing and sequences loss across most repeat types. This is particularly evident amongst the Ty3-*gypsy* retroelements in which all families identified are underrepresented in *N. tabacum*, as is 35S ribosomal DNA. Analysis of all repetitive DNA sequences indicates the T-genome of *N. tabacum* has experienced greater sequence loss than the S-genome, revealing preferential loss of paternally derived repetitive DNAs at a genome-wide level. Thus, the three genomes of *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum* have experienced different evolutionary trajectories, with genomes that are dynamic, stable, and downsized, respectively.

Key words: next generation sequencing, allopolyploidy, genome downsizing, transposable elements, retroelements, paternal genome, *Nicotiana tabacum*, *Nicotiana sylvestris*, *Nicotiana tomentosiformis*.

Introduction

Angiosperm evolution has been heavily impacted by polyploidy, which has occurred in the ancestry of most, if not all, species (Soltis et al. 2009). Polyploidy itself may induce revolutionary changes in genome composition in early generations (Leitch and Leitch 2008), a phenomenon explored here. Interspecific hybridization combined with whole-genome multiplication (allopolyploidy) provides a natural experiment in genome perturbation where the fate of DNA sequences can be examined by studying the descendants of the two progenitor species and their allopolyploid offspring. McClintock (1984) first proposed that allopolyploidy can induce "genomic shock" and we now know that changes can occur at the DNA sequence, epigenetic, karyotypic, and transcription levels (Wendel 2000). Moreover, polyploid-associated genetic change has been observed to occur rapidly in some species, occurring after only a few generations, leading many to envisage a "genome revolution" where perturbation of the progenitor genomes is induced by their unification (Wendel 2000; Comai et al. 2003; Liu and Wendel 2003; Feldman and Levy 2009).

Repetitive DNA sequences, which comprise a large proportion of the genomes of many plant species, may be subject to change in sequence, copy number, and/or epigenetic profile following allopolyploidy (Matyasek et al. 2002, 2003; Adams and Wendel 2005; Leitch et al. 2008). However, there are only a few examples of allopolyploid-associated or interspecific hybridization–associated changes for the transposable elements (TEs; Parisod et al. 2010). Such evidence includes 1) the activation and movement of retroelements in natural (Petit et al.

2007) and synthetic *Nicotiana tabacum* (Petit et al. 2010) in addition to the loss of some retroelements which may occur rapidly (within a few generations). 2) The activation of retroelements and miniature inverted-repeat transposable elements in rice following alien DNA introgression from related wild species (Liu and Wendel 2000), in which activation was transient, involving amplification of a few copies (10–20 copies) and methylation of the new insertions, which were stably inherited in subsequent generations. 3) Allopolyploid induced activation of Wis2 retroelement transcription in synthetic crosses of *Aegilops sharonensis* × *Triticum monococcum*, although this was not associated with any observed increase in copy number or element mobility (Kashkush et al. 2003).

There are examples where repeat sequence activation following allopolyploidy is not apparent, notably in *Gossypium* synthetic allopolyploids (Liu et al. 2001; Hu et al. 2010) and in recently formed (within last 150 years) natural *Spartina anglica* (Ainouche et al. 2009). Similarly, sequence-specific amplified polymorphism analysis of *Arabidopsis thaliana* × *Arabidopsis lyrata* revealed the CAC family of transposons was not activated in neotetraploids (Beaulieu et al. 2009). However, there were substantial epigenetic changes influencing establishment of nucleolar dominance and degree of cytosine methylation at 25% of loci examined as well as a large chromosomal deletion.

Genome size estimates have indicated that many allopolyploids have undergone genome downsizing (Dolezel et al. 1998; Leitch and Bennett 2004; Beaulieu et al. 2009). Certainly, the balance between retrotransposition and DNA deletion will influence genome size and turnover of DNA sequences (Leitch and Leitch 2008). Indeed, analysis of rice bacterial artificial chromosome clones revealed that retroelement insertions may only have a half-life of a few million years, an indication of the speed with which these retroelement replacement mechanisms can operate (Ma et al. 2004). Such turnover of sequences may explain why genomic in situ hybridization fails, even in some relatively young *Nicotiana* allopolyploids where loss of homology with progenitor species can be detected after only approximately 5 million years of divergence (Clarkson et al. 2005; Lim et al. 2007).

It is apparent that the dynamism of plant genomes is not restricted to changes in DNA sequence. In *Spartina* and *Dactylorhiza* allopolyploid hybrids, epigenetic alterations have been shown to occur rapidly; such changes are often associated with TEs and can be specific to the maternally derived portion of the genome (Parisod et al. 2009; Paun et al. 2010).

The genus *Nicotiana* (Solanaceae) provides an excellent model group for studies on the consequences of polyploidy, because the genus consists of approximately 70 species, and ~40% of which are documented to be allotetraploids derived from six independent polyploidy events (Clarkson et al. 2005, 2010; Leitch et al. 2008). The allopolyploid species studied here, *N. tabacum* (tobacco), is particularly worth studying because it is relatively young species (less than 200,000 years old; Leitch et al. 2008) and is derived from known ancestors that are related to *Nicotiana sylvestris* (the maternal genome donor, the S-genome component of *N. tabacum*) and *Nicotiana tomentosiformis* (the paternal genome donor, the T-genome component of *N. tabacum*).

Previous molecular and cytogenetics studies have suggested that for noncoding tandemly repeated DNA, *N. tabacum* is typically additive for its two diploid parents (Murad et al. 2002; Koukalova et al. 2010), exceptions being for 35S nuclear ribosomal DNA (rDNA), a satellite called NTRS and A1/A2 repeats derived from the intergenic spacer (IGS) of 35S rDNA. The IGS in *N. tabacum* has experienced near complete replacement with a novel unit most closely resembling the *N. tomentosiformis* type (Volkov et al. 1999; Lim, Kovarik, et al. 2000). In addition, A1/A2 repeats that are found within the IGS and scattered across the *N. tomentosiformis* genome have fewer than expected dispersed copies in *N. tabacum* (Lim et al. 2004). Similarly, for Tnt2 retroelements, there is evidence for the gain of new insertion sites as well as element loss (Petit et al. 2007). Other variation includes translocations between the S- and T-genomes, some of which appear ubiquitous, and probably fixed, whereas others are specific to particular *N. tabacum* cultivars (Lim, Matyasek, et al. 2004). In generation S3 of synthetic *N. tabacum*, a similar translocation to the one fixed in *N. tabacum* is observed in some plants, suggesting a fitness advantage for such a change (Skalicka et al. 2005). Furthermore, in some synthetic *N. tabacum* lines, there is already replacement of several thousand rDNA units with a novel unit type (Skalicka et al. 2003) and evidence for the loss of *N. tomentosiformis*–derived Tnt1 insertion sites (Petit et al. 2010). These events suggest a rapidly diverging genome, perhaps responding to the genomic shock of allotetraploidy (McClintock 1984).

The emergence of next generation sequencing technologies (Margulies et al. 2005) has enabled, for the first time, the possibility of studying in detail and at modest cost, the repetitive elements of any genome. Using DNA sequence data produced with 454 pyrosequencing and a genome coverage of ~1%, Macas et al. (2007) have been able to calculate copy number and genome proportions of well-represented repeat sequences in pea (*Pisum sativum*). In addition, Swaminathan et al. (2007) have used a similar approach to classify the repeats present in soybean, whereas others have investigated the genome of barley (Wicker et al. 2006, 2009). More recently, Hribova et al. (2010) have used 454 read-depth analysis to characterize the repeat component of the banana genome. However, these studies did not focus on addressing the question of how repeat sequences respond to allopolyploidy, the principal objective of this paper.

Here, we compare the genomes of *N. tabacum* and the extant lineages most closely related to its two diploid progenitors by using 454 GS FLX Titanium Technology, sequencing in each case at least 0.5% of the genome. Such data combined with clustering based repeat identification and abundance estimates using established approaches (Novak et al. 2010) enabled us to analyze patterns of

evolution subsequent to polyploidy for abundant repetitive sequences. We present here our analysis of the nuclear ecology (sensu Brookfield 2005) and population dynamics of repeat sequences associated with the divergence of allotetraploid *N. tabacum*.

Materials and Methods

Plant Material

Nicotiana sylvestris Speg. & Comes (ac. ITB626) was obtained from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France. *Nicotiana tomentosiformis* Goodsp. (ac. NIC 479/84) was from the Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. *Nicotiana tabacum* cv. SR1, Petit Havana, was obtained from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France. The *N. tomentosiformis* accession was selected because it is the most similar of the accessions to the T-genome of *N. tabacum*, with which it shares several cytological markers (Murad et al. 2002) and amplified fragment length polymorphisms (MA Grandbastien, unpublished data). Of the *N. sylvestris* accessions available, none is particularly more suitable than any other as they are all closely related (Petit et al. 2007).

DNA Extraction and 454 Sequencing

To reduce organellar contamination of reads, genomic DNA was isolated from purified nuclei prepared from fresh leaf tissue as described in Fojtova et al. (2003). Extracted DNA was checked for integrity by gel electrophoresis. Approximately 5 μg of genomic DNA was submitted for sequencing at the NERC Biomolecular Analysis Facility—Liverpool, United Kingdom. DNA was randomly sheared by nebulization and sequenced using a 454 GS FLX Instrument with Titanium reagents (Roche Diagnostics). For each species, we used one-eighth of a 70 × 75 picotiter plate. Sequence reads were submitted to the NCBI sequence read archive (SRA) under the study accession number SRA023759.

Preparation and Analysis of 454 Reads

Using custom Perl scripts sequence reads and associated quality files, the first ten bases were clipped to remove any associated adapter sequences. The stand-alone Blast program (http://www.ncbi.nlm.nih.gov/) was used to screen 454 reads for similarity to the appropriate plastid genome (*N. sylvestris* NCBI#: NC_007500.1, *N. tomentosiformis* NCBI#: NC_007602.1, and *N. tabacum* NCBI#: NC_001879.2). Reads with significant hits (e-value $< e^{-6}$) to plastid DNA were excluded from further analysis, whereas the remaining 454 reads were considered nuclear in origin.

Comparative Genome Analysis Using Blast

The stand-alone Blast program was used to assess sequence similarity at the genome-wide level. Complete pairwise analysis was performed on the *N. tabacum* data set and the proportion of reads with significant hits (*E* value $< e^{-8}$) was recorded for each sequence. All other Blast parameters were set to default throughout the analysis. The same analysis was repeated using *N. tabacum* sequences to probe the *N. sylvestris* and *N. tomentosiformis* data sets, and for each *N. tabacum* read, the number of sequences (from the progenitor data set) with significant sequence similarity hits to the *N. tabacum* reads was recorded. Due to the number of reads in each data set being unequal, the number of hits recorded in all cases was standardized to the *N. tabacum* data set, where hit numbers were scaled up or down depending on the difference in the number of reads between data sets. For example, the *N. tabacum* data set consists of 70,616 reads, whereas the *N. tomentosiformis* data set has 65,858 reads and to standardize these data, the number of hits recorded for *N. tomentosiformis* was multiplied by 1.072 (number of reads in the *N. tabacum* data set/number of reads in the *N. tomentosiformis* data set). We constructed an in silico *N. tabacum* consisting of a random set of 35,000 reads from each of the parental data sets. An equal contribution from the parents was used to reflect the equivalence of genome size in the progenitor species. A control analysis consisting of the in silico *N. tabacum* in place of the 454 *N. tabacum* reads was performed (supplementary fig. 1, Supplementary Material online).

Individual *N. tabacum* 454 reads, for annotation purposes, were subjected to sequence similarity searches to known repeat elements, including those submitted to RepBase and a custom database consisting of known satellite and rDNA repeats from the three *Nicotiana* species. We also annotated the reads with known protein domains by sequence similarity to the pfam database. The resulting data were plotted in the R statistical package (R Development Core Team 2010; fig. 1 and supplementary fig. 1, Supplementary Material online).

Genome-Wide Analysis of Mean Similarity of Related Sequences

We analyzed the genomes of parental and progenitor species by comparisons of the mean sequence similarity of related sequences. Blast analysis with an *e*-value cutoff of e^{-6} was used to identify related sequences. Custom Perl scripts and the R statistical package were used to calculate the mean sequence similarity for each sequence with a Blast hit under the condition that the HSP was above 80 bp in length.

Clustering, Contig Assembly, Graph Visualization, and Species-Specific Reference Assembly

Repeat sequence assembly was performed with a combined data set of 454 reads from all three species using a graph-based clustering approach as described in Novak et al. (2010). Briefly, the reads were subjected to a complete pairwise sequence comparison, and their mutual similarities were represented as a graph in which the vertices corresponded to sequence reads; overlapping reads were connected with edges and their similarity scores were expressed as edge weights. Distances between a given node (a single sequence) and other related nodes are determined, in part, by the bit score (edge weight) of a Blast analysis between sequences and a Fruchterman–Reingold

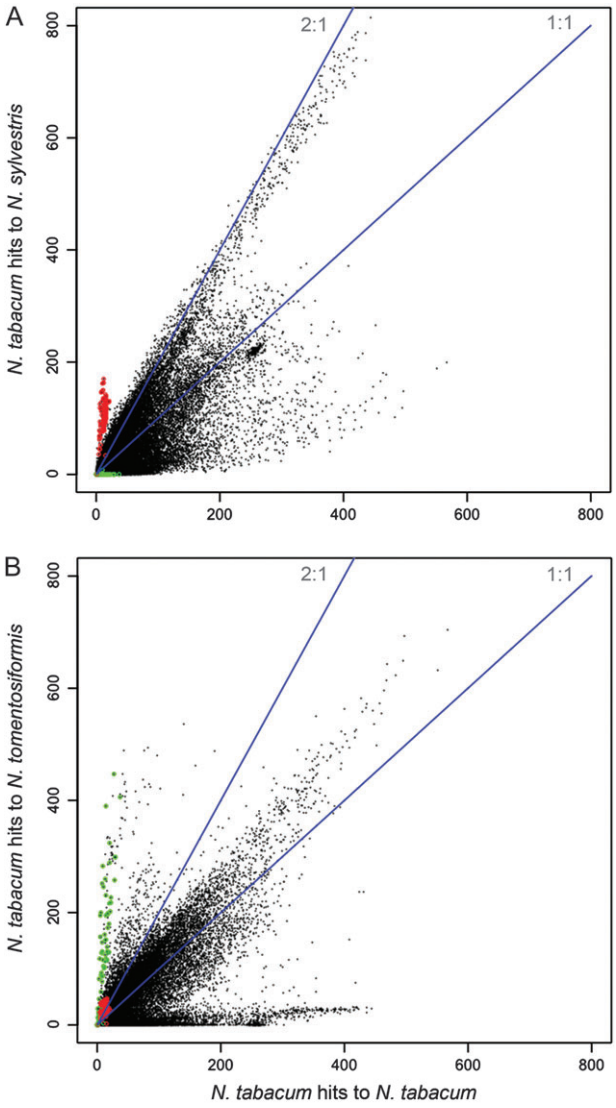

FIG. 1. Genome comparisons using pairwise similarity analysis of individual 454 reads. *Nicotiana tabacum* reads compared with the *N. tabacum* data set (*x* axis) and *Nicotiana sylvestris* (*A*) or *Nicotiana tomentosiformis* (*B*) data sets (*y* axis) using the Blast program with *e* value cutoff of 10^{-8}. The number of similarity hits was normalized to take into account the varying size of each data set. Reads highlighted red and green are rDNA and NTRS sequences; respectively, 1:1 and 2:1 lines are labeled and indicated in blue. In (*A*), reads on the 2:1 line are likely from *N. sylvestris* with reduced frequency in *N. tabacum* caused by the unification (as a result of allopolyploidy) with the *N. tomentosiformis* genome. Reads on the 1:1 line in (*A*) and (*B*) are sequences inherited from both parents where they occur in similar copy numbers.

algorithm is used to position the nodes. This results in more similar sequences being placed closer together, whereas more distantly related reads are placed further apart. The graph structure was analyzed using custom-made programs in order to detect clusters of frequently connected nodes representing groups of similar sequences. These clusters, corresponding to families of genomic repeats, were separated and analyzed with respect to the number of reads they contained (which is proportional to their genomic abundance) and similarity to known repeats. Graphs of selected clusters were also visually examined using the SeqGrapheR program (Novak et al. 2010) in order to assess structure and variability of the repeats.

We then used the CLC Genomics Workbench v. 3 to independently map reads derived from each species to reference sequences derived from the clustering and assembly algorithm described above. Default parameters of at least 80% sequence identity along 50% of the sequence read were used. This approach allowed us to estimate the average read depth along the length of the contig (RD), genome representation (GR, the average RD × the length of the contig), and genome proportion (GP, calculated as (GR/database size in bp) × 100) for all reference sequences in each species. Sequence similarity searches and custom Perl scripts were used to sort resulting clusters and contigs according to sequence type, RD, and GR. Clusters were annotated using sequence similarity (BlastN and BlastX) searches to the entire RepBase (edition 14.10, accessed 9/1/2009), using an *e* value cutoff score of e^{-6}. Additional annotation using the Blast function on the GyDB was required in order to establish the clade to which Ty3-*gypsy*-like elements belonged (Llorens et al. 2008). The total GR and GP of a given repeat was calculated by summing all GR and GP estimates for clusters associated with that repeat type.

All scripts are available on request.

Results

454 High-Throughput DNA Sequencing

454 GS FLX Titanium sequencing of genomic DNA of *N.* sylvestris, *N. tomentosiformis*, and *N. tabacum* returned between 68,000 and 75,000 reads per species, with an average read length of 360–370 bp. This totals 22–29 Mb of DNA sequence per species. Filtering for plastid contaminants and trimming of primer sequences resulted in 19–25 Mb of DNA sequence for each accession. This amounts to ~0.9% coverage of the *N. sylvestris* (1C genome size of 2,650 Mb) genome, ~0.8% coverage of the *N. tomentosiformis* (1C genome size of 2,650 Mb) genome, and ~0.5% coverage for *N. tabacum* (1C genome size of ~5,100 Mb) (Leitch et al. 2008). Sequence reads were submitted to NCBI SRA under the study accession number SRA023759.

Genome-Wide Comparisons via 454 Read Similarity Analysis

To estimate abundance of sequences within and between species, we conducted pairwise sequence similarity searches. The data are shown as 2D plots where the number of sequence similarity hits in *N. tabacum* is plotted against the number of hits in each parent (fig. 1*A* and *B*). The output reflects the abundance of sequences in the *Nicotiana* genomes. We would expect those sequences that were faithfully inherited in *N. tabacum* exclusively from one parent to fall on a 2:1 line. This is because these sequences will be twice as abundant in the parent as in *N. tabacum*, given the normalized data sets and the effective dilution by

the other parental genome. Those sequences falling above the 2:1 line are likely underrepresented in *N. tabacum* and derived from the parent in the analysis. Similarly, sequences falling on a 1:1 line are expected to be in similar abundance in both parents.

Figure 1*A* and *B* show complete pairwise sequence similarity analysis of individual 454 reads from *N. tabacum* against all reads in the *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis* data sets. In figure 1*A*, which shows the analysis of *N. tabacum* against *N. sylvestris*, there is a distinct clustering of reads on or close to a 2:1 line, suggesting these are *N. tabacum* reads that have been inherited solely or predominantly from *N. sylvestris*. Few reads in this category were identifiable using similarity searches to RepBase or pfam domains. In comparison, when the same analysis is conducted using the *N. tomentosiformis* data set, sequences on the 2:1 line are less abundant (fig. 1*B*).

The analysis in figure 1*A* also shows a spike of sequences that reach substantially higher copy number (i.e., higher frequency of sequence similarity hits) in *N. sylvestris* than in *N. tabacum*. We found that these sequences are predominantly rDNA (highlighted red in fig. 1*A*). The corresponding sequences are also highlighted red in figure 1*B*. This spike was absent in a control genome, generated in silico from an equal mixture (35,000 reads) from each parental data set (totaling 70,000 reads) (supplementary fig. 1*A* and *B*, Supplementary Material online).

In figure 1*B*, 3,078 sequence reads have a higher frequency of sequence similarity hits in *N. tomentosiformis* than would be expected from their observed frequency in *N. tabacum* (i.e., sequences that fall above the 2:1 line in fig. 1*B*). This pattern is absent in the in silico *N. tabacum* (supplementary fig 1*B*, Supplementary Material online). For the reads above the 2:1 line (i.e., they are underrepresented in *N. tabacum* relative to expectation), the mean and sum of the residuals (i.e., deviation from the line) was 22.2 and 68,332, respectively. In *N. sylvestris*, there are 5,919 sequences above the 2:1 line, but the mean of residual of these sequences is only 9.6 and the sum totals 56,822. Amongst the reads above the 2:1 line in figure 1*B* (plotting *N. tabacum* against *N. tomentosiformis*), there are NTRS-like repeat sequences (highlighted in green), previously shown to occur in *N. tomentosiformis*, other species of section *Tomentosae* and *N. tabacum* but not in *N. sylvestris* (Matyasek et al. 1997). The remainder of the reads had few significant hits to RepBase, but several were related to retrotransposon gag (retrotrans-gag) and reverse transcriptase (RVT) pfam domains (data not shown).

Comparison of Genome-Wide Sequence Similarity

For all reads in each of the three data sets, we calculated the mean sequence similarity between that read and all related sequences within the same data set. Figure 2 shows histograms of mean sequence similarity in *N. tabacum*, its diploid progenitors and an in silico *N. tabacum*. A major peak with a mean of ~0.86 is seen in all three species, and in the in silico *N. tabacum*. *Nicotiana sylvestris* has a secondary peak where the mean sequence similarity is close to one (fig. 2*A*). These latter sequences are likely either highly constrained by selection or have experienced recent expansion/homogenization. The sequences from *N. sylvestris* with a mean sequence similarity (to other related reads in the *N. sylvestris* data set) above 0.98 and a coefficient of variation less than one (2,248 reads in total) were identified (supplementary table 1, Supplementary Material online) and shown to include 5S and 35S rDNA sequences (totaling 353 of the reads). In addition, there were a number of *gypsy*-like repeat sequences, although they were in considerably lower abundance than rDNA repeats. *Nicotiana tomentosiformis* (fig. 2*B*) lacks such an abundance of sequences with a high mean sequence similarity. The in silico *N. tabacum* (fig. 2*C*) exhibits a secondary peak of high mean sequence similarity as seen in *N. sylvestris* (fig. 2*A*), but the peak is absent in natural *N. tabacum* (fig. 2*D*).

Clustering and Contig Assembly

We combined all 454 high-throughput DNA sequencing reads from the three *Nicotiana* species and subjected this combined data set (>70 Mb of DNA sequence) to a clustering based repeat identification procedure, leading to partitioning of sequencing data into groups of overlapping reads representing individual repeat families as described in Novak et al. (2010). As the average read depth in each cluster reflects the genomic proportions of the corresponding repeat, read-depth analysis was used to estimate the repeat composition in the genomes of the species studied. Details of the repeat identification and assembly output are given in table 1. The normalized (by total number of reads in the data set) contribution of each species to the 30 largest clusters is shown in supplementary fig. 2 (Supplementary Material online). Clusters were then assembled to provide reference sequences that were used as a scaffold for the independent mapping of reads for each of the three *Nicotiana* species (table 1 and fig. 3). This allowed characterization of the average read depth along the length of the contig (RD), genome representation (GR, calculated as RD × contig length), and genome proportion (GP, calculated as (GR/total size of the data set in base pairs) × 100) for each of the three species. GP is therefore the percentage of the data set (and therefore the genome) that can be attributed to a given repeat. This allowed characterization of the most abundant repeats in the three genomes. Others have used similar approaches to measure repeat sequence abundance (Macas et al. 2007; Swaminathan et al. 2007; Hribova et al. 2010).

An example of the output of the clustering and assembly procedure is provided for cluster CL2, which contains reads from all three species (fig. 4) and has sequence similarity to Ogre-like LTR retroelements (Macas and Neumann 2007). Each node within the network corresponds to a single 454 read and similar reads are placed more closely together than more distantly related sequences. We observe in this network that most reads fall along a contiguous line, similar to an assembly into a single contig. However, it is clear that some related reads deviate from this main axis and become

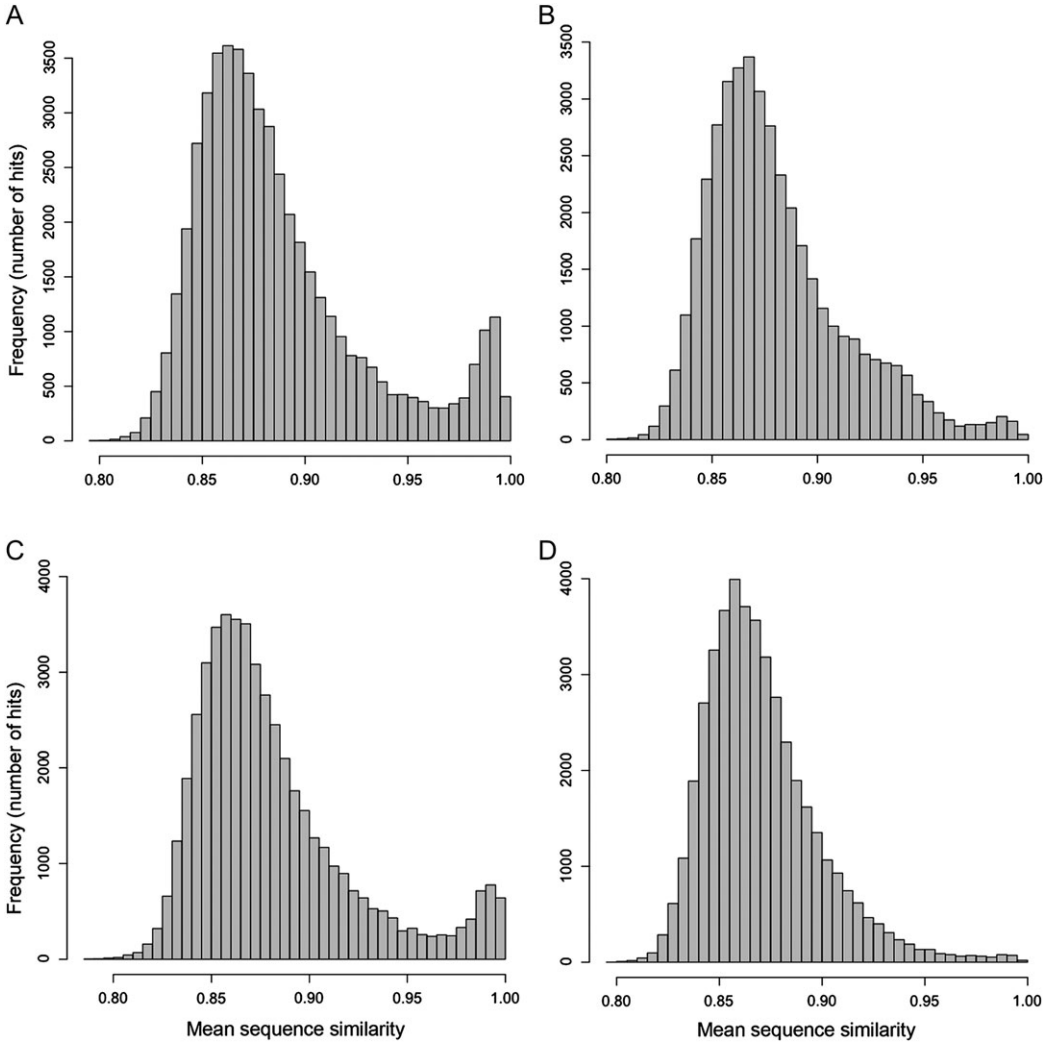

FIG. 2. Histogram showing frequency of mean sequence similarity between each read in a 454 data set and all related sequences in the same data set in (*A*) *Nicotiana sylvestris*, (*B*) *Nicotiana tomentosiformis*, (*C*) an in silico *Nicotiana tabacum*, and (*D*) natural *Nicotiana. tabacum*. In (*A*), many reads have high mean sequence similarity values generating a secondary peak. This peak is much reduced in *N. tomentosiformis* (*B*) and is absent in *Nicotiana tabacum* (*D*).

a linked but separate string of sequences (boxed in fig. 4). These are likely to be alternative variants of this repeat, one of which is found in the genome of *N. sylvestris* and another in *N. tomentosiformis* (red and blue in fig. 4, respectively). Both repeat variants are present in *N. tabacum*.

To check the validity of the contigs developed in silico (as described above), we cloned and sequenced a region of cluster 3 contig 8 and found clones sharing between 92% and 96% identity with the consensus (data not shown).

Characterizing *Nicotiana* Genomes

Table 2 shows GR and GP estimates for the major repeat sequence fraction of the *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis* genomes. The sequence type with largest GPs in all three species was the retroelements, comprising at least 20.52%, 27.17%, and 22.90% of the genomes of *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis*, respectively. In *N. tabacum*, comparison of observed GPs with expected percentages (average of the parents) reveals the GP

Table 1. Details of the Output of the Clustering and Assembly Algorithm for the Combined Data Set (producing the reference sequences) and the Species-Specific Read Mapping Analyses[a].

		Number of Clusters in Assembly	Number of Contigs in Assembly	Minimum/Maximum Contig Length (bp)	Percentage of reads mapped to contigs
Combined assembly		**16,229**	**17,443**	**107/9,632**	**N/A**
	Nicotiana tabacum	**8,496**	**10,464**	**109/5,198**	**44**
Species-specific	***Nicotiana sylvestris***	**9,791**	**11,446**	**107/5,198**	**63**
read mapping	***Nicotiana tomentosiformis***	**6,131**	**7,378**	**108/9,362**	**53**

NOTE.—N/A, not available.

[a] The repeat identification algorithm is described in detail in Novak et al. (2010).

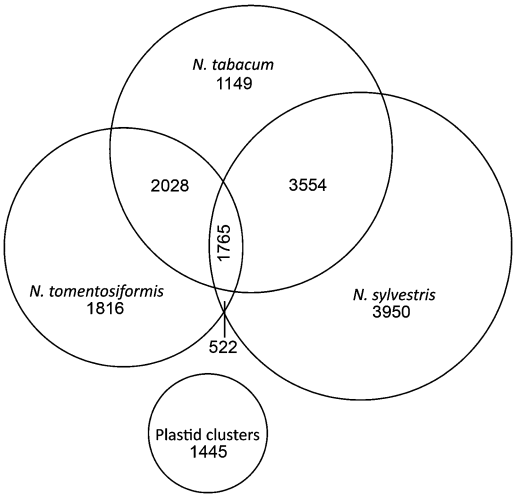

FIG. 3. Venn diagram where the area of each circle (and the intersections) is proportional to number of 16,229 clusters that have significant similarity to sequence reads from the species as indicated. Absolute numbers are given in each section.

of retroelements to be reduced by over 18% from expectation. The majority of retroelements are Ty3-*gypsy*-like (estimates ranging from 17% to 23% in the three species), and in *N. tabacum*, there is a reduction in their GP by 19.8% from expectation. Figure 5*A* shows the contribution of the major groups of the Ty3-*gypsy*-like elements present in all three species. The group with the highest GP in *N. sylvestris* is Tat, which includes the large Ogre and Atlantys elements. This group is also well represented in *N. tabacum* but less so in *N. tomentosiformis*, where the largest group are the Del (Chromovirus) elements. All families of Ty3-*gypsy* have a GP lower in *N. tabacum* than would be expected based on the proportions observed in the diploid progenitors (fig. 5*B*), indicating that sequence loss may have occurred subsequent to allotetraploidy.

Estimates of 35S rDNA abundance in the three *Nicotiana* species have shown that these repeats make up a substantial fraction of the *N. sylvestris* genome (1.70%). The observed abundance of 35S rDNA in *N. tomentosiformis* is lower (0.48%), whereas in *N. tabacum* it is lower still (0.17%), which is more than an 80% reduction in GP compared with that expected. We also observed that one cluster (CL3) is particularly abundant in *N. tomentosiformis* (GP = 1.91%), whereas in *N. tabacum*, the abundance of this repeat was considerably lower (GP = 0.1%). In addition, pararetrovirus-like sequences are more abundant in the *N. tomentosiformis* genome (0.54%) than they are in both *N. tabacum* (0.25%) and *N. sylvestris* (0.22%), revealing a 34% reduction in GP from expectation (table 2).

Comparing Observed with Expected Genome Proportions in *N. tabacum*

We used linear regression to compare GP estimates of 14,634 repeat clusters in *N. tabacum* against those in the two progenitor species (fig. 6). If the *N. tabacum* genome was an equal mixture of the two progenitors, the slope of the regression would have been 0.5 for each (fig. 6). The actual estimate of the slope (fig. 6) was close to, although significantly different from, the expected slope of 0.5 for *N. tabacum* versus *N. sylvestris* (0.472, standard error [SE] 0.004) whereas the GP contribution from *N. tomentosiformis* in *N. tabacum* was considerably lower and significantly different from expectation (0.300, SE 0.003). In fig. 6, notice that 1) the fitted surface (red) through the observed data falls below the expected surface (green), which assumes *N. tabacum* has inherited sequences faithfully from the progenitors; 2) repetitive DNAs inherited from *N. tomentosiformis* appear underrepresented along the length of the data range, and 3) this discrepancy is greatest for the most common repeat elements in *N. tomentosiformis*.

Discussion

454 Titanium Sequencing to Estimate Repeat Abundance

The major repeat composition of the genomes of three *Nicotiana* species has been characterized using 454 GS FLX pyrosequencing, providing between ~0.5% and 1% coverage of

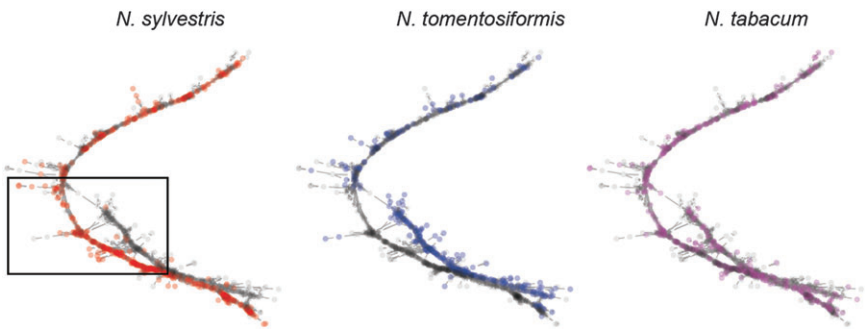

FIG. 4. An example of the output of the clustering based repeat assembly algorithm (Novak et al. (2010) shows a network of sequence reads in Cluster 2 (CL2), where nodes represent sequence reads. Reads with sequence similarity are connected by edges (lines). The graph is reproduced for each species, with the reads highlighted in red (*Nicotiana sylvestris*), blue (*Nicotiana tomentosiformis*), and purple (*Nicotiana tabacum*). There are distinct variants of the repeat in each of the progenitor genomes (this region is boxed in the *N. sylvestris* plot), evident by the splitting of reads into separate strings of sequence, where one string contains reads form *N. sylvestris* and the other from *N. tomentosiformis*. For CL2, *N. tabacum* has both these strings and is additive of the parents.

Table 2. Genome Representation (GR) and Genome Proportion (GP) of Major Repeat Classes within the *Nicotiana tomentosiformis*, *Nicotiana sylvestris*, and *Nicotiana tabacum* Genomes.

			N. sylvestris		*N. tomentosiformis*		*N. tabacum*			
Class	Order	Superfamily	GR	GP (%)	GR	GP (%)	GR	GP (%)	Deviation from Parental Average	Difference as percentage of Expected GP
Retroelement			6870681	27.17	4340540	22.90	49879834	20.52	−4.49	−18.03
	LTR		6614318	26.16	4219779	22.26	4843776	19.93	−4.28	−17.68
		Gypsy	5694116	22.52	3800409	20.05	4148861	17.07	−4.22	−19.8
		Copia	864519	3.47	419370	2.26	668687	3.10	0.23	8.01
		Unknown	55682	0.22	21937	0.12	25477	0.10	<0.1	−41.18
	LINE		212097	0.84	74864	0.39	122512	0.50	−0.11	−18.7
		L1	100815	0.40	21044	0.11	29646	0.12	−0.14	−52.95
		RTE	111282	0.44	53820	0.28	77699	0.32	<0.1	−11.11
		Unknown	44266	0.18	N/A	N/A	15167	0.06	<0.1	−33.34
	SINE	TS/TS2	N/A	N/A	23960	0.13	21696	0.09	<0.1	38.46
DNA transposon			450969	1.78	327076	1.73	410725	1.69	<0.1	−3.7
	Helitron		129168	0.51	128361	0.68	145124	0.60	<0.1	0.84
	Ac		40235	0.20	38761	0.2	46353	0.19	<0.1	−5
	MuDR		132438	0.52	73586	0.39	95050	0.39	<0.1	−14.29
	EnSpm		84257	0.33	45352	0.24	63659	0.26	<0.1	−8.77
	Harbinger		231	<0.01	N/A	N/A	3075	0.01	<0.1	2551
	TIR	hAT	41797	0.17	26123	0.14	33887	0.14	<0.1	−9.68
	Unknown		12595	0.03	14895	0.08	23577	0.10	<0.1	81.82
35S rDNA[a]			430296	1.70	90945	0.48	42313	0.17	−0.92	−84.4
5S rDNA[ab]			1575	0.01	1750	0.013	4200	0.031	<0.1	150
Satellite		NTRS	0	0	402601	0.21	~80	<0.0001	−0.15	−99
		SYL2[b]	~6750	0.03	0	0	~4730	0.02	<0.1	33

NOTE.—N/A, not available.

[a] The genic but not intergenic regions were estimated here.

[b] These estimates were based on Blast read depth to cloned sequences.

the genome per species. Such coverage theoretically allows the reconstruction of repeat units that have copy number in excess of ~1,000 copies per 1C genome. This approach provides estimates of repeat abundance (copy number and genome proportion estimates) that are in broad agreement with other experimental approaches and have been used to characterize repeats in pea (Macas et al. 2007), soybean (Swaminathan et al. 2007), barley (Wicker et al. 2009), and banana (Hribova et al. 2010).

We observed a relatively low abundance of TEs in the *Nicotiana* genomes analyzed (table 2). *Hordeum vulgare* with a genome size similar to *N. tabacum* (1C = 5,439 Mb) has higher proportions of Ty1-*copia* (~16%) and Ty3-*gypsy* (~30%) elements than *N. tabacum* (Wicker et al. 2009). Likewise, *Zea mays*, with a genome size similar to the progenitor diploids of *N. tabacum*, has a total TE abundance of around 84% of the genome (~46% gypsy elements) (Schnable et al. 2009). However, *N. tabacum* has similar abundance of TEs to *P. sativum* (1C = 4,778 Mb), with 5% and 24% of the genome consisting of Ty1-*copia* and Ty3-*gypsy* elements, respectively (Macas et al. 2007), whereas for the much smaller genome (1C = 490 Mb) of *Oryza sativa* ~4% of the DNA consists of Ty1-*copia* elements and ~11% of Ty3-*gypsy* (International Rice Genome Sequencing Project 2005).

It is possible that our GP estimates of repeat sequences could be low because of an abundance of a diverse range of low copy repeats in the genome. Furthermore, we may be unable to identify repeats because RepBase has only 67 Solanaceae entries (October 2010), including complete and incomplete Ty1-*copia* and Ty3-*gypsy* elements. However, these interpretations seem unlikely as conserved domains are normally detectable across wide phylogenetic distances. Importantly, on lowering the Blast threshold to $e\ \ 10^{-2}$, we observed only a small (3–5%) increase in the total proportion of the genome that was annotated (data not shown). Further support for our estimates of TE abundance in *N. tabacum* (~22% of genome) is provided by reassociation kinetics of long DNA fragments, where only 25% of the genome is shown to consist of repetitive DNA in excess of 1,500 bp in length (Zimmerman and Goldberg 1977).

The satellite repeat content of *Nicotiana* genomes seems to be underestimated in 454 data sets (6-fold for NTRS and >10-fold for HRS60) as compared with Southern blot estimates (Koukalova et al. 1989; Matyasek et al. 1997). There is also a 2- and 10-fold variation in read depths along the NTRS and HRS60 monomers, respectively (supplementary fig. 3, Supplementary Material online). Nevertheless, although the absolute numbers of tandem repeats may be underestimated in our 454 data sets, the relative abundance of these repeats in the three species of *Nicotiana* is fully concordant (data not shown).

Angiosperm Genome Dynamism

The genomes of angiosperms are thought to be highly dynamic in relation to other eukaryotes, in particular mammals (Kejnovsky et al. 2009), and many studies have

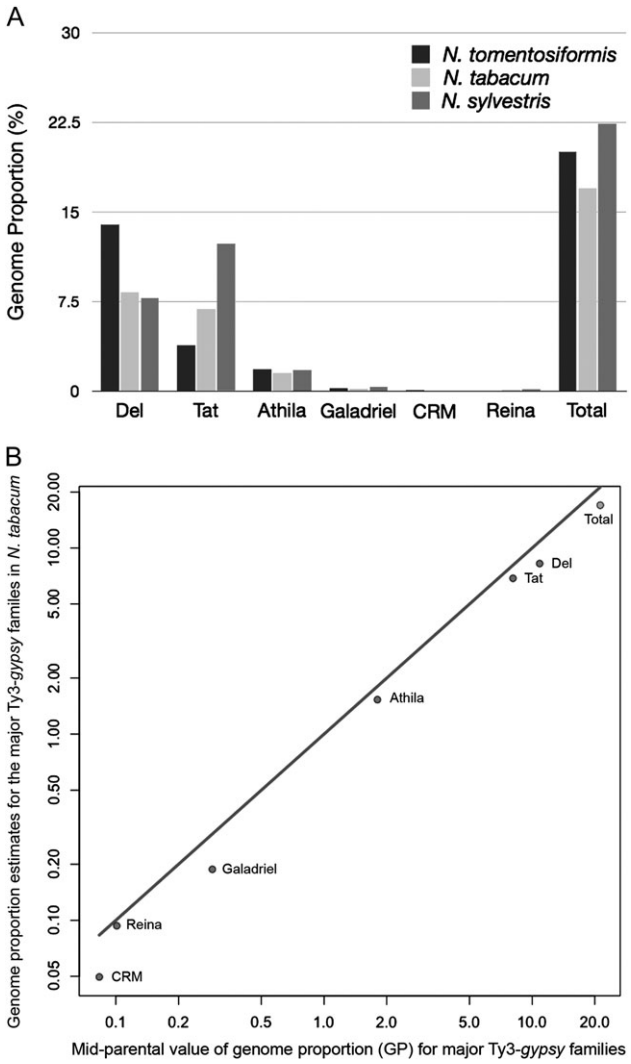

FIG. 5. (*A*) Histogram showing genome proportions of the major Ty3-*gypsy* families detected in the genomes of three *Nicotiana* species. The Ty3-*gypsy* clades are indicated along the *x* axis, with the proportion in each species given as a percentage of the genome on the *y* axis. (*B*) Genome proportion values of Ty3-*gypsy* families for *Nicotiana tabacum* compared with the expected value (parental average, blue line), plotted on a log scale. Genome proportion estimates are below the expected value for all Ty3-*gypsy* families observed.

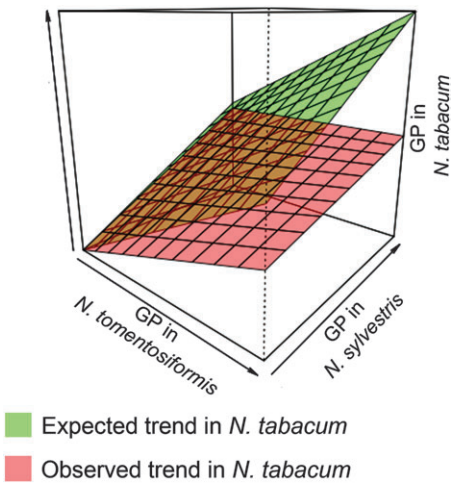

FIG. 6. We used linear regression analysis to generate a smoothed contour plot of genome proportion (GP) values for clusters in each of the three *Nicotiana* species (red surface, for a description of this analysis, see Materials and Methods). The expected plane assuming faithful inheritance of sequences in *Nicotiana tabacum* is shown (green surface). The observed trend is for repeat clusters in *N. tabacum* to be underrepresented compared with expected, particularly from clusters that are abundant in *Nicotiana tomentosiformis*. The data range is for clusters with a GP between 0% and 0.5% on each axis.

suggested that hybridization can trigger extensive genetic change in some young hybrids and allopolyploids (Comai et al. 2003; Adams and Wendel 2005; Leitch and Leitch 2008; Feldman and Levy 2009; Parisod et al. 2009; Paun et al. 2010; Petit et al. 2010), although the extent of change is variable between taxa (Baumel et al. 2002).

Occurrence of repeats with a high degree of sequence similarity can be indicative of either repeat expansion or homogenization, and previous research has used this signature to identify recent episodes of repeat turnover (Kim et al. 1998; Jordan and McDonald 1999). There is an indication that expansion of repeats and/or homogenization is occurring in the diploid genome of *N. sylvestris* because there are many repeats with high mean sequence similarity to related sequences in the same genome (fig. 2*A*). Some sequences in *N. sylvestris* showing evidence of recent expansion belong to the Ty3-*gypsy* superfamily, other TE groups, and rDNA sequences (supplementary table 1, Supplementary Material online). We know that rDNA units can be homogenized astonishingly rapidly in *Nicotiana*; in synthetic *N. tabacum*, thousands of units were converted to a novel type in just a few generations (Skalicka et al. 2003).

Importantly, only a few sequences have the signature of high mean sequence similarity in the genomes of *N. tomentosiformis* (fig. 2*B*) and *N. tabacum* (fig. 2*D*). The low levels of these sequences in *N. tabacum* perhaps indicates homogenization/expansion of repeats in *N. sylvestris* after the formation of *N. tabacum* or that these sequences have been eliminated from the *N. tabacum* genome. There appears not to be large-scale repeat expansion following polyploidy in *N. tabacum*, as also observed for retroelements in the polyploid *Gossypium hirsutum* (cotton) which formed 1–2 million years ago (Hu et al. 2010). Some reads that do have high mean sequence similarity in *N. tabacum* are rDNA sequences. Such an observation is expected considering that 35S rDNA has been recently homogenized in this species (Volkov et al. 1999; Lim, Kovarik, et al. 2000; Kovarik et al. 2008). However, the number of rDNA repeats is much lower in *N. tabacum* compared with its progenitors (table 2), explaining why we find only a few copies.

Genome Downsizing in Polyploids

The 1C genome size (GS) of *N. tabacum* (5,100 Mb) is ~ 3.7% less than would be expected by summing the sizes of the two parental genomes (*N. sylvestris* = 2,650 Mb, *N. tomentosiformis* = 2,650 Mb), which has been proposed as evidence of genome downsizing in this species (Bennett

and Leitch 2005; Leitch et al. 2008). Analysis of more than 3,000 diploid and polyploid angiosperm species indicated that genome downsizing following polyploidy may be a common, although not ubiquitous occurrence (Leitch and Bennett 2004). Indeed one allotetraploid of *Nicotiana* has been shown to have a larger that predicted based on the sum of their parent's genomes (Leitch et al. 2008).

The following sequence types have materially lower GP in *N. tabacum* than expected compared with the diploid progenitors. 1) Retroelements overall are reduced in *N. tabacum* from expected GP by 18.0% (expectation is the mean GP of the two diploid parents, reflecting the equivalence of GS in the progenitors) of which Ty3-*gypsy* elements are lower by 19.8% (table 2). All major Ty3-*gypsy*-like families detected exhibit a reduction in GP from expectation (fig. 5*B*), indicating a general reduction in their abundance in *N. tabacum* (or increase in one or both progenitors). 2) 35S rDNA exhibits a reduction of GP in *N. tabacum* by 84% from expectation (table 2). All observations are in line with the genome downsizing hypothesis and the disparity between the expected genome size of *N. tabacum* (1C = 5,300 Mb) and that estimated by flow cytometry (1C = 5,100 Mb) (Leitch et al. 2008). 3) In addition, using Southern hybridization data, Skalicka et al. (2005) determined the proportion of NTRS satellite repeats to be 3% of the genome in *N. tomentosiformis*, 0% in *N. sylvestris*, and 0.5% in *N. tabacum* (cultivar SR1), a 85% negative deviation from expectation assuming *N. tabacum* inherited units in the abundance found in the diploids.

Nicotiana tabacum may have lost ~200 Mb of DNA, much of which could be accounted for by the sequences described here. Regression analysis of GPs in *N. tabacum* and the progenitor species reveals the general trend is for repeats to be underrepresented in *N. tabacum* compared with expected (fig. 6) These data indicate that sequence loss seems to be particularly prevalent among the most common repeat types. The large underrepresentation of the Ty3-*gypsy*-like sequences may simply be a consequence of their high abundance.

Loss of Repeats from Paternally Derived T-Genome of *N. tabacum*

From cytogenetic analysis of chromosomes, Gill (1991) proposed in the nuclear cytoplasmic interaction hypothesis that the paternally inherited genome of an allopolyploid is more prone to genetic change than the maternally derived genome. In support of this hypothesis, Southern blot (Skalicka et al. 2005) and cytogenetic (Koukalova et al. 1989; Lim, Matyasek, et al. 2000) data have revealed that in *N. tabacum*, the subtelomeric satellite repeat HRS60, inherited from *N. sylvestris*, does not deviate from expectation. In contrast, Skalicka et al. (2005) reported that four families of repeats inherited from *N. tomentosiformis* were in lower copy number in *N. tabacum*. In addition, there is preferential loss of Tnt1 insertions from the *N. tomentosiformis*–derived genome in synthetic *N. tabacum* (Petit et al. 2010) and of Tnt2 elements in natural *N. tabacum* (Petit et al. 2007). We report here that the most well-represented cluster in *N. tomentosiformis* (GP = 1.91%) is CL3, and this cluster is in much lower abundance in *N. tabacum* (GP= 0.1%), perhaps indicative of sequence loss in the T-genome of *N. tabacum*.

On a genome-wide scale, we provide evidence that the T-genome of *N. tabacum* has experienced preferential sequence loss compared with the S-genome. Cross-species sequence similarity analyses (fig. 1*B*) showed many sequences in *N. tomentosiformis* that deviate substantially from the 2:1 line and are less well represented in *N. tabacum* than expected. Although *N. sylvestris* has a greater number of sequences that are underrepresented in *N. tabacum* (fig. 1*A*), they do not deviate substantially from expectation and the sum of residuals is less than that observed for *N. tomentosiformis*, indicating that repeats derived from *N. sylvestris* may not to be affected to the same degree as those of the T-genome.

We compared a smoothed contour plot that best describes the observed GPs in *N. tabacum* with a theoretical surface that assumes an average GP of that found in the parents GPs (fig. 6). Linear regression analysis revealed unequal contribution from the progenitor species to the *N. tabacum* genome, with particular underrepresentation of repeats from *N. tomentosiformis* (fig. 6), regardless of their abundance in *N. sylvestris*. The degree to which a repetitive DNA sequence is underrepresented in *N. tabacum* increases with the corresponding abundance in *N. tomentosiformis*. In contrast for clusters that were absent or had a low abundance in *N. tomentosiformis*, GP values were close to expectation (fig. 6) indicating faithful inheritance from *N. sylvestris*.

The lack of evidence for repeat expansion in *N. tomentosiformis* (fig. 2) indicates that the underrepresentation of *N. tomentosiformis*–derived repeat sequences in *N. tabacum* is likely to be the result of sequence erosion in the latter rather than sequence gains in the former postallotetraploidy. Collectively, our data are congruent with the loss of repeats derived from the T-genome of *N. tabacum* and the maintenance of repeats from the S-genome.

Instability of a particular genomic component of an allopolyploid has also been observed in crosses between *Brassica nigra* and *Brassica rapa*, where directional loss of restriction fragments was apparent in some lines (Song et al. 1995). Mechanisms that induce the preferential elimination of paternally derived sequences are unknown. Potentially, they could involve the 24 nt class of siRNAs that are known to be highly abundant and uniparentally expressed in the endosperm (Baulcombe 2009). Perhaps these small RNAs provide an opportunity for the epigenetic modification of *N. sylvestris*–derived repeats, which were maternally inherited. The small RNAs may have acted to enhance the stability and reduce the frequency of DNA loss from the S-genome of *N. tabacum* in early generations after its formation.

Conclusion

What is apparent from these data is that not all repeats have responded in the same way subsequent to allopolyploidy. The processes giving rise to repeat copy number

changes may be stochastic or directed. Evidence for the latter may be the preferential loss of repeats from the *N. tomentosiformis*–derived T-genome of *N. tabacum*. Our data suggest that the three species studied have experienced distinctive patterns of genome evolution. *Nicotiana sylvestris* and *N. tomentosiformis* probably diverged early in the evolution of genus *Nicotiana* which split from *Symonanthus* around 15 million years ago (Chase et al. 2003; Clarkson et al. 2005). *Nicotiana sylvestris*, the maternal progenitor of *N. tabacum*, appears to have many features indicating recent expansion of repeats. In contrast, this signature is much less apparent in *N. tomentosiformis*, suggesting that its genome has not experienced substantial rounds of homogenization/amplification in its recent evolutionary history. In *N. tabacum*, which formed less than 0.2 million years ago (Clarkson et al. 2005) there is evidence for a third pattern of genome evolution, that of genome reduction. In a broad sense, the three genomes of *Nicotiana* studied here are experiencing different evolutionary trajectories, and the same families of repeats, in *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum* are dynamic, stable, and downsizing, respectively, indicating the same families of repeats can have varying fates in different, yet closely related lineages.

Supplementary Material

Supplementary figures 1–3 and table 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank NERC and the grants OC10037 from the Ministry of Education, Youth and Sports of the Czech Republic and AVOZ50510513 from the Academy of Sciences of the Czech Republic for supporting the project. We thank two anonymous referees for their valuable comments, R. Verity for assistance with the plotting of some of the data using R. In addition, we thank B. Chalhoub, R. Buggs, and L. Kelly for their insightful comments and useful discussion.

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol*. 8:135–141.

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien M-A, Fukunaga K, Ricou M, Misset MT. 2009. Hybridization, polyploidy and invasions: lessons from *Spartina* (Poaceae). *Biol Invasions*. 11:1159–1173.

Baulcombe D. 2009. The diverse roles of small, non-coding RNA in plants. *Mech Dev*. 126:S24.

Baumel A, Ainouche M, Kalendar R, Schulman AH. 2002. Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* CE Hubbard (Poaceae). *Mol Biol Evol*. 19:1218–1227.

Beaulieu J, Jean M, Belzile F. 2009. The allotetraploid *Arabidopsis thaliana–Arabidopsis lyrata* subsp *petraea* as an alternative model system for the study of polyploidy in plants. *Mol Genet Genomics*. 281:421–435.

Bennett MD, Leitch IJ. 2005. Angiosperm DNA C-values database. Richmond, UK.

Brookfield JFY. 2005. The ecology of the genome—mobile DNA elements and their hosts. *Nat Rev Genet*. 6:128–136.

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

Clarkson JJ, Kelly LJ, Leitch AR, Knapp S, Chase MW. 2010. Nuclear glutamine synthetase evolution in *Nicotiana*: phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Mol Phylogenet Evol*. 55:99–112.

Clarkson JJ, Lim KY, Kovarik A, Chase MW, Knapp S, Leitch AR. 2005. Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol*. 168:241–252.

Comai L, Madlung A, Josefsson C, Tyagi A. 2003. Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philos Trans R Soc Lond B Biol Sci*. 358:1149–1155.

Dolezel J, Greilhuber J, Lucretti S, Meister A, Lysak MA, Nardi L, Obermayer R. 1998. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann Bot*. 82:17–26.

Feldman M, Levy AA. 2009. Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. *J Genet Genomics*. 36:511–518.

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

Gill BS. 1991. Nucleocytoplasmic interaction (NCI) hypothesis of genome evolution and speciation in polyploid plants. In: Sasakuma T, Kinoshita T, editors. Kihara Memorial International Symposium on cytoplasmic engineering in wheat. Yokohama (Japan): Kihara Memorial Foundation. p 48–53.

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

Hu GJ, Hawkins JS, Grover CE, Wendel JF. 2010. The history and disposition of transposable elements in polyploid *Gossypium*. *Genome* 53:599–607.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800.

Jordan IK, McDonald JF. 1999. Tempo and mode of Ty element evolution in *Saccharomyces cerevisiae*. *Genetics* 151:1341–1351.

Kashkush K, Feldman M, Levy AA. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet*. 33:102–106.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol Evol*. 24:572–582.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res*. 8:464–478.

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

Koukalova B, Reich J, Matyasek R, Kuhrova V, Bezdek M. 1989. A BamHI family of highly repeated DNA sequences of *Nicotiana tabacum*. *Theor Appl Genet*. 78:77–80.

Kovarik A, Dadejova M, Lim YK, Chase MW, Clarkson JJ, Knapp S, Leitch AR. 2008. Evolution of rDNA in *Nicotiana* allopolyploids: a potential link between rDNA homogenization and epigenetics. *Ann Bot*. 101:815–823.

Leitch AR, Leitch IJ. 2008. Perspective genomic plasticity and the diversity of polyploid plants. *Science* 320:481–483.

Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biol J Linn Soc*. 82:651–663.

Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann Bot*. 101:805–814.

Lim KY, Kovarik A, Matyasek R, Bezdek M, Lichtenstein CP, Leitch AR. 2000. Gene conversion of ribosomal DNA in *Nicotiana tabacum* is associated with undermethylated, decondensed and probably active gene units. *Chromosoma* 109:161–172.

Lim KY, Kovarik A, Matyasek R, Chase MW, Clarkson JJ, Grandbastien MA, Leitch AR. 2007. Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytol*. 175:756–763.

Lim KY, Matyasek R, Kovarik A, Leitch AR. 2004. Genome evolution in allotetraploid *Nicotiana*. *Biol J Linn Soc*. 82:599–606.

Lim KY, Matyasek R, Lichtenstein CP, Leitch AR. 2000. Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section *Tomentosae*. *Chromosoma* 109:245–258.

Lim KY, Skalicka K, Koukalova B, Volkov RA, Matyasek R, Hemleben V, Leitch AR, Kovarik A. 2004. Dynamic changes in the distribution of a satellite homologous to intergenic 26-18S rDNA spacer in the evolution of *Nicotiana*. *Genetics* 166:1935–1946.

Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. 2001. Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome* 44:321–330.

Liu B, Wendel JF. 2000. Retrotransposon activation followed by rapid repression in introgressed rice plants. *Genome* 43:874–880.

Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol*. 29:365–379.

Llorens C, Futami R, Bezemer D, Moya D. 2008. The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res*. 26:38–46.

Ma JX, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res*. 14:860–869.

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

Macas J, Neumann P, 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.

Margulies M, Egholm M, Altman WE, et al. (56 co-authors). 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.

Matyasek R, Fulnecek J, Lim KY, Leitch AR, Kovarik A. 2002. Evolution of 5S rDNA unit arrays in the plant genus *Nicotiana* (Solanaceae). *Genome* 45:556–562.

Matyasek R, Gazdova B, Fajkus J, Bezdek M. 1997. NTRS, a new family of highly repetitive DNAs specific for the T1 chromosome of tobacco. *Chromosoma* 106:369–379.

Matyasek R, Lim KY, Kovarik A, Leitch AR. 2003. Ribosomal DNA evolution and gene conversion in *Nicotiana rustica*. *Heredity*. 91:268–275.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* 226:792–801.

Murad L, Lim KY, Christopodulou V, Matyasek R, Lichtenstein CP, Kovarik A, Leitch AR. 2002. The origin of tobacco's T genome is traced to a particular lineage within *Nicotiana tomentosiformis* (Solanaceae). *Am J Bot*. 89:921–928.

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

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol*. 186:37–45.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *New Phytol*. 184:1003–1015.

Paun O, Bateman RM, Fay MF, Hedren M, Civeyrel L, Chase MW. 2010. Stable epigenetic effects impact adaptation in allopolyploid orchids (*Dactylorhiza: Orchidaceae*). *Mol Biol Evol*. 27:2465–2473.

Petit M, Guidat C, Daniel J, et al. (11 co-authors). 2010. Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytol*. 186:135–147.

Petit M, Lim KY, Julio E, Poncet C, de Borne FD, Kovarik A, Leitch AR, Grandbastien MA, Mhiri C. 2007. Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). *Mol Genet Genomics*. 278:1–15.

R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Schnable PS, Ware D, Fulton RS , et al. (157 co-authors) 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115.

Skalicka K, Lim KY, Matyasek R, Koukalova B, Leitch AR, Kovarik A. 2003. Rapid evolution of parental rDNA in a synthetic tobacco allotetraploid line. *Am J Bot*. 90:988–996.

Skalicka K, Lim KY, Matyasek R, Matzke M, Leitch AR, 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.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot*. 96:336–348.

Song KM, Lu P, Tang KL, Osborn TC. 1995. Rapid genome chane in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proc Natl Acad Sci U S A*. 92:7719–7723.

Swaminathan K, Varala K, Hudson ME. 2007. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*. 8:132–145.

Volkov RA, Borisjuk NV, Panchuk II, Schweizer D, Hemleben V. 1999. Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. *Mol Biol Evol*. 16:311–320.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol Biol*. 42:225–249.

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*. 7:275.

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, 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.

Zimmerman JL, Goldberg RB. 1977. DNA sequence organization in the genome of *Nicotiana tabacum*. *Chromosoma* 59:227–252.