Next generation sequencing reveals genome downsizing in allotetraploid

Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs

Research Article

Simon Renny-Byfield¹, Michael Chester¹,², Ales Kovařík³, Steven C Le Comber¹, Marie-Angèle Grandbastien⁴, Marc Deloger⁴, Richard Nichols¹, Jiri Macas⁵, Petr Novák⁵, Mark W. Chase⁶, Andrew R. Leitch¹*

¹ Queen Mary University of London, School of Biological and Chemical Sciences, Mile End Road, London, E1 4NS, UK; ² Laboratory of Molecular Systematics and Evolutionary Genetics, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611-7800 USA; ³ Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i Kralovopolska 135, CZ-612 65 Brno; ⁴ Institute Jean-Pierre Bourgin, INRA-Versailles, 78026, France; ⁵ Biology Centre ASCR, Institute of Plant Molecular Biology, Branišovská 31, České Budějovice, CZ-37005, Czech Republic; ⁶ Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK.

*author for correspondence: Andrew R Leitch

e-mail: a.r.leitch@qmul.ac.uk

telephone: +44 (0)20 7882 5294

Keywords: Next generation sequencing, allopolyploidy, genome-downsizing, transposable elements, retroelements, paternal genome, Nicotiana tabacum, Nicotiana sylvestris, Nicotiana tomentosiformis

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
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, *N. sylvestris* (maternal, S-genome donor) and *N. 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 homogenisation, 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.

**Accession numbers:** 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.
Supplementary information: Three Figures
One Table
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 hybridisation 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; Matyasek et al. 2003; Adams and Wendel 2005; Leitch et al. 2008). However there are only a few examples of allopolyploid- or interspecific hybridisation-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 N. 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
MITEs 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 (SSAP) analysis of *Arabidopsis thaliana × A. lyrata* revealed the CAC family of transposons was not activated in neo-tetraploids (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 BAC 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* hybridisation (GISH) fails, even in some
relatively young *Nicotiana* alloployploids where loss of homology with progenitor species can be detected after only c. 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 *Dactylorthiza* allopolyplploid 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. 2009b; Paun et al. 2010).

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

Previous molecular and cytogenetics studies have suggested that for non-coding 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 et al. 2000a). 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. 2004b). 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 et al. 2004a). 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; Wicker et al. 2009). More recently, Hribova et al. (2010) have used 454 read-depth analysis to characterise 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 analyse 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 (AFLPs, 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, UK. DNA was randomly sheared by nebulization and sequenced using a 454 GS FLX Instrument with Titanium reagents (Roche Diagnostics). For each species we used 1/8th of a 70 x 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 <10^{-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 pair-wise analysis was performed on the N. tabacum dataset and the proportion of reads with significant hits (E-value < 10^{-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 datasets, and for each N. tabacum read the number of sequences (from the progenitor dataset) with significant sequence similarity hits to the N. tabacum reads was recorded. Due to the number of reads in each dataset being unequal, the number of hits recorded in all cases was standardised to the N. tabacum dataset, where hit numbers were scaled up or down depending on the difference in the number of reads between datasets. For example the N. tabacum dataset consists of 70,616 reads, whereas the N. tomentosiformis dataset has 65,858 reads and to standardise these data the number of hits recorded for N. tomentosiformis was multiplied by 1.072 (number of reads in the N. tabacum dataset/number of reads in the N. tomentosiformis dataset). We constructed an *in silico* N. tabacum consisting of a random set of 35,000 reads from each of the parental datasets. 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).

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 Figure 1).

**Genome-wide analysis of mean similarity of related sequences**

We analysed the genomes of parental and progenitor species by comparisons of the mean sequence similarity of related sequences. BLAST analysis with an e-value cut-off of $10^{-6}$ 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 visualisation and species-specific reference assembly**

Repeat sequence assembly was performed with a combined dataset 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 pair-wise 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 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 analysed 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 analysed 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 cut-off score of 10⁻⁶. Additional annotation using the BLAST function on the GyDB was required in order to establish
the clade to which Ty1 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 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 2650 Mb) genome, ~0.8% coverage of the *N. tomentosiformis* (1C genome size of 2650 Mb) genome and ~0.5% coverage for *N. tabacum* (1C genome size of ~5100 Mb) (Leitch et al. 2008). Sequence reads were submitted to NCBI sequence read archive (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. 1A and B). The output reflects the abundance of sequences in the *Nicotiana* genomes. We would expect those sequences that were faithfully inherited by *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 normalised datasets and the effective dilution by the other parental genome. Those sequences falling above the 2:1 line are likely unrepresented 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.

Fig. 1 A and B show complete pair-wise sequence similarity analysis of individual 454 reads from *N. tabacum* against all reads in the *N. tabacum*, *N. sylvestris* and *N. tomentosiformis* datasets. In Fig. 1 A, which shows the analysis of *N. tabacum* against the *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* dataset, sequences on the 2:1 line are less abundant (Fig. 1 B).

The analysis in Fig. 1A 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 Fig. 1 B. This spike was absent in a control genome, generated *in silico* from an equal mixture (35,000 reads) from each parental dataset (totalling 70,000 reads) (Supplementary Fig. 1, A, B).

In Fig. 1 B, 3078 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 Figure 1, B). 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 68332, respectively. In *N. sylvestris* there are 5919 sequences above the 2:1 line, but the mean of residual of these sequences is only 9.6 and the sum totals 56822. Amongst the reads above the 2:1 line in Fig. 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 datasets we calculated the mean sequence similarity between that read and all related sequences within the same dataset. 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* dataset) above 0.98 and a coefficient of variation less than one (2,248 reads in total) were identified (Supplementary Table 1) and shown to include 5S and 35S rDNA sequences (totalling 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 dataset (>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 normalised (by total number of reads in the dataset) contribution of each species to the 30 largest clusters is shown in Supplementary Figure 2. 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 characterisation of the average read-depth along the length of the contig (RD), genome representation (GR, calculated as RD \times contig length) and genome proportion (GP, calculated as (GR/total size of the dataset in base pairs) \times 100) for each of the three species. GP is therefore the percentage of the dataset (and therefore the genome) that can be attributed to a given repeat. This allowed characterisation of the most abundant repeats in the three genomes. Others have used similar approaches to
measure repeat sequence abundance in the genomes of pea and soybean (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 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 92-96% identity with the consensus (data not shown).

Characterising 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 were 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) reveal the GP of retroelements to be reduced by over 18% from expectation. The majority of retroelements are Ty3-gypsy-like (estimates...
ranging from 17-23% 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 to 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 14634 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. 6B, C). The actual estimate of the slope (Fig. 6A, C) was close to, although significantly different from, the expected slope of 0.5 for *N. tabacum* versus *N. sylvestris* (0.472, 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 under-represented along the length of the data range, and. [3] this discrepancy is greatest for the most common repeat elements in *N. tomentosiformis*. 
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 the genome per species. Such coverage theoretically allows the reconstruction of repeat units that have copy number in excess of ~1000 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 characterise 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 analysed (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 Pisum sativum (1C = 4778 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 $1 \times 10^{-2}$ 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 1500 bp in length (Zimmerman and Goldberg 1977).

The satellite repeat content of *Nicotiana* genomes seems to be underestimated in 454 datasets (6-fold for NTRS and >10-fold for HRS60) as compared to Southern blot estimates (Koukalova et al. 1989; Matyasek et al. 1997). There is also a two- and ten-fold variation in read depths along the NTRS and HRS60 monomers, respectively (Supplementary Figure 3). Nevertheless, although the absolute numbers of tandem repeats may be underestimated in our 454 datasets, the relative abundance of these repeats in the three species of *Nicotiana* are 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 suggested that hybridisation 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 homogenisation, 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 are occurring in the diploid genome of *N. sylvestris* since 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). We know that rDNA units can be homogenised 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. 2D). 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 homogenised in this species (Volkov et al. 1999; Lim et al. 2000a; Kovarik et al. 2008). However the
number of rDNA repeats is much lower in *N. tabacum* compared to its progenitors (Table 2), explaining why we find only a few copies.

**Genome downsizing in polyploids**

The 1C genome size (GS) of *N. tabacum* (5100 Mb) is ~ 3.7% less than would be expected by summing the sizes of the two parental genomes (*N. sylvestris* = 2650 Mb, *N. tomentosiformis* = 2650 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 3000 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 than 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 to 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 = 5300 Mb) and that estimated by flow cytometry (1C = 5100 Mb) (Leitch et al. 2008). [3] In addition, using
Southern hybridisation 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 to 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.

$\textit{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 et al. 2000b) 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 under representation of repeats from *N. tomentosiformis* (Fig. 6), regardless of their abundance in *N. sylvestris*. The degree to which a repetitive DNA sequence is under-represented 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 post allotetraploidy. Collectively, our data are congruent with the loss of repeats derived from the T-genome of \textit{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 \textit{Brassica nigra} and \textit{B. 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 24nt class of siRNAs that are known to be highly abundant and uni-parentally expressed in the endosperm (Baulcombe 2009). Perhaps these small RNAs provide an opportunity for the epigenetic modification of \textit{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 \textit{N. tabacum} in early generations after its formation.

\textit{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 \textit{N. tomentosiformis} derived T-genome of \textit{N. tabacum}. Our data suggest that the three species studied have experienced distinctive patterns of genome evolution. \textit{Nicotiana sylvestris} and \textit{N. tomentosiformis} probably diverged early in the evolution of genus \textit{Nicotiana} which split from \textit{Symonanthus} around 15 million years ago (Chase et al. 2003; Clarkson et al. 2005). \textit{Nicotiana sylvestris}, the maternal progenitor of \textit{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 homogenisation/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 down-sizing, respectively indicating the same families of repeats can have varying fates in different, yet closely related lineages.

**Acknowledgements**

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.
Literature cited


satellite homologous to intergenic 26-18S rDNA spacer in the evolution of *Nicotiana*. Genetics **166**:1935-1946.


Macas, J., P. Neumann, and A. Navratilova. 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**.


Table 1
Details of the output of the clustering and assembly algorithm for the combined dataset (producing the reference sequences) and the species specific read mapping analyses.

<table>
<thead>
<tr>
<th></th>
<th>Number of clusters in assembly</th>
<th>Number of contigs in assembly</th>
<th>Minimum/Maximum contig length(bp)</th>
<th>% of reads mapped to contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>combined assembly</td>
<td>16229</td>
<td>17443</td>
<td>107/9632</td>
<td>N/A</td>
</tr>
<tr>
<td>N. tabacum</td>
<td>8496</td>
<td>10464</td>
<td>109/5198</td>
<td>44</td>
</tr>
<tr>
<td>Species specific read mapping</td>
<td>N. sylvestris</td>
<td>9791</td>
<td>11446</td>
<td>107/5198</td>
</tr>
<tr>
<td>N. tomentosiformis</td>
<td>6131</td>
<td>7378</td>
<td>108/9362</td>
<td>53</td>
</tr>
</tbody>
</table>

1The repeat identification algorithm is described in detail in Novak et al. (2010).
Table 2 Genome proportion (GP) of major repeat classes within the N. tomentosiformis, N. sylvestris and N. tabacum genomes.
<table>
<thead>
<tr>
<th>Class</th>
<th>Order</th>
<th>Superfamily</th>
<th>N. sylvestris Genome Representation (%)</th>
<th>N. tomentosiformis Genome Representation (%)</th>
<th>N. tabacum Genome Representation (%)</th>
<th>Deviation from parental average (%)</th>
<th>Difference as % of expected GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retroelement</td>
<td></td>
<td></td>
<td>6870681</td>
<td>4340540</td>
<td>49879834</td>
<td>-4.49</td>
<td>-18.03</td>
</tr>
<tr>
<td>LTR</td>
<td></td>
<td></td>
<td>6614318</td>
<td>4219779</td>
<td>4843776</td>
<td>-4.28</td>
<td>-17.68</td>
</tr>
<tr>
<td>Gypsy</td>
<td></td>
<td></td>
<td>5694116</td>
<td>3800409</td>
<td>4148861</td>
<td>-4.22</td>
<td>-19.8</td>
</tr>
<tr>
<td>Copia</td>
<td></td>
<td></td>
<td>864519</td>
<td>419370</td>
<td>668687</td>
<td>0.23</td>
<td>8.01</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td>55682</td>
<td>21937</td>
<td>25477</td>
<td>&lt;0.1</td>
<td>-41.18</td>
</tr>
<tr>
<td>LINE</td>
<td></td>
<td></td>
<td>212097</td>
<td>74864</td>
<td>122512</td>
<td>-0.11</td>
<td>-18.7</td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td></td>
<td>100815</td>
<td>21044</td>
<td>29646</td>
<td>-0.14</td>
<td>-52.95</td>
</tr>
<tr>
<td>RTE</td>
<td></td>
<td></td>
<td>111282</td>
<td>53820</td>
<td>77699</td>
<td>&lt;0.1</td>
<td>-11.11</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td>44266</td>
<td>N/A</td>
<td>15167</td>
<td>&lt;0.1</td>
<td>-33.34</td>
</tr>
<tr>
<td>SINE TS/TS2</td>
<td>DNA Transposon</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>23960</td>
<td>0.13</td>
<td>38.46</td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td></td>
<td>129168</td>
<td>128361</td>
<td>145124</td>
<td>&lt;0.1</td>
<td>0.84</td>
</tr>
<tr>
<td>Region</td>
<td>Genic</td>
<td>Intergenic</td>
<td>Genic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>------------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ac</td>
<td>40235</td>
<td>38761</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MuDR</td>
<td>132438</td>
<td>73586</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EnSpm</td>
<td>84257</td>
<td>45352</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harbinger</td>
<td>231</td>
<td>N/A</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIR hAT</td>
<td>41797</td>
<td>26123</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>12595</td>
<td>14895</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35S rDNA</td>
<td>430296</td>
<td>90945</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5S rDNA</td>
<td>1575</td>
<td>1750</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satellite</td>
<td>0</td>
<td>0</td>
<td>~80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYL2²</td>
<td>~6750</td>
<td>0</td>
<td>~4730</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 The genic, but not intergenic regions were estimated here.
2 These estimates were based on BLAST read depth to cloned sequences.
Figure 3

- **N. tabacum**: 1149
- **N. tomentosiformis**: 1816
- **N. sylvestris**: 3950
- **Plastid clusters**: 1445

Numbers inside the circles indicate the number of corresponding elements.
Figure 6

- **Green** Expected trend in *N. tabacum*
- **Red** Observed trend in *N. tabacum*
Figure Legends

Figure 1

Genome comparisons using pair-wise similarity analysis of individual 454 reads. *Nicotiana tabacum* reads compared to the *N. tabacum* dataset (x-axis), and *N. sylvestris* (A) or *N. tomentosiformis* (B) datasets (y-axis) using the BLAST program with e-value cutoff of 1e-8. The number of similarity hits was normalised to take into account the varying size of each dataset. Reads highlighted red and green are rDNA and NTRS sequences; respectively 1:1 and 2:1 lines are labelled 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.

Figure 2

Histogram showing frequency of mean sequence similarity between each read in a 454 dataset and all related sequences in the same dataset in (A) *N. sylvestris*, (B) *N. tomentosiformis*, (C) an *in silico* *N. tabacum* and (D) natural *N. 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 *N. tabacum* (D).
Figure 3

Venn Diagram where the area of each circle (and the intersections) is proportional to number of 16229 clusters that have significant similarity to sequence reads from the species as indicated. Absolute numbers are given in each section also.

Figure 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 (N. sylvestris), blue (N. tomentosiformis) and purple (N. 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 from N. sylvestris and the other from N. tomentosiformis. For CL2, N. tabacum has both these strings and is additive of the parents.

Figure 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 N. tabacum compared to 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.
Figure 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, see methods for a description of this analysis). The expected plane assuming faithful inheritance of sequences in *N. tabacum* is shown (green surface). The observed trend is for repeat clusters in *N. tabacum* to be under-represented compared to expected, particularly from clusters that are abundant in *N. tomentosiformis*. The data range is for clusters with a GP between 0 – 0.5% on each axis.

**Supplementary Figure Legends**

Supplementary Figure 1

Genome comparisons via complete pair-wise sequence similarity analysis of 454 reads showing the number of sequence similarity hits per sequence. (A) The *in silico* *N. tabacum* dataset against itself on the x-axis, and the *in silico* *N. tabacum* against *N. sylvestris* on the y-axis (standardised to the size of the *N. tabacum* dataset). (B) The *in silico* *N. tabacum* dataset against itself on the x-axis, and the *in silico* *N. tabacum* against *N. tomentosiformis* on the y-axis.

Supplementary Figure 2

Histogram showing the contribution of reads from each species in the 30 largest clusters (as a percentage of the total reads in the combined dataset). These were developed from the clustering algorithm described in detail in Novak et al. (2010). The contribution of the individual species was normalised to reflect the differences in
the sizes of the three species datasets, and RepBase annotations are added to each cluster.

Supplementary Figure 3

Comparisons of copy number estimates (based on 454 read depth) along monomers of two families of satellite repeats (NTRS, T-genome derived; SYL, S-genome derived) in the three *Nicotiana* species.

Supplementary Table 1

A total of 2,248 reads from *N. sylvestris* with a mean sequence similarity > 0.98 and a coefficient of variation < 1 to other related reads in the *N. sylvestris* dataset (see Fig. 2) were subjected to repeat identification. The total of each repeat type identified is given.