Independent, Rapid and Targeted Loss of Highly Repetitive DNA in Natural and Synthetic Allopolyploids of *Nicotiana tabacum*

Simon Renny-Byfield¹, Ales Kovarík², Michael Chester¹,³, Richard A. Nichols¹, Jiri Macas⁴, Petr Novák⁴, Andrew R. Leitch¹* 

¹ School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom, ² Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic, ³ Laboratory of Molecular Systematics and Evolutionary Genetics, Florida Museum of Natural History, University of Florida, Gainesville, Florida, United States of America, ⁴ Biology Centre ASCR, Institute of Plant Molecular Biology, České Budějovice, Czech Republic

**Abstract**

Allopolyploidy (interspecific hybridisation and polyploidy) has played a significant role in the evolutionary history of angiosperms and can result in genomic, epigenetic and transcriptomic perturbations. We examine the immediate effects of allopolyploidy on repetitive DNA by comparing the genomes of synthetic and natural *Nicotiana tabacum* with diploid progenitors *N. tomentosiformis* (paternal progenitor) and *N. sylvestris* (maternal progenitor). Using next generation sequencing, a recently developed graph-based repeat identification pipeline, Southern blot and fluorescence *in situ* hybridisation (FISH) we characterise two highly repetitive DNA sequences (NicCL3 and NicCL7/30). Analysis of two independent high-throughput DNA sequencing datasets indicates NicCL3 forms 1.6–1.9% of the genome in *N. tomentosiformis*, sequences that occur in multiple, discontinuous tandem arrays scattered over several chromosomes. Abundance estimates, based on sequencing depth, indicate NicCL3 is almost absent in *N. sylvestris* and has been dramatically reduced in copy number in the allopolyploid *N. tabacum*. Surprisingly elimination of NicCL3 is repeated in some synthetic lines of *N. tabacum* in their forth generation. The retroelement NicCL7/30, which occurs interspersed with NicCL3, is also under-represented but to a much lesser degree, revealing targeted elimination of the latter. Analysis of paired-end sequencing data indicates the tandem component of NicCL3 has been preferentially removed in natural *N. tabacum*, increasing the proportion of the dispersed component. This occurs across multiple blocks of discontinuous repeats and based on the distribution of nucleotide similarity among NicCL3 units, was concurrent with rounds of sequence homogenisation.

**Introduction**

Polyploidy, where an individual possesses more than a diploid complement of chromosomes, is a fundamental process in the evolution of land plants [1,2,3,4]. Molecular evidence suggests a whole genome duplication (WGD) at the base of all seed plants and another at the base of the angiosperms, and with many lineages having additional WGDs events in their ancestry [5]. Many of these polyploid events are associated with major radiations of land plants [3,5].

The phenomenon of polyploidy is often associated with interspecific hybridisation (allopolyploidy), where divergent genomes are unified within a single nucleus. It has been suggested that this process can induce rapid, reproducible and directional changes to the progenitor sub-genomes [6,7,8,9,10,11]. Analysis of wheat F₁ hybrids has revealed preferential loss of sequences from one of the progenitor genomes, as well as reproducible loss of DNA sequences across independently synthesised neo-tetraploids [12]. For example second-generation neo-tetraploids of a cross

between *Aegilops tauschii* × *Triticum turgidum* have shown elimination of a sequence derived from *A. tauschii* in a tissue specific manner, likely to have occurred during embryo development [13]. Similarly, abundance estimates for repetitive DNA in the genomes of the allopolyploid *Nicotiana tabacum* (formed 20,000–200,000 years ago) and its diploid progenitors indicate the preferential elimination of paternally derived DNA, contributing to genome downsizing thought to have occurred in this species [14,15]. A comparable pattern is observed in synthetic *N. tabacum* Th37 lines, produced in the 1970s [16].

The emergence of high throughput DNA sequencing [17] has allowed the analysis of highly repetitive sequences in the genomes of several angiosperm species including banana, pea, soybean, barley, *Silene latifolia* as well as allopolyploid *N. tabacum* and its diploid progenitors [15,18,19,20,21,22]. Here we examine *Nicotiana tabacum* and progenitors *N. sylvestris* (maternal S-genome donor) and *N. tomentosiformis* (paternal T-genome donor) focusing on the genomic organisation and abundance of two novel repeat families, NicCL3 and NicCL7/30. We used high throughput DNA
sequencing to determine if these repeats are inherited in an additive manner, and to assess any changes in their organisation following allopolyploidy.

Materials and Methods

Plant material

The following accession were used; [1] Nicotiana tabacum cv. SR1 Petit Havana and cv. 095-55, [2] Nicotiana sylvestris Speg. & Comes ac. ITB626 both originating from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France. [3] Nicotiana tomentosiformis Goodsp. ac. NIC 479/84 (Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany), TW142 (USD, North Carolina State University, Raleigh, NC, USA) and Nee et al. 51771 (New York Botanic Gardens). [4] Synthetic N. tabacum Th37 lines in generation four, generated by Burk [23] and characterized previously [16,24]. [5] Synthetic N. tabacum TR1-A in generation S0, generated and characterized by Lim et al. [25]. [6] Diploid species in section Tomentosa [26] derived from USDA: N. tomentosa Ruiz and Pav.; N. kawakamii Y. Ohashi; N. tomentosiformis Goodsp.; and N. glauca Lindl. (section Undulatae, formerly section Tomentosa [27,28]).

High-throughput sequencing of genomic DNA

We used Roche 454 FLX pyrosequencing (454 sequencing) as generated in Renny-Byfield et al. [15]. Sequence reads are deposited in the NCBI sequence read archive (SRA) under the study accession number SRA023759. We sequenced here, using the Illumina Genome Analyzer xII at The Genome Centre Queen Mary University of London, between 47–61% of the genomes of N. tomentosiformis (ac. NIC 479/84), N. sylvestris (ac. ITB626), N. tabacum (ac. SR1) and the synthetic N. tabacum TR1-A (details of the sequencing output can be found in Table S1; sequence reads were submitted to the NCBI SRA under the study accession number: SRA045794). We choose the N. tomentosiformis accession NIC 479/84 because it most closely resembles the T-genome of N. tabacum [29,30]. There is no N. sylvestris accession that is considered to be more closely related to the N. tabacum S-genome than any other [30].

Clustering, contig assembly and sequence analysis

A graph-based clustering approach described in was used to identify and reconstruct, in silico, the major repeat types present in the genomes of N. tabacum, N. sylvestris and N. tomentosiformis as described in Renny-Byfield et al. [15]. A combined dataset of 454 sequence reads from all three species was used to generate clusters and contigs representing repetitive DNA sequences. Mutual similarities can then be visualised in graph form (Fig. 1 a and Fig 2 b) in which nodes correspond to sequence reads, and a Fruchterman-Reingold algorithm is used to position nodes. Reads that are most similar are placed closest together whilst those that are less closely related are more distant (described in detail in Novak et al. [31]. Contig assembly is performed with reads from each cluster and the contigs are named according to the number of the cluster from which they derive (X) and Nt designates Nicotiana, i.e. Nic CL3. Each cluster typically generates multiple contigs, each of which is designated a number (Y), giving a format NtCL3 contigY. All contigs assembled in this work are available via our websites: http://webspace.qmul.ac.uk/sbyfield/Simon_Renny-Byfield/Data.html and http://webspace.qmul.ac.uk/arleitch/Site/Home.html.

We estimated the genomic abundance of all contigs in each cluster using the “map reads to reference” function of CLC Genomics workbench version 4, requiring 80% sequence similarity over 50% of the sequencing read (any given sequence from the Illumina or Roche 454 datasets was mapped only once). The depth at which reads are mapped reflects the genomic proportion (GP) of the corresponding repeat and so provides a measure of its abundance within the genome. To obtain the GP of a given cluster, all GPs for contigs within that cluster were summed. For Roche 454 and Illumina datasets the average read-depth along each contig (RD), genome representation (GR, calculated as RD × contig length) and genome proportion (GP, calculated as (GR/total size of the dataset in base pairs) × 100) were calculated for each species independently. For the synthetic N. tabacum line TR1-A only Illumina sequence reads were used to calculate GP.

Clusters were then subjected to sequence similarity searches against RepBase [32] in order to identify, where possible, the repeat type from which they derive.

Analysis of NicCL3 using paired-end reads

We analysed paired-end Illumina data from N. tomentosiformis and N. tabacum to assess the occurrence of sequences where one of the paired reads hits NicCL3 and the partner read did not. Reads were screened for quality and both reads of the pair were removed from the dataset if one or other of the reads failed the following quality checks: the read was at least 95 bp long and with no more than five unidentified nucleotides (Ns). All reads passing the quality checks were then trimmed to 95 bp in length. Illumina reads were subjected to similarity searches (requiring 90% sequence similarity along 53% of the sequence read) against contig 8 of NicCL3. More stringent settings were used in this instance to compensate for the shorter read length of Illumina reads. The proportion of pairs where both reads hit (termed a dual High-scoring Segment Pair (HSP)) was recorded. Subsequently unmatched sequences from a pair, where only one read matches contig 8 of NicCL3 (termed solo HSPs), were subjected to further sequence similarity searches to all other clusters. Those that hit other contigs in the NicCL3 cluster were reassigned as dual HSPs. The distribution of solo HSPs was then plotted as a proportion of total HSPs along the length of the monomer of contig 8 of NicCL3.

Sequence similarity in NicCL3

We compared sequence similarity of NicCL3 derived sequences in N. tomentosiformis and N. tabacum using 454 reads described in Renny-Byfield et al. [15]. Reads deriving from N. tomentosiformis and N. tabacum were analysed by BLASTn analysis using the stand alone BLAST program [33] with default parameters with the exception of the following: -e 1e-5, -v 80,000, -b 80,000, -F F. Reads from each species were analysed separately in a pair-wise fashion. Custom BioPerl scripts were used to extract the sequence similarity of all hits to a given read (excluding the query sequence hitting itself). In addition we analysed a mix of all the NicCL3 derived reads from both of the progenitor species. Pair-wise similarity scores for NicCL3 sequences from N. tabacum and N. tomentosiformis and the mix of the progenitor species were plotted as frequency distributions and density estimates using the R statistical package [34].

We used BLASTn to analyse the proportion of NicCL3 reads that matched the consensus sequence (NicCL3, contig 8) at any given nucleotide for N. tomentosiformis and N. tabacum using custom BioPerl scripts.

PCR

DNA was amplified from 50 ng of N. tomentosiformis (ac. NIC 479/84) genomic DNA using Bioline Taq DNA polymerase (San Francisco, USA) supplemented with 1× Bioline NH4 Buffer,
Figure 1. Structure and copy number of NicCL3. (a) Graphical 2D projection of a three dimensional network where each node represents a single 454 sequence within NicCL3. Nodes are placed according to sequence similarity, where similar sequences are placed close together, and more distantly related sequences further away. Sequence similarity is indicated by edges (connecting lines). Red nodes represent sequence reads. 
1.5 mM MgCl₂, 0.2 mM of each dNTP and 0.2 μM of each primer pair.

(i) Primer pair 1 (forward: 5′-GGTAGAGTAGTGATGAGG-3′ reverse: 5′-TGGTGAGATTGGG-3′, Fig. 1 b, filled arrows). PCR primers were designed from NicCL3 contig 8. PCR involved an initial denaturation step of 3 min at 94°C, followed by 36 cycles of 94°C for 40 s, 48°C for 40 s and 72°C for 45 s, followed by a final extension step of 72°C for 3 min.

(ii) PCR analysis with primer pair 2 (5′-TAAAACTCCCAAATCCG-3′ and reverse 5′-TGGGTATAGTGAGACGA-3′, Fig. 1 b, open arrows). PCR primers were designed against a second region of NicCL3 contig 8. PCR used an initial denaturation of 3 min at 94°C, followed by 36 cycles of 94°C for 50 s, 48°C for 1 min and 72°C for 3 min, followed by a final extension of 72°C for 7 min.

(iii) Primer pair 3 (forward 5′-TGTGAGTGAGTGACTTGTTTTT and reverse 5′-CTTGCTGCTCTCTTGACT-3′). PCR primers were designed against NicCL7 contig 7. PCR followed that described for primer pair 1.

Cloning and sequencing

PCR products of NicCL3 and NicCL7 were cleaned using the Qiagen PCR purification kit and cloned using a TOPO H cloning kit with the pCR H 2.1 vector by Invitrogen according to the manufacturer’s instructions. Positive clones were sequenced using T7 forward and M13 reverse primers at Eurofins MWG|operon. The clones sharing highest similarity with the appropriate contig were selected and used to produce probes for fluorescent in situ hybridisation (FISH) and Southern blot hybridisation.

Probes for FISH

(1) Probes were prepared from a clone (number 9; NCBI accession JQ899200) of NicCL3 (using primer pair 1) and from a clone (number 1; NCBI accession JQ899201) of NicCL7 (using primer pair 3). PCR amplification used the conditions described above for primer pair 1.

Figure 2. The cluster NicCL7/30. (a) Cluster NicCL7/30 shown as a graph. Individual sequence reads are represented as nodes on the graph and for simplicity edges representing similarity hits are not shown. The position of nodes was calculated using the Fruchterman-Reingold algorithm. (b) The same graph but with sequences highlighted depending on the progenitor species from which they derive. (c) Another representation of NicCL7/30 but indicating sequence similarity to conserved coding domains (CCD) including protease (PROT), reverse transcriptase (RT), RNaseH (RH), integrase (INT), chromovirus chromo-domain (CHDII) and gag-pol (GAG). (d) Estimated copy-numbers from 454 read-depth analysis, along the length of the most abundant contig in the merged cluster NicCL7/30. A region between ~500 and 3200 bp is more abundant than the remaining contig and likely represents the LTR region of this retroelement, where higher abundance may be due to the presence of solo-LTRs. The position of PCR primers used to make probes for this sequence are indicated with arrows.

doi:10.1371/journal.pone.0036963.g002
with the addition of 0.5 mM digoxigenin-11-dUTP or biotin-16-dUTP labelled nucleotides. All probes were cleaned using Qiagen PCR Purification Kit according to the manufacturers instructions. (2) An 18S nuclear ribosomal DNA (rDNA) sequence cloned from Allium cernum [35] was used to generate a probe as detailed above, with the exception that the extension step of the PCR was at 72 °C for 2 min and the final extension was for 7 min. The following primers were used, 18S2F 5'-CGGAGATTAGGTTCCGATTCC-3' and AB101R 5'-ACGAAATTATGCTACCCGGT-3', the latter modified from Sun et al. [36]. (3) Total genomic DNA for genomic in situ hybridisation (GISH) from N. tomentosiformis (as NIC 479/84) and N. sylvestris was labelled with biotin-16-dUTP and digoxigenin-11-dUTP, respectively, by using the Roche Nick Translation Kit according to the manufacturers instructions.

**Fluorescence in situ hybridisation (FISH)**

Metaphases were accumulated in freshly harvested root-tips by pre-treatment in saturated Gammexane® (hexachlorocyclohexane, Sigma) in water for 4 h. Root-tips were fixed for 24 h in 3:1 absolute ethanol:glacial acetic acid and stored at −20 °C in 100% ethanol. Root-tip metaphases were spread onto glass slides after enzyme digestion as described in Lim et al. [37], and checked for quality using phase contrast microscopy.

FISH followed the protocol described in Lim et al. [38]. Briefly, probe DNA (delivering 50 ng of cloned probe or 100 ng of genomic probe per slide) was added to the probe hybridisation mix (50% (v/v) formamide, 10% (w/v) dextran sulphate, 0.1% (w/v) sodium dodecyl sulphate in 2x SSC (0.3 M NaCl, 0.03 M sodium citrate, pH 7.0)). About 50 μl of the probe mixture was added to each chromosomal preparation and the material denatured with a dyad slide heating-block at 70 °C for 2 min. After overnight hybridisation at 37 °C, slides were washed in 20–25% (v/v) formamide in 0.1x SSC at 42 °C at an estimated hybridisation stringency of 85–89%. Sites of probe hybridisation were detected with 20 μg ml⁻¹ fluorescein conjugated anti-digoxigenin IgG (Roche Biochemical Ltd.) and 5 μg ml⁻¹ Cy3 conjugated streptavidin (Amersham Biosciences). Chromosomes were counterstained using Vectashield with DAPI (4',6-diamidino-2-phenylindole; Vector Laboratories). Material was photographed using a Hamamatsu Orca ER camera and a Leica DMRA2 epifluorescent microscope. Images were processed with Improvision Openlab software and Adobe Photoshop CS2, adjusting for colour balance, contrast, and brightness uniformly.

For multiple probe labelling, preparations were stripped of probe and signal by a 10 min wash at 110% stringency (60% (v/v) formamide, 0.1x SSC at 42 °C). Slides were checked to ensure no signal could be visualised. Slides were then subjected to a second round of FISH using alternative probes and re-photographed.

**Southern blot hybridisation**

DNA was extracted from fresh young leaves according to Kovarik et al. [39], digested with restriction endonucleases (5 U μg⁻¹ DNA, twice for 6 h), fractionated by gel electrophoresis and transferred to GE-Healthcare Hybond XL membranes using alkaline capillary transfer. Membranes were hybridized with 32P-labelled DNA probe (DecaLabel DNA Labeling Kit, MBI Fermentas). Southern blot hybridisation was carried out in a 0.25 M sodium phosphate buffer (pH 7.0) supplemented with 7% (w/v) sodium dodecyl sulphate (SDS) at 65 °C [40]. Membranes were washed with 2x SSC, 0.1% SDS (twice for 5 min) and then with 0.2x SSC and 0.1% SDS (twice for 15 min at 65 °C). The membranes were exposed to a Storage Phosphor Screen, scanned (Storm, GE-Healthcare) and the signal was quantified using Image Quant (GE-Healthcare). The DNA probe was a ~500 bp insert of clone 9 of NiCL3 used in the FISH experiments.

All materials and data are available on request.

**Results**

Clustering, contig assembly and repeat abundance estimates

A combined dataset of 454 reads from the three *Nicotiana* species totalling >70 Mb of DNA was subjected to a clustering based repeat identification procedure as described in the Materials and Methods section, and in detail in Novak et al. [31]. Briefly sequence reads are subjected to pair-wise sequence similarity analysis where related sequences are gathered into clusters. These clusters correspond to families of repetitive DNA sequences and the reads therein are further assembled into contigs. The depth at which Roche 454 or Illumina reads map to these sequences allows estimation of genomic proportion (GP) of the corresponding repeat. Moreover *N. tabacum* is a symmetrical hybrid since both ancestors have roughly the same genome size (~2,650 Mbp/1C [41]). Therefore for a uniparentally inherited repeat, the expected genome proportion (GP) in *N. tabacum* is 0.5 of the parental GP.

Read-depth analysis revealed two clusters (NiCL3 and NiCL7) to be highly abundant in the genome of *N. tomentosiformis*. Illumina sequencing read-depth across NiCL3 indicates a genome proportion (GP) of 1.60%, while similar analysis with 454 data indicate a GP of 1.91%. The corresponding values in *N. tabacum* are 0.10 and 0.09% respectively, both markedly lower than the abundance (0.80%/0.95%) that would be predicted given additivity of the parents (Table 1).

A graphical representation of sequence relationships in the cluster containing NiCL3 is shown in Figure 1 A. Reads form a circle-like pattern indicative of direct terminal or tandem repeats. NiCL3 is a tandem repeat (see below). With this graphical analysis, tandem repeats often have a globular shape in 3D-networks, particularly if the monomer size is small. The reason that NiCL3 does not have this pattern is due to its length (2.2 kb). Not all of the ~360 bp reads that make up the graph share sequence similarity (i.e. reads in different regions of the monomer will not overlap, as with a short monomer). The read connections (edges) are largely ‘linear’ until reaching either end of the monomer, where reads can bridge adjacent monomers, forcing the ends of the network to close up in a wheel like pattern. Copy number estimates along the most abundant contig of NiCL3 is shown in Figure 1 B. A MGBLAST search was conducted using the consensus NicCL3 monomer as a query to *N. tabacum* genome survey sequences (GSSs) (e-value < 1e-15). This produced 741 hits along the whole length of the NiCL3 monomer, with 381 hits showing 95% to 100% similarity, supporting the restriction digest, sequencing and clustering/assembly data.

Sequence similarity searchers of NiCL3 to RepBase returned a small region (positions 717–953 with 40% amino acid identity) with similarity to GYPSODEI1.1 a Ty3/gypsy-like retroelement identified in *Solanum demissum* [32] while searches against the Pfam conserved protein domain database returned no matches. Regions with similarity to *Nicotiana tomentosiformis* endogenous pararetrovirus (NtOEPRV) insertion sites [42] were identified and indicated in black (Fig. 1 b).

The NiCL7 cluster is closely related to cluster 30 (NiCL30) and they are likely derived from the same repeat family. Therefore, they were merged in to a single cluster, hereafter called NiCL7/
30, shown graphically to be circle-like (Fig. 2 a–c). Protein BLAST searches indicate that reads within NicCL3/30 have sequence similarity to reverse transcriptase (RT), integrase (INT), RNaseH (RH), protease (PROT) and GAG domains of LTR retroelements, as well as a chromovirus specific chromatin-remodeling domain (CHDII). We therefore suggest NicCL3/30 is likely to be a chromoviruses-like (Ty3/gypsy retroelements, 70% amino acid similarity to reverse transcriptase (RT), integrase (INT), RNaseH (RH), protease (PROT) and GAG domains of LTR retroelements, as well as a chromovirus specific chromatin-remodeling domain (CHDII). We therefore suggest NicCL3/30 is likely to be a chromoviruses-like (Ty3/gypsy retroelements, 70% amino acid identity along 297 bp) family of repetitive DNA, although the repetitive sequence is not formally classified. Reference sequences for NicCL3 and NicCL7/30 are available at the following websites: http://webspace.qmul.ac.uk/sbyfield/Simon_Renny-Byfield/data.html and http://webspace.qmul.ac.uk/arleitch/Site/Home.html.

We analysed Illumina paired-end data to assess the proportion of paired sequences where one read hits NicCL3 and the other member of the pair did not (solo HSPs). In N. tomentosiformis and N. tabacum, 3.16% and 3.80% of paired reads had only one match (solo HSPs) to the NicCL3 respectively. In N. tomentosiformis we observed 95 instances where one sequence of a pair matched NicCL3 while the other matched NicCL7/30. In N. tabacum comparisons of the distribution of solo HSPs along the length of NicCL3 revealed regions of the sequence with high proportions of solo HSPs (Fig. 1 c), a similar pattern was observed in N. tomentosiformis, although it was less apparent. It is noteworthy that the irregular profile of copy number estimates along NicCL3 corresponds closely with the distribution of solo HSPs (compare Fig. 1 b with 1 c).

Cloning regions of NicCL3 and NicCL7

PCR using primer pair 1 (thick black arrows in Fig. 1 b) against the consensus of NicCL3 amplified the region between position 109 and 581 bp. Cloning of the PCR product resulted in four sequences sharing between 92–96% identity with the in silico consensus. PCR using primer pair 3 against the region between 1488 and 1926 bp of NicCL7 produced a band of the expected size. The PCR products were cloned and five clones chosen for sequencing, each had sequence similarity varying between 92 and 96% against the in silico consensus. Clone 9 for NicCL3 and clone 1 for NicCL7 were chosen for further analysis.

FISH

FISH using the NicCL3 clone 9 to metaphase spreads of N. tomentosiformis (ac. NIC 479/84 and Nee et al. 51771) reveals loci on eight of the large sub-metacentric chromosomes (Fig. 3 a, c and Table 2). The signal is highly localized and is exclusive to the distal region of the long arm of four chromosome pairs. The 18S rDNA-bearing chromosome (chromosome 3, following the nomenclature of Lim et al. [43]) lacks any detectable signal. In contrast there is NicCL3 signal at an interstitial locus on the orthologous 18S rDNA-bearing chromosome of the diploid relative N. kawakami (Fig. 3 i). NicCL3 signal is also observed on chromosome T3 of N. tabacum, although it is restricted to the most distal regions of the long arm (boxed in Fig. 3 c). All NicCL3 loci in N. tabacum are noticeably smaller than those in the progenitor N. tomentosiformis and the diploid N. kawakami.

Metaphase chromosomes of several synthetic N. tabacum lines (Th37-3, -7 and -14) reveal only two NicCL3 signals, on a single pair of large submetacentric chromosomes (Fig. 3 b, d, Table 2). The loss of signal is not caused by the absence of N. tomentosiformis-derived chromosomes as GISH to metaphase spreads of Th37-3 reveal a full complement of N. tomentosiformis chromosomes (24 red chromosomes in Fig. 3 f, g). The S0 generation synthetic N. tabacum TR1-A has eight NicCL3 signals as expected (Fig. 3 j).

NicCL7 has a dispersed signal on all N. tomentosiformis chromosomes (Fig. 3 a), although some regions bind the probe more efficiently producing a band-like pattern on large submetacentric chromosome pairs, particularly evident on the 18S rDNA-bearing chromosome. NicCL7 signal is associated with all NicCL3 signals in N. tomentosiformis, Th37 and TR1-A. Th37-3 has NicCL7 signal on 24 of the 48 chromosomes (Fig. 3 j); it is likely these derive from N. tomentosiformis. We were unable to detect any signals of NicCL3 and NicCL7 in N. sylvestris (data not shown).

Southern blot hybridisation

Southern blot hybridisation was carried out using NicCL3 as a probe. For each species 1–2 µg of genomic DNA was digested with BamHI and SpeI enzymes (Fig. 4, Table 2), which have a single restriction site within NicCL3. A ladder pattern of bands was evident in N. kawakami, N. tomentosiformis (TW142 and NIC 479/84), natural N. tabacum (095-55 and SR1), synthetic N. tabacum, Th37-3, -5, -6, -7 and -8. The bands are indicative of tandemly arranged satellite repeats arranged head to tail. The fastest migrating band corresponded to the satellite monomer (2.2 kb), contained within the 2.9 kb in silico reconstruction (Fig. 1 b). There was no signal detected in Th37-1, N. sylvestris N. glutinosa or N. otophora. Other species (N. setchellii and N. tomentosa) have trace amounts of background signal but lack any detectable ladder pattern (Table 2).

In natural N. tabacum, Th37 and N. tomentosiformis digestion of the unit is inhibited when the methylation sensitive restriction enzyme Hhal is used (with one restriction site in the monomer), indicating cytosine methylation of NicCL3 at the restriction site in these species (Fig. 4).

The in silico consensus of NicCL3 sequence includes terminal repeats (Fig. 1 B) and to confirm that these arise because the consensus includes a whole monomer and part of a second monomer in the tandem array, we designed PCR primer pair 2 (open arrows in Fig. 1 b). PCR analysis generated a product of ~1400 bp, consistent with a monomer length of 2.2 kb (data not shown). Sanger sequencing of a clone of this PCR product

### Table 1. Estimated abundance of two families of repetitive DNA sequences in the genomes of N. tabacum and progenitor species N. sylvestris and N. tomentosiformis.

| cluster name | most abundant contig (length in bp) | N. tomentosiformis | N. sylvestris | parental additivity | N. tabacum | TR1-A (50 synthetic tobacco) |
|--------------|-------------------------------------|--------------------|---------------|---------------------|------------|-----------------------------|
| NicCL3       | contig 8 (2926)                     | 1.91/1.60          | <0.01/0.01    | 0.95/0.80           | 0.10/0.09  | NA/0.77                     |
| NicCL7/30    | CL7 (contig 74759)                  | 1.40/1.27          | 0.15/0.20     | 0.78/0.74           | 0.52/0.56  | NA/0.71                     |

doi:10.1371/journal.pone.0036963.t001
Figure 3. FISH of NicCL3 and NicCL7/30. Fluorescence in situ hybridisation (FISH) to metaphase chromosomes of (a, c) N. tomentosiformis (ac. NIC 479/84); (b, d) Th37-3; (e, h) N. tabacum (ac. 095-55); (i) N. kawakamii and; (j) TR1-A. The probes used were 18S rDNA (blue; a-e and h-i only), NtCL7 (green) and NicCL3 probes (red) counter stained with DAPI (grey). Inset (e) shows enlarged chromosome T3 with NicCL3 signal at the distal end of the long arm (arrow heads). (f, g) Genomic in situ hybridisation (GISH) to chromosomes of Th37-3, showing the N. tomentosiformis sub-genome (red) and N. sylvestris sub-genome (green). (i) Note that chromosome 3 of N. kawakamii (18S rDNA bearing) has a large NicCL3 signal proximal to the centromere (arrows). (j) TR1-A an S0 synthetic N. tabacum with the expected number of NicCL3 (red) signals and highly localised NtCL7 (green) signals. Scale bar is 5 μm. doi:10.1371/journal.pone.0036963.g003
Sequence similarity in NicCL3

In order to detect evidence for rounds of amplification and/or homogenisation of NicCL3, we compared sequence similarity of NicCL3 derived 454 reads in N. tomentosiformis, N. tabacum and N. sylvestris. Reads deriving from N. tomentosiformis and N. tabacum were analysed separately. In addition, we analysed a dataset consisting of reads from N. sylvestris and N. tomentosiformis (representing parental additivity). However because there were so few reads from N. sylvestris the output was nearly identical to that from N. tomentosiformis alone (data not shown). Pair-wise similarity scores for NicCL3 sequences from N. tabacum and N. tomentosiformis were plotted as frequency distributions and kernel density estimates (Fig. 5). This analysis revealed a peak of identical sequences in both N. tomentosiformis and separately in N. tabacum. In addition a major peak of reads with sequence similarity close to 0.95 is evident in N. tomentosiformis. In N. tabacum six separate peaks are visible and the N. tabacum genome contains proportionally more reads with lower sequence similarity compared with N. tomentosiformis (Fig. S1). A two-sample Wilcoxon test revealed a significant difference (p<0.00001) between mean sequence similarity of NicCL3 derived-sequences in N. tomentosiformis (0.93) and N. tabacum (0.90). We also examined the proportion of sequence reads from N. tabacum or N. tomentosiformis matching the consensus (NicCL3, contig 8) for each nucleotide along its length (Figure S1 A). We plotted the average proportion of bases identical to the consensus over consecutive 20 bp windows (Figure S1 B). The data indicate that a similar proportion of bases match along the length of the consensus in both species, with the exception of a region towards the end of NicCL3, where the reads are more divergent.

Discussion

NicCL3, an abundant repetitive sequence

Data presented here indicate that next generation sequencing, even with low genome coverage, is an effective way to characterise novel repeats and to compare their evolutionary dynamics between related species. We show that one of the most abundant repeats in the N. tomentosiformis genome, NicCL3 (Table 1), is predominantly arranged in tandem (Fig. 1 c, Fig. 4), has a unit length of 2.2 kb (Fig. 4, Table 2) and is localized in N. tomentosiformis, N. kawakamii, several synthetic N. tabacum lines and natural N. tabacum (Fig. 5). However the sequence is not a typical tandem repeat like the Nicotiana satellites belonging to the HRS60 family [44] for the following reasons. (1) Typically tandem repeat monomers in angiosperms are ~180 bp in length [45]. Even the long monomer pSc250 in Secale cereale is only 550 bp [46]. (2) Satellite blocks usually occur in long arrays of similar units. However NicCL3 also includes a substantial component that is dispersed (c. 3% in N. tomentosiformis and 9% in N. tabacum and Fig. 1 c), some of which is associated with NicCL7/30. In N. tomentosiformis, Th37 and N. tabacum NicCL3 digestion is almost entirely inhibited when using a methylation sensitive restriction enzyme (Fig. 4). These findings indicate that NicCL3 loci are likely to be heavily methylated. However we observed reads derived

Table 2. Repetitive DNA family NicCL3 in the genomes of Nicotiana tabacum, N. tomentosiformis and members of the section Tomentosae.

| accession                  | Southern hybridisation | number of FISH signals |
|----------------------------|------------------------|------------------------|
| N. tomentosiformis (TW142) | +                      | n.s                    |
| N. tomentosiformis (NIC 479/83) | + (Fig. 3)            | 8 (Fig. 3 a, c)        |
| N. tomentosiformis (Nee et al. 51771) | n.s                  | 8                      |
| N. tabacum (SR1)           | + (Fig. 3)             | 8                      |
| N. tabacum (095-55)        | +                      | 8 (Fig. 3 h, e)        |
| Th37* 1                    | –                      | n.s                    |
| 3                         | + (Fig. 3)             | n.s                    |
| 6                         | +                      | n.s                    |
| 7                         | +                      | 2                      |
| 8                         | +                      | n.s                    |
| 9                         | n.s                    | –                      |
| 14                        | n.s                    | 2                      |
| TR1-A                     | n.s                    | 8 (Fig. 3 j)           |
| N. sylvestris              | –                      | –                      |
| N. kawakamii              | +                      | 8 (Fig. 3 i)           |
| N. otophora                | –                      | –                      |
| N. setchellii              | trace amounts          | –                      |
| N. glutinosa               | trace amounts          | n.s                    |

*Indicates a ladder like pattern following restriction digestion and Southern blot analysis.

n.s not screened.

*Groups of Th37 plants as described in [24].

doi:10.1371/journal.pone.0036963.t002
Figure 4. Tandem arrangement of \textit{NicCL3}. Southern blot hybridisation of genomic DNA from (TH) synthetic tobacco Th37-5, (TO) \textit{N. tomentosiformis} ac. NIC 479/84 and (TA) \textit{N. tabacum} ac. SRI1 digested with SpeI, BamHI and HaeIII (a methylation sensitive isoschizomer of BamHI) and probed with \textit{NicCL3}. Size indicators on the left are in kb. Digestion with BamHI and SpeI results in a ladder like pattern, typical of tandem repeats. Digestion is inhibited when using HaeIII, indicating extensive CG methylation of tandem units.

\textit{NicCL3} sequence similarity patterns are typical of tandem repeats. Digestion is inhibited when using \textit{Bam}HI and \textit{Hae}III, indicating extensive CG methylation of tandem units. The unusually long tandem sequence (2.2 kb) and a small flanking \textit{NtoEPRV} (endogenous pararetrovirus) insertions [42] results in a ladder like pattern, typical of tandem repeats. Digestion is inhibited when using \textit{Hae}III, indicating extensive CG methylation of tandem units.

Elimination of \textit{Nic}CL3 in synthetic and natural \textit{N. tabacum}

Next generation sequence (using both Illumina and Roche 454) and FISH analysis have revealed the genome of \textit{N. tabacum} to have a much lower abundance of \textit{NicCL3} than expected given its abundance in \textit{N. tomentosiformis}, suggesting large-scale losses (Table 1). A reduction in copy number of \textit{NicCL3}, amounting to thousands of units has also been observed in fourth generation synthetic \textit{N. tabacum} (Th37). Our supposition that \textit{NicCL3} has experienced dramatic loss in natural and synthetic lines is evidenced by:

(1) One of the \textit{N. tomentosiformis} accession analysed here is the closest known diploid relative of the T-genome of \textit{N. tabacum} and the actual paternal progenitor lineage of Th37 (acc NIC 479/84; reference Murad et al. [29]), and this accession has \textit{NicCL3} in high abundance (Table 1 and Fig. 3).

(2) A total of three \textit{N. tomentosiformis} accessions all show similar and strong \textit{NicCL3} hybridisation patterns in either FISH (Nee et al. 51771 and NIC 479/84; Fig. 3 and Table 2) and/or Southern blot hybridisation (TW142 and NIC 479/84; Fig. 4 and Table 2).

(3) \textit{N. kawakamii} and \textit{N. tomentosiformis} are sister taxa in phylogenetic analysis [26] and both have strong \textit{NicCL3} probe binding in FISH and Southern blot analysis (Fig. 3, 4 and Table 2).

Together (1), (2) and (3) indicate \textit{NicCL3} was probably abundant in the common ancestor of \textit{N. kawakamii} and \textit{N. tomentosiformis} as well as the true paternal ancestor of \textit{N. tabacum}. Therefore the discrepancy between the expected GP and observed GP in \textit{N. tabacum}, as well as the loss of \textit{NicCL3} loci in Th37, is likely to be due to sequence reduction in the allopolyploids rather than expansion in the progenitor post allopolyploidy.

We have shown that, in synthetic \textit{N. tabacum} lines Th37-3, -7 and -14 the number of large blocks of \textit{NicCL3} signal is reduced from eight signals to two (Fig. 3 b, d). Several lines of Th37 (3,5,6,7 and 8) show low, but detectable levels of \textit{NicCL3} following Southern blot analysis. It is clear that whole loci carrying many thousands of \textit{NicCL3} units have been lost from synthetic lines. In addition two synthetic \textit{N. tabacum} lines (Th37-1, 9, Table 2) lack any detectable \textit{NicCL3} signal both in Southern blot and FISH analysis indicating that this sequence has been completely (or near
Mechanisms of NicCL3 loss

The loss of NicCL3 in synthetic N. tabacum Th37-1, 3, 8 and 9 cannot be attributed to incomplete chromosomal contribution from N. tomentosiformis as GISH to metaphase spreads show the expected number of N. tomentosiformis-derived chromosomes (Fig. 3 and Skalicka et al. [16]). Repeats arranged in tandem, for example rDNA, are thought to alter their copy number via unequal crossing-over, although the exact mechanisms are still obscure [51,52]. Homoeologous chromosome pairing has been proposed as a mechanism of sequence and chromosome loss [53], and compelling evidence exists for such chromosomal rearrangements in synthetic Brassica hybrids [54,55] and recently formed Tragopogon allopolyploids [56,57]. However Salina et al. [58] suggested that changes in copy number of tandem repeats Spelt1 and Spelt52 in synthetic wheats, were not a consequence of interspecific recombination during meiosis, as lines with or without the Ph1 locus show similar patterns of copy number change (Ph1 mutants have increased frequency of homoeologous pairing). Similarly in Nicotiana there is no evidence for extensive homoeologous pairing [9,27,59], and so an alternative explanation is needed. Striking sequence homologies exist between different chromosomes of the same species: essentially the same repeats form large blocks of heterochromatin on multiple chromosomes of both S and T genomes (this study and Lim et al. [9]). Hence, it is possible that recombination between large blocks at homologous and non-homologous loci carrying NicCL3 may explain its elimination. Indeed the higher proportion of solo HSPs in paired-end data in N. tabacum compared to N. tomentosiformis is consistent with the preferential loss of the tandem repeated component of NicCL3 in the allopolyploid (Fig. 1c).

The outcome of such process would be the generation of chromosomes with either extremely large arrays and/or chromosomes with large deletions of repeats. Indeed if small deletions within the unit were responsible for lowering the genome proportion of NicCL3 in N. tabacum, then one might expect to see a smear towards smaller molecular weight fragments in Southern blot analysis (Fig. 4), however this was not observed. Instead the relatively sharp bands suggest that the removal of whole units within the tandem array is responsible for the reduced abundance of NicCL3 in N. tabacum.

Recombination mechanisms are thought to be responsible for homogenisation of sequences arranged in tandem and there is evidence that this has occurred in NicCL3 (Fig. 5). Both N. tabacum and N. tomentosiformis have a peak in the number of sequences with a nucleotide similarity of one. Nicotiana tabacum has a series of peaks each with progressively less sequence similarity (Fig. 5), perhaps indicative of more ancient rounds of homogenisation. It is possible that these events coincide with NicCL3 unit loss.

We examined the possibility that different regions of NicCL3 may be more variable than others, a pattern that would explain the series of peaks in Fig. 5. However analysis of the sequence similarity of reads against the consensus failed to provide any evidence of such a pattern in either N. tabacum or N. tomentosiformis (Fig. S1). Hence a hypothesis of repeated rounds of sequence homogenisation seems a better explanation for the series of peaks.

Our study is significant in providing evidence of multiple large-scale deletions, occurring repeatedly in both natural and synthetic material. This has resulted in the removal of almost all of the continuous arrays of NicCL3 in N. tabacum. We have hypothesized that the loss is most likely the result of multiple unequal recombination events between tandem components of NicCL3 and the maintenance of dispersed units of NicCL3 suggests they are more stable than those in tandem array.

Supporting Information

Figure S1  Sequence similarity of BLASTn hits to the consensus of NicCL3 (contig 8) calculated by examining the proportion of hits that match the consensus over a given nucleotide. (a) All the data points for each nucleotide in the consensus and (b) the data averaged over consecutive 20 bp windows.

Table S1  Dataset size and average read length for the four Illumina runs used in this analysis following the removal of plastid sequences.

Acknowledgments

We would like to thank Laura Kelly, Richard Buggs, Mark Chase, Bolous Chalhoub, Marie-Angele Grandbastien, Marc Deloger and Stephen C. LeComber.

Author Contributions

Conceived and designed the experiments: ARL SRB RAN AK. Performed the experiments: SRB MC JM PN AK. Analyzed the data: SRB PN JM AK AR L. Contributed reagents/materials/analysis tools: AK JM PN SRB. Wrote the paper: SRB ARL AK JM PN MC RAN.

References

1. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Current Opinion in Plant Biology 8: 135–141.
2. Leitch AR, Leitch IJ (2008) Perspective – Genomic plasticity and the diversity of polyploid plants. Science 320: 481–483.
3. Solheim DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, et al. (2009) Polyploidy and angio...
associated with an inverse correlation between gene copy number and expression. Genetics 176: 2509–2519.

12. Shakhd H, Kashkush K, Ozkan H, Feldman M, Levy AA (2001) Sequence elimination and cytosome methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell 13: 1749–1759.

13. Khasdan V, Yaakov B, Kraitshtein Z, Kashkush K (2010) Developmental timing of DNA elimination following allopolyploidization in wheat. Genetics 185: 307–316.

14. Leitch JI, Hanson L, Lim KY, Kovarik A, Chase MW, et al. (2008) The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). Annals of Botany 101: 805–814.

15. Renny-Byfield S, Chester M, Kovarik A, Le Comber SC, Grandbastien MA, et al. (2011) Next generation sequencing reveals genome downsizing in allopolyploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNA. Molecular Biology and Evolution 28: 2043–2051.

16. Skalicek K, Lim KY, Matyasek R, Matuze M, Leitch AR, et al. (2005) Preferential elimination of repeated DNA sequences from the paternal, Nicotiana tomentosiformis genome donor of a synthetic, allopolyploid tobacco. New Phytologist 166: 291–303.

17. Margulis L, Egelhorn M, Alhme WM, Araya S, Badger JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

18. Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, et al. (2010) Repetitive part of the banana (Musa acuminate) genome investigated by low-throughput 454 sequencing. BMC Plant Biology 10.

19. 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.

20. Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11.

21. Wicker T, Taudien S, Houben A, Keller B, Graner A, et al. (2009) A whole-genome shotgun sequence snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. The Plant Journal 59: 712–722.

22. Macas J, Krejnovsky E, Neumann P, Novak P, Kohlízkova A, et al. (2011) Next generation sequencing-based analysis of repetitive DNA in the model dicotous plant Silene latifolia. PLoS One 6: e27335.

23. Burk LG (1973) Partial self-fertility in theoretical amphiploid progenitor of X. tabuana Journal of Heredity 64: 345–350.

24. Shigeto T, Lim KY, Matyasek R, Koucha R, Leitch AR, et al. (2005) Rapid evolution of parental rDNA in a synthetic tobacco allopolyploid. American Journal of Botany 90: 908–906.

25. Lim KY, Koucha-Skalcika K, Sarasvan V, Clarkson J, Chase MW, et al. (2006) A genetic recombination-based analysis of repetitive DNA in the model dicotous plant Silene latifolia. BMC Genomics 8: 132–145.

26. Wicker T, Taudsen S, Houben A, Keller B, Graner A, et al. (2009) A whole-genome shotgun sequence snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. The Plant Journal 59: 712–722.

27. Macas J, Krejnovsky E, Neumann P, Novak P, Kohlízkova A, et al. (2011) Next generation sequencing-based analysis of repetitive DNA in the model dicotous plant Silene latifolia. PLoS One 6: e27335.

28. Burk LG (1973) Partial self-fertility in theoretical amphiploid progenitor of X. tabuana Journal of Heredity 64: 345–350.

29. Shigeto T, Lim KY, Matyasek R, Koucha R, Leitch AR, et al. (2005) Rapid evolution of parental rDNA in a synthetic tobacco allopolyploid. American Journal of Botany 90: 908–906.

30. Kelly LJ, Leitch AR, Clarkson J, Hunter RB, Knapp S, et al. (2010) Intragenic recombination events and evidence for hybrid speciation in Nicotiana (Solanaceae). Molecular Biology and Evolution 27: 781–799.

31. Goodspeed T (1954) The genus Nicotiana. Chron Bot 16: 1–536.

32. Knapp S, Chase MW, Clarkson J (2004) Nomenclatural changes and a new sectional classification in Nicotiana (Solanaceae). Taxon 53: 73–82.

33. Murad E, Lim KY, Christophel F, Matyasek R, Lichtenstein CP, et al. (2002) The origin of tobacco’s T genome is traced to a particular lineage within Nicotiana tomentosiformis (Solanaceae). American Journal of Botany 89: 921–928.

34. Petri M, Lim KY, Jolicoe J, Beatrice JM, de Borne FD, et al. (2007) Differential impact of retrotransposon populations on the genome of allopolyploid tobacco (Nicotiana tabacum). Molecular Genetics and Genomics 278: 1–15.

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

36. Jurka J, Kupintov VV, Pavlick A, Klonowski P, Kohany O, et al. (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenetics and Genome Research 110: 462–467.

37. Ahsan MF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

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