Evolution of repetitive DNA in angiosperms:

Examples from *Nicotiana* allopolyploids

Simon Renny-Byfield

School of Biological and Chemical Sciences,
Queen Mary University of London,
Mile End Road,
London, E1 4NS

Advisor: Professor A.R. Leitch

A thesis submitted to the
University of London
for the degree of
Doctor of Philosophy
July 2012

Declaration

I certify that this thesis, and the research conducted therein, are the product of my own work, and that all ideas and quotation from other works, published or otherwise, are fully acknowledged in accordance with standard refereeing practices in the biological sciences.

I acknowledge particular data acquisition and analytical assistance as follows:

Steve LeComber helped develop the script to produce Figure 2.1, originally made in Mathematica, which I have reproduced here in R.

Ales Kovarik performed the majority of Southern blot analysis in Chapter Three and Michael Chester assisted me with the remaining Southern blots also in Chapter Three.

Petr Novak and Jiri Macas carried out the clustering in Chapter Two and helped produce Figures 2.3 and 2.5.

Acknowledgements

First and foremost I would like to express my sincere gratitude to my PhD supervisor Andrew R. Leitch. During my time as a PhD student Andrew has been an absolutely outstanding supervisor and mentor. He has succeeded in making my journey through academic research as smooth as possible and without his guidance, patience, expertise and friendship I would not have been able to complete this thesis.

My sincere thanks is extended to Michael Chester, who was an enormous help and expertly instructed me on the basics of molecular biology at the beginning of my studies, despite the fact he was writing his own thesis at the time. Michael's efforts in helping me were gratefully appreciated. Members of my review panel, Richard A. Nichols and Steve C. LeComber, have assisted me greatly, making panel meetings interesting valuable and encouraging, thank you to you both. I would also like to thank my Czech colleagues Jiri Macas and Petr Novak, as well as Mahesh Pancholi who helped me immensely while I was setting up the bioinformatics pipeline here at QMUL. I wish to acknowledge and express my sincere gratitude to Ales Kovàrik, whose expertise in molecular biology is second to none, his comments and guidance have been invaluable. Heike Brinkman has provided excellent support during the period of study and has made life in the lab easy and productive, thank you Heike. All of these colleagues have made this thesis possible, thank you to you all.

I also wish thank Richard Buggs, Ilia Leitch, Laura Kelly, Bolous Chalhoub, Mark W. Chase, Mike Fay, Elizabeth McCarthy, Marie Angele-Grandbastien and Marc Deloger for useful comments and discussion during the course of my research.

Last but not least I would like to thank Sarah Tuffin, Gemma Renny-Byfield, Eileen Byfield, Diane Doyle and my parents Ronald F. Byfield and Antonia Renny-Byfield for their continued support and encouragement.

Abstract

Allopolyploidy, interspecific hybridisation coupled with genome multiplication, is a prevailing force in the evolution of angiosperms. This thesis examines the consequences of allopolyploidy at the genomic level. The genus *Nicotiana* is an ideal model system for such studies as it includes allopolyploids formed over widely different time frames (recent to millions of years). The global genome composition of several diploid and allopolyploid species was analysed using a graph-based clustering approach, grouping next generation sequencing reads into clusters (families) of repetitive DNA. Such analysis enables examination of genome size change and diploidisation processes post-allopolyploidy.

I compared the abundance of >14,000 repeats in the young allopolyploid *N. tabacum* (less than 0.2 million years old) with relatives of the diploid progenitors, *N. tomentosiformis* (paternal genome donor) and *N. sylvestris* (maternal genome donor). Repetitive DNA from the paternal genome tends to be eliminated, whereas DNA from the maternal line remains largely unchanged. A newly described tandem repeat (*Nic*CL3) paternally inherited in *N. tabacum,* is a striking example. Despite a predicted abundance of ~1% *Nic*CL3 now accounts for only 0.1% of the genome in the allopolyploid, a loss repeated in some synthetic lines of *N. tabacum* after only four generations.

Nicotiana section *Repandae* formed from a single hybridisation event between relatives of *N. sylvestris* and *N. obtusifolia* c 5 million years ago. Subsequent

diversification has produced four species where genome size varies by 33%; *N. repanda* showing genome upsizing and *N. nudicaulis* showing genome downsizing compared with the expected genome size. There was evidence for the erosion of low copy-number repetitive DNA in both allopolyploids. However in *N. repanda* genome downsizing has been counteracted by the expansion of a few repeat types. Notably these processes are concurrent with the failure to distinguish progenitor chromosome sets, which I argue is part of the diploidisation process.

Contents

List of Figures

List of Tables

Chapter 1 General Introduction

The importance and occurrence of allopolyploidy in angiosperms

The evolution of angiosperms is inextricably linked with whole genome duplication (WGD) and hybridisation, together known as allopolyploidy (Adams & Wendel, 2005; Leitch & Leitch, 2008; Soltis *et al.*, 2009; Jiao *et al.*, 2011). Phylogenetic analysis of orthologous and paralogous genes across diverse plant groups indicate a WGD at the base of all seed plants with an additional WGD event at the root of the angiosperms (Jiao *et al.*, 2011). Several WGDs have occurred within the flowering plants, including the γ event at the base of the eudicot lineage and several within the monocots (Soltis *et al.*, 2009). In addition to these palaeopolyploid events many more *local* events are evident, including several WGDs in the crucifer lineage (Soltis *et al.*, 2009). Whole genome duplications have been followed by major species radiations, indicating the potential importance of such processes in the diversification of land plants (Soltis *et al.*, 2009). It has been estimated that as many as 15% of speciation events within angiosperms are associated with polyploidy, increasing to 31% in ferns (Wood *et al.*, 2009). However comparisons of species richness in sister genera, separated by shifts in ploidy, provide no evidence that polyploidy *per se* induces an overall increase in diversification (Wood *et al.*, 2009). Mayrose *et al.* (2011) analysed the diversification, speciation and extinction rates of polyploid and diploid plant species and found that newly arisen polyploids diversified at lower rates than diploids. Furthermore WGD events were disproportionately distributed at the tips of phylogenetic tress indicating that neo-polyploids generally fail to establish.

Despite these recent observations the fact that all major groups of land plants have experienced a WGD during their ancestry suggests that polyploidy has played a crucial role in the generation of the major taxonomic groups we see today. However the reason for the ubiquitous nature of WGD in angiosperms remains unclear. Does polyploidy play an important role because of its high frequency, or because polyploids, once established, are more successful at giving rise to major species radiations?

While ancient polyploid events have only recently been identified, it has long been clear that chromosomal polyploidy (hereafter termed polyploidy) in contemporary plants is ongoing. In this case the close link between polyploidy and hybridisation (allopolyploidy) is more apparent. Allopolyploids of recent origin include *Spartina anglica* (Ainouche *et al.*, 2004), *Tragopogon mirus*, *T. miscellus* (Soltis *et al.*, 2004) and *Senecio cambrensis* (Ashton & Abbott, 1992). In addition many of the world's most economically important crop species, including wheat, tobacco and several *Brassica* species, are allopolyploids.

The ubiquitous nature of ancient WGD, numerous examples of modern allopolyploids and the prevalence of polyploidy in crop plants have led researchers to ask several key questions; what role has allopolyploidy played in the generation of genetic diversity, introgressive traits and the origin of major taxonomic groups? Do hybridisation and WGD have distinct impacts on plant genomes? At what tempo do these changes occur, what sequences do these changes affect and are alterations predictable? What are the impacts of allopolyploidy on gene expression, both immediately and over longer time scales? Research with these questions in mind has led to the identification of several key processes affecting allopolyploid genomes.

Allopolyploidy and homoploid hybridisation: gene evolution and expression

It has long been noted that the WGD introduces new genetic material from which evolutionary novelty can develop (Stebbins, 1950; Stephens, 1951; Ohno, 1970; Force *et al.*, 1999; Wendel, 2000; Leitch & Leitch, 2008). An example being immediate redundancy at the gene level, where the fate of duplicate genes can vary. One possibility is non-functionalisation, where selection on one of the duplicate genes is relaxed, resulting in the accumulation of mutations that prevent gene function in one of the duplicates, leading to the generation of psuedogenes (Lynch & Conery, 2000). Another outcome is neo-functionalisation, where a novel function evolves in one of the duplicate genes (Vandenbussche *et al.*, 2003). Alternatively the function of an ancestral gene is divided and partitioned between the two duplicate copies, a process known as sub-functionalisation (Force *et al.*, 1999), or both copies can be maintained by purifying selection, as seen in *Xenopus* (Hughes & Hughes, 1993). Finally, one of the duplicate genes can be lost (Buggs *et al.*, 2010a). Such processes of divergence in gene function are thought to introduce the phenotypic and physiological novelty in polyploid organisms and are cited as one of the key factors in the success of polyploid plants (Comai *et al.*, 2000; Wendel, 2000; Adams & Wendel, 2005; Leitch & Leitch, 2008). Zhang *et al.* (2011) demonstrated duplication, sub-functionalisation and hyperfunctionalisation of the Q homeoalleles in allopolyploid wheat. Moreover diversification of these homeoalleles is a crucial process in the development of the free-threshing character (the easy removal of seeds from their surrounding chaff), indicating that polyploidy has played a crucial role in generating the genetic novelty used in successful domestication.

Freeling (2009) proposed a 'dosage balance' hypothesis to explain retention of duplicated genes through selection for balanced stoichiometry of interacting gene products. Genes in dosage balance may be those whose products are involved in heterodimer formation, or are closely linked in protein networks. Buggs *et al.* (2011) tested this hypothesis by studying young (originating ~80 years ago), independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). The authors were able to detect similar patterns of gene retention in multiple independent populations. They concluded dosage sensitivity was a predictor of duplicate gene retention, supporting Freeling's hypothesis (Buggs *et al.*, 2011). Importantly some genes that were lost were grouped in the same Gene Ontogeny (GO) categories as those lost in other allopolyploids in Asteraceae (Barker *et al.*, 2008), suggesting a tendency towards repeatability.

In addition to the myriad of outcomes at the genic level it has become evident that many allopolyploids exhibit alterations to the transcriptome when compared to the progenitor species. An early indication of this phenomenon was the observation of nucleolar dominance (reviewed in Pikaard 2000), where one set of orthologous rRNA genes (rDNA) are expressed while the alternative set are silenced. It is now apparent that alterations to gene expression following allopolyploidy are more widespread (Adams *et al.*, 2003; Auger *et al.*, 2005; Hegarty *et al.*, 2006; Flagel *et al.*, 2008; Rapp *et al.*, 2009; Buggs *et al.*, 2010b; Chague *et al.*, 2010). For example cDNA-single-stranded conformation polymorphism in natural and synthetic lines of *Gossypium hirsutum* revealed unequal contribution of homeologues to the transcriptome in a tissue specific manner (Adams *et al.*, 2003). This revealed not only that allopolyploidy can induce changes in gene expression, but that these alterations were dependent on the gene, homeologue and tissue in question. These observations are consistent with the hypothesis of rapid sub-functionalisation, previously thought to be a far more protracted process involving accumulation of complimentary mutations (Lynch & Force, 2000). However the scale of the resulting allopolyploid-induced gene expression perturbation and the speed at which it occurs remained obscure, as did the differential effects of hybridisation versus genome doubling.

In *Helianthus* analysis of gene expression in homoploid hybrid *H. deserticola* also revealed that ~12% of over 3000 EST showed altered expression compared to the progenitor species (Lai *et al.*, 2005). The relative impact of hybridisation versus genome doubling was investigated by Hegarty *et al.* (2006). They compared the young allopolyploid *Senecio cambrensis* and an F$_1$ hybrid (*Senecio × baxterii*) with their progenitor species using microarray analysis. Noticeable changes to patterns of gene expression occurred in the F$_1$ hybrid, while expression levels in the natural allopolyploid were more similar to the progenitors. Colchicine duplication of chromosomes in the F_1 hybrid resulted in gene expression levels that more closely resembled the parents, indicating that WGD can dampen the effects of hybridisation, perhaps due to the return of 'gene balance' between homeologues. Thus in *Senecio* it appears hybridisation, rather than genome doubling, is responsible for the majority of alterations to the transcriptome.

Although Hegarty *et al*. (2006) were able to tease apart the contribution of WGD and hybridisation, the sensitivity of the microarray analysis was insufficient to distinguish the relative contribution of parental homeologues. However Buggs *et al*. (2010a) were able to compare large numbers of loci using Roche 454 (454) pyro-sequencing of expressed sequence tags (ESTs). They identified single nucleotide polymorphisms (SNPs) diagnostic for *T. pratensis* and *T. dubius*, progenitors of the allopolyploid *T. miscellus*. Highly redundant Illumina EST data was then used to compare the relative contribution of each parental homeologue to the *Tragopogon miscellus* transcriptome. Buggs and co-workers showed that 69% of genes had roughly equal contribution from each homeologue, 22% were expressed in a bias manner in favour of one of the homeologues and ~8.5% lacked any detectable level of either homeologue. Assaying a sample of silenced homeologues revealed that 85% of the genes lacking expression were actually lost from the genome of the allopolyploid. Therefore Buggs *et al*. (2010a) were able to demonstrate that silencing of these genes was not epigenetic but the result of gene loss.

Several examples of alterations to gene expression associated with allopolyploidy have been documented, however in some instances this occurs to only a minority of genes. For example in two independently synthesised *Arabidopsis* allopolyploids (*A. thaliana* × *A. arenosa*) only 5.2% and 5.6% of transcripts assayed exhibited expression patterns differing from the mid-parent value (Wang *et al.*, 2006). Similarly when comparing expression changes between the progenitor species *Senecio squalidus* and *S. vulgaris* with their progeny *S.* × *baxterii* (homoploid hybrid) and *S. cambrensis* (allopolyploid), only 5-7% of transcripts demonstrated altered expression levels (Hegarty *et al.*, 2005). In wheat only ~7% of transcripts in allopolyploid *Triticum aestivum* differed from parental additivity when compared to a control of a mix of progenitor RNAs (Chague *et al.*, 2010) and in newly synthesised allotetraploid cotton only 5% of transcripts showed altered expression patterns (Adams *et al.*, 2004).

Assessing gene expression patterns by comparing allopolyploids to a mid-parent value can result in crucial alterations in expression profiles being overlooked. Rapp *et al.* (2009) invoked a phenomenon of "genome dominance" in allopolyploid *Gossypium*: For most genes the contribution of each homeologue was biased in favour of one of the progenitor species, although only 1-4% of genes showed expression levels significantly different from the mid-parent value.

Repetitive DNA and plant genomes

Angiosperm genomes are characterised by incredible range in genome size, ranging from 63 Mbp/1C in *Genlisea margaretae* (Greilhuber *et al.*, 2006) to 148,852 Mbp/1C in *Paris japonica* (Pellicer *et al.*, 2010), a 2363-fold variation. Much of this can be attributed to increases in ploidy level and repetitive DNA content. However even some of the smallest plant genomes have significant proportions of repetitive DNA. *Arabidopsis thaliana*, with a genome size of 157 Mbp/1C (Bennett *et al.*, 2003), harbours transposons totalling 14% of the genome (Kaul *et al.*, 2000). In rice (489 Mbp/1C; Bennett and Smith 1976) the proportion of repetitive DNA is 38-55% (Goff *et al.*, 2002; Barabaschi *et al.*, 2012), while in wheat it has been reported as high as 70-80% (Bennetzen *et al.*, 1998; SanMiguel *et al.*, 2002).

Much of the repetitive DNA in plant genomes is comprised of various transposable elements (TEs), which are grouped into two broad categories, class I (retroelements) and class II (DNA transposons), based on their mode of transposition. Class I TEs include the long terminal repeat (LTR) retroelements which are characterised by a 'copy and paste' mode of transposition. The action of retroelement encoded reverse transcriptase (RT) converts retroelement-derived RNA to DNA for insertion into host DNA (Kumar & Bennetzen, 1999). Ty3/*Gypsy* and Ty1/*Copia* elements are common in plants and often constitute the majority of LTR retroelements found within the genome (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker & Keller, 2007; Wicker *et al.*, 2009; Hribova *et al.*, 2010; Macas *et al.*, 2011). Retroelements that lack LTR regions include long interspersed nuclear elements (LINEs), which encode RT and envelope protein

(ENV) but are often 5′ truncated (Kumar & Bennetzen, 1999). Short interspersed nuclear elements (SINEs) are the simplest of all plant retroelements, lack any detectable protein coding regions or LTRs, are non-autonomous and transposition occurs via the action of enzymes provided by other retroelements.

Class II DNA transposons are also found in plants, transpose via a cut and paste mechanism and are generally less abundant than Class I TEs in plant genomes (Wicker *et al.*, 2007).

Transposable elements are major drivers of genome divergence, particularly in plants, where gene capture, illegitimate recombination and insertions into regulatory sequences can pave the way for significant genome restructuring and gene evolution (Kazazian, 2004). High levels of TE proliferation have the potential to be deleterious. To counteract this retroelement activity is controlled by epigenetic modification and RNA directed mechanisms (Steimer *et al.*, 2000; Cantu *et al.*, 2010). Retroelement transcription is generally prevented by epigenetic marks induced by siRNA pathways, ameliorating retroelement expansion.

To counter balance expansion of TEs mechanisms exist for the removal of repetitive DNA from the genome, likely to involve ectopic and illegitimate recombination (Kejnovsky *et al.*, 2009). The rate of 'genome turnover' (the reshuffling and replacement of genomic DNA over time) is governed by the rate of repetitive DNA expansion and its subsequent removal by recombination mechanisms. Therefore the shifting balance between these two phenomena can generate genome expansion or contraction. Given control and removal mechanisms exist, it is of importance to understand the properties of genomes that allow such large volumes of repetitive DNA to establish and persist and what governs their expansion and contraction.

Genome divergence following hybridisation and allopolyploidy

It seems clear that WGDs coupled with hybridisation, both in the distant past and more recently, ae important in generating the diversity required for the radiation of major plant lineages (Leitch & Leitch, 2008; Soltis *et al.*, 2009). Indeed McClintok (1984) proposed that allopolyploidy can generate "genomic shock" whereby the unification of genomes long ago diverged results in rapid alterations at the DNA sequence and karyotypic level.

It has been argued that chromosomal and sequence alterations triggered by "genomic shock" help establish fertility in neo-allopolyploids (Feldman *et al.*, 1997; Kovarik *et al.*, 2011). Indeed rapid karyotype alterations can be seen in some synthetic allopolyploids of *Nicotiana*, which exhibit extensive chromosomal re-arrangements after only 4 generations (Lim *et al.*, 2006b). Similarly natural allopolyploids of *Nicotiana* have undergone inter-genomic translocations, homogenisation of tandem arrays as well as losses and gains of repetitive DNA over longer time-frames (Koukalova *et al.*, 1989; Lim *et al.*, 2000a; Matyasek *et al.*, 2003; Skalicka *et al.*, 2003; Kovarik *et al.*, 2004; Clarkson *et al.*, 2005; Lim *et al.*, 2007; Kovarik *et al.*, 2008; Koukalova *et al.*, 2010; Kovarik *et al.*, 2011).

In synthetic *Brassica* allopolylpoids restriction fragment length polymorphisms (RFLPs) revealed both losses and gains of fragments after only five generations

(Song *et al.*, 1995). The timing of these changes is in good agreement with chromosomal data (Lim *et al.*, 2006b; Gaeta *et al.*, 2007) and together these indicate genetic alterations following allopolyploidy can be rapid. In wheat amplified fragment length polymorphisms (AFLP), assayed across synthetic hybrids as well as derived allopolyploids, revealed repeatable loss of markers (Shaked *et al.*, 2001). Furthermore losses occurred in F_1 hybrids and after chromosome doubling, revealing the timing of these changes can be variable.

Particular classes of sequences can alter in response to allopolyploidy (Petit *et al.*, 2007; Parisod *et al.*, 2010; Petit *et al.*, 2010). For example Petit *et al.* (2007) demonstrated using sequence specific amplified polymorphisms (SSAPs) that Tnt1 and Tnt2 Ty1/*Copia*-like retroelements were non-additive of the progenitors in allopolyploid *N. tabacum*. In this case non-additivity included losses and gains of bands, corresponding to retroelement deletions and transpositions respectively. Losses and gains were also observed in fourth generation synthetic *N. tabacum*, indicating the same sequences were undergoing deletion and transposition in both natural and synthetic allopolyploids (Petit *et al.*, 2010). Such observations suggest some form of targeted and repeatable sequence elimination, perhaps influenced by the epigenetic state of the sequence in question and/or the interaction of small RNAs (Cantu *et al.*, 2010). Turnover of retroelements and other repetitive DNA sequences is likely the reason genomic *in situ* hybridisation (GISH) fails in allopolyploids of *Nicotiana* formed c. 5 million years ago (Lim *et al.*, 2007).

Although numerous examples of genome perturbation following allopolyploidy exist, other cases seem to indicate little genetic change has occurred (Liu *et al.*, 2001). In *Gossypium* amplified fragment length polymorphism (AFLP) assays of more than 22,000 loci revealed that in newly synthesised allopolyploids additivity of the AFLP bands was observed in almost all cases. Using Southern hybridisation and known retroelements as probes, the authors were able to demonstrate a sample of six newly synthesised allopolyploids showed complete parental additivity, indicating little change at retroelement insertion sites. The neo-allopolyploid *Spartina anglica* also exhibits complete parental additivity when assayed using inter-retrotransposon amplified polymorphism (IRAP) and retrotransposon-microsatellite amplified polymorphism (REMAP) analyses (Baumel *et al.*, 2002). In contrast to previous examples, it seems much of the genome in these allopolyploids is stable.

However genome size (GS) data across the angiosperms indicate a general trend for lower GS in polyploids compared to the sum of the progenitor diploids (Leitch & Bennett, 2004). Indeed analysis of the Kew C-value database indicates the distribution of GS in angiosperms is heavily skewed towards small genomes (mode = 6 Mbp/1C, mean = 5900 Mbp/1C; Leitch and Leitch 2012). The volume of 'missing' DNA is often substantial and is likely caused by a reduction in the repetitive DNA component in polyploid genomes. Modelling studies of retroelement copy-number expansion and contraction in *Gossypium* suggests, in diploids with small genomes, increased rates of removal of *Gorge3* retroelements counter balanced their expansion (Hawkins *et al.*, 2009). Recent studies in *Arabidopsis* indicate that GS reduction can be achieved by the accumulation of hundreds of thousands of small deletions (Hu *et al.*, 2011). These observations indicate potential mechanisms for GS reduction in polyploids. Indeed similar studies of Tnt1 and Tnt2 retroelements in *N. tabacum* provided evidence for loss of insertion sites in natural and synthetic allopolyploids (Petit *et al.*, 2007; Petit *et al.*, 2010). In the long-term it is likely such deletions accumulate to equate to substantial DNA loss.

Only recently have researchers been able to tease apart the impact of hybridisation compared to WGD on the genome organisation of hybrid species. By using genetic linkage analysis in interspecific hybrids of *Helianthus* (homoploid hybrids), Lai *et al.*, (2005) demonstrated extensive chromosome re-arrangements despite a lack of WGD. Further analysis revealed reshuffling of the homoploid hybrid genome can continue for ~100 generations, indicating the time frame over which these changes can occur (Buerkle and Rieseberg, 2008). These findings suggest, at least in some cases, hybridisation is a driver of genome re-arrangements.

However flow cytometry analysis of DNA content in first and sixth generation synthetic *Helianthus* homoploid hybrids indicates that GS is additive of the progenitors, suggesting that amplification/elimination of TEs and other repetitive DNA families does not occur immediately after hybridisation (Baack *et al.*, 2005). Furthermore the GS of wild homoploid hybrids deviated from progenitor additivity by as much as 50%, largely the results of TE amplification (Ungerer *et al.*, 2006) indicating that changes in GS may not be a direct result of hybridisation. From work in *Helianthus* it is clear that similar phenomena affecting nuclear DNA are observed in both homoploid hybrids and allopolyploids. More data are needed to elucidate the patterns and processes, specific and universal that affect both modes of hybrid speciation.

Nicotiana as a model of allopolyploid evolution

The genus *Nicotiana* (*Solanaceae*) is endemic to South America, Australia and Africa and comprises ~76 species (Knapp *et al.*, 2004). Evolution in the genus is characterised by frequent hybridisation, which has been proposed as a major driving force within the group (Goodspeed, 1954). Several allopolyploid events have given rise to polyploid sections, and as a result almost half of the 76 described species of *Nicotiana* are allotetraploids (Knapp *et al.*, 2004). Polyploid sections include *Nicotiana*, *Undulatae*, *Repandae*, *Polydicliae*, *Rusticae* and a section native to Australia, the *Suaveolentes.*

For many of these sections the closest living relatives of the progenitor species are known allowing comparative evolutionary studies (Knapp *et al.*, 2004). Interestingly polyploid events in *Nicotiana* have occurred over large time scales, from the young allopolyploid *N. tabacum* (<200,000 mya) to members of the section *Suaveolentes* (~10 mya; Leitch *et al.*, 2008). Thus *Nicotiana* provides a useful means of comparing allopolyploid evolution over varying time scales within the framework of a single genus. For this reason members of *Nicotiana* have become models for the study of allopolyploidy in angiosperms, with particular focus on the young allotetraploid *N. tabacum* ($2n = 4x = 48$).

GISH and molecular studies have revealed the progenitors of *N. tabacum* to be close relatives of extant *N. sylvestris*, the maternal progenitor, the S-genome donor, and *N. tomentosiformis*, the paternal progenitor, the T-genome donor (Chase *et al.*, 2003; Clarkson *et al.*, 2010). This knowledge has allowed comparisons between the repetitive DNA content of the progenitor genomes with the allopolyploid (Koukalova *et al.*, 1989; Kenton *et al.*, 1993; Matyasek *et* *al.*, 1997; Volkov *et al.*, 1999; Lim *et al.*, 2000a; Fulnecek *et al.*, 2002; Murad *et al.*, 2002; Skalicka *et al.*, 2003; Clarkson *et al.*, 2004; Melayah *et al.*, 2004; Murad *et al.*, 2004; Lim *et al.*, 2007; Petit *et al.*, 2007; Petit *et al.*, 2010; Renny-Byfield *et al.*, 2011; Renny-Byfield *et al.*, 2012). A shared integration of gemini-virus related DNA (GRD) specific to a single line of *N. tomentosiformis* and the T-genome of *N. tabacum*, allowed the ancestry of the allopolyploid to be traced to a particular lineage of *N. tomentosiformis* (Murad *et al.*, 2002). Thus we can be reasonably confident we are analysing the extant lineage closest to the true paternal ancestor of *N. tabacum*.

Previous research has focused on the evolution of repetitive DNA in *N. tabacum*, where rDNA sequences are well characterised and probably the best understood (Volkov *et al.*, 1999; Lim *et al.*, 2000a; Fulnecek *et al.*, 2002; Kovarik *et al.*, 2004). By comparing Southern blot hybridisation patterns of rDNA units in *N. tabacum* and its diploid progenitors Volkov *et al.* (1999) identified a novel rDNA sequence in the allopolyploid, similar to the paternally derived unit. Moreover this unit has replaced almost all other rDNA units, an observation congruent with rapid homogenization of these sequences (Lim *et al.*, 2000a). In addition to rDNA the evolution of a number of tandem repeat families has been described in detail (Koukalova *et al.*, 1989; Matyasek *et al.*, 1997; Lim *et al.*, 2004b; Koukalova *et al.*, 2010). These include the elimination of intergenic spacer (IGS) like satellite repeats (A1/A2) and NTRS satellite repeats in *N. tabacum*. Although the evolution of tandem repeats in *Nicotiana* is well studied, the contribution of other repetitive DNA sequences is less well characterised.

Next generation sequencing (NGS) and plant genomes

The earliest efforts to sequence and characterise plant and animal genomes used traditional cloning and Sanger sequencing (Sanger *et al.*, 1977). Whilst several improvements in reaction chemistry allowed for longer reads and higher throughput, sequencing even small genomes was a time-consuming, expensive and labourious undertaking (Blattner *et al.*, 1997; Cole *et al.*, 1998; Kaul *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001). The development of high-throughput NGS, where as many as 100 million DNA fragments are sequenced simultaneously, has allowed cheap, effective and in-depth analysis of the genomes of non-model eukaryotes including plants.

The new technology beginning to revolutionise biology includes, among others; [1] Roche 454 pyrosequencing, which uses DNA templates that are amplified by emulsion PCR whilst covalently attached to beads. Subsequently the beads are deposited in picotitre wells and the extension of DNA templates is recorded via the fluorescence emitted when a nucleotide is added to the growing molecule (Margulies *et al.*, 2005). Hundreds of thousands of DNA sequences, up to ~700 bp in length, can be analysed in a single run. [2] The various Illumina platforms use sequencing by synthesis to generate many millions of shorter DNA fragments (up to 108 bp) giving greater throughput at the expense of read length. These sequencing platforms have been used to *re-sequence* one thousand human genomes to fully realise the genetic diversity of our species (Overbeek *et al.*, 2005; Zhang & Dolan, 2010), and to sequence *de-novo* another one thousand species of economic, conservation or medical interest at the Beijing Genomics

Institute (BGI). Only ten years ago such aims would have been considered impossible to achieve within any realistic time frame or budget. Indeed the cost of sequencing per base has decreased so dramatically that now it is possible to sequence the *A. thaliana* genome at 1000x coverage for as little as £32,000 (Glenn, 2011).

Next generation sequencing technology has been used to revolutionise our understanding of gene expression in many systems (Roeding *et al.*, 2009; Trick *et al.*, 2009; Yassour *et al.*, 2009; Buggs *et al.*, 2010b; Buggs *et al.*, 2010c; Buggs *et al.*, 2012). It is now conceivable to assess almost the complete transcriptome of a species *de-novo*, both simply and cheaply. Moreover read-depth of a given gene is proportional to its abundance within the transcriptome allowing the quantification of gene expression with relative ease. The combined use of 454 and Illumina data has led to the successful characterisation and quantification of the ESTs in *Tragopogon* allopolyploids (Buggs *et al.*, 2010c). This approach has proved to be an order of magnitude more economical when compared with more traditional approaches (Buggs *et al.*, 2012) and promises to provide the much needed data to understand gene expression changes in hybrid plants.

Recent studies using Roche 454 sequencing data have encompassed the analysis of repetitive DNA in several angiosperm species including *Glycine max, Medicago truncatula, Pisum sativum* and *Hordeum vulgare* (Wicker *et al.*, 2006; Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker *et al.*, 2009; Novak *et al.*, 2010; Macas *et al.*, 2011). In these studies a 'genome skimming' (Straub *et al.*, 2012) approach, where a random sample of the genome is sequenced, has allowed the characterisation of repetitive DNA families. By taking only a small sample of the genome (typically 0.5-10%), the data consist mostly of repetitive sequences. Genic regions in one or a few copies throughout the genome are unlikely to be represented in such a dataset. On the other hand repeat sequences present in tens of thousands of copies will be detected numerous times. Identification, categorisation and the relative proportion of sequence types in the dataset allow the characterisation of the repeat component of the genome. For plant genomics NGS technology provides a quick and cost effective approach to old problems, where low throughput was prohibitive in the past (Kelly & Leitch, 2011).

Aim and scope of the thesis

Although recent studies have used NGS to analyse the repeat component of fairly large diploid plant genomes, none have focused on the phenomenon of repetitive DNA evolution and genome size change in allopolyploids. In this thesis I use, and extend, a repeat identification and quantification pipeline developed in the laboratory of Jiri Macas (Macas *et al.*, 2007; Hribova *et al.*, 2010; Novak *et al.*, 2010) and provide the first example of NGS technology applied to repetitive DNA evolution in allopolyploids.

The aim of this thesis is to assess alterations that occur to the genome following allopolyploidy. More specifically the basal objective was to use graph-based clustering to compare repeats in the progenitors and the allopolyploids, so that comparisons and evolutionary inferences can be made. This approach runs through the thesis, with each chapter containing a description of the repeat component of one or more allopolyploids and the diploid progenitors. I was particularly interested in the fate of each progenitor genome: Do the progenitor sub-genomes respond differently in allopolyploids? This question is explored in detail in chapters two and three using the young allopolyploid *Nicotiana tabacum* as a model. In chapter four I explore characteristics of diploidisation (where an allopolyploid becomes more diploid like) and genome size change in *Nicotiana* section *Repandae*. The last chapter is a brief overview of the novel findings of this thesis and contains a synthesis of its importance to the understanding of allopolyploid evolution in general.

Chapter 2 Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum,* predominantly through the elimination of paternally derived repetitive DNAs

Publication information

This chapter is based on a paper published in *Molecular Biology and Evolution,* for which I was the lead author. Some of the co-authors contributed to data analysis: Clustering was performed in the lab of Jiri Macas, Steve C LeComber helped produce Figure 2.1 and Petr Novak and Jiri Macas helped produce Figure 2.3 and 2.5. All authors on this paper contributed to editing, proof reading and commenting on the original manuscript.

Renny-Byfield S, Chester M, Kovařík A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novák P, W. Chase M, ***et al.*** **2011.** Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28**: 2843-2854.

Summary

Next generation sequencing was used to characterise 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 representative repetitive DNA sequences in the genomes of the progenitor species and *N. tabacum* indicate the presence of all major types of retroelements previously 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. In the derived allotetraploid *N. tabacum*, there is evidence of genome downsizing and sequence loss across most repeat types. This is particularly evident amongst the Ty3/*Gypsy* retroelements in which all families identified are under-represented, as is 35S ribosomal DNA. Comparative analysis of repetitive DNA sequences indicates the T-genome of *N. tabacum* has experienced greater sequence loss than the S-genome, revealing preferential elimination of paternally derived repetitive DNAs at a genome-wide level. Thus the three genomes of *N. sylvestris, N. tomentosiformis* and *N. tabacum* are dynamic, stable and downsized respectively.

Introduction

Angiosperm evolution has been heavily impacted by polyploidy, which has occurred in the ancestry of most, if not all, species (Soltis *et al.*, 2009; Jiao *et al.*, 2011). Polyploidy itself may induce revolutionary changes in genome composition in early generations (Leitch & Leitch, 2008), a phenomenon explored here. Interspecific hybridisation combined with whole genome duplication (allopolyploidy) provides a natural experiment in genome perturbation. 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'. We now know allopolyploidy can perturb the epigenome and transcriptome (Pikaard, 2000; Madlung *et al.*, 2002; Adams *et al.*, 2003; Hegarty *et al.*, 2005; Hegarty *et al.*, 2006; Rapp *et al.*, 2009; Buggs *et al.*, 2010a; Buggs *et al.*, 2010b; Buggs *et al.*, 2010c; Chague *et al.*, 2010; Paun *et al.*, 2010) and lead to divergence in chromosome number and structure (Lim *et al.*, 2004a; Gaeta *et al.*, 2007; Lim *et al.*, 2007; Lim *et al.*, 2008; Chester *et al.*, 2012) and repetitive DNA content of the genome (Koukalova *et al.*, 1989; Song *et al.*, 1995; Lim *et al.*, 2000a; Shaked *et al.*, 2001; Kovarik *et al.*, 2004; Salina *et al.*, 2004; Gaeta *et al.*, 2007; Petit *et al.*, 2007; Koukalova *et al.*, 2010; Petit *et al.*, 2010; Jiang *et al.*, 2011; Renny-Byfield *et al.*, 2011; Renny-Byfield *et al.*, 2012). Moreover polyploid-associated genetic change occurs rapidly in some species, 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 & Wendel, 2003; Feldman & 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; Liu & Wendel, 2003; Matyasek *et al.*, 2003; Adams & Wendel, 2005; Salmon *et al.*, 2005; Leitch *et al.*, 2008). However there are only a few examples of allopolyploid- or interspecific hybridisation-associated changes for the transposable elements (TEs; reviewed in 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 miniature inverted repeat transposable elements (MITEs) in rice following alien DNA introgression from related wild species (Liu & Wendel, 2000). Activation was transient, involved 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 in which TE activation following allopolyploidy is not apparent, notably in *Gossypium* synthetic allopolyploids (Liu *et al.*, 2001) 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).

Genome size estimates have indicated that many allopolyploids have undergone genome downsizing (Dolezel *et al.*, 1998; Leitch & Bennett, 2004; Beaulieu *et al.*, 2009). Certainly the balance between transposition and DNA deletion will influence genome size and turnover of DNA sequences (Leitch & 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 cause genomic *in situ* hybridisation (GISH) to fail, as in some relatively young *Nicotiana* allopolyploids (ca. 5 million years of divergence; Lim *et al.*, 2007).

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

The genus *Nicotiana* (Solanaceae) provides an excellent model group for studies on the consequences of allopolyploidy, since the genus consists of c. 70 species, ~40% of which are documented to be allotetraploids derived from six independent polyploid events (Clarkson *et al.*, 2005; Leitch *et al.*, 2008). The allopolyploid species studied here, *Nicotiana tabacum* (tobacco), is particularly appropriate for study as it is relatively young (less than 200,000 yrs old; Leitch *et al.*, 2008) and is derived from ancestors 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 cytogenetic 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). 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 fourth generation synthetic *N. tabacum*, a similar translocation to the one fixed in natural *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 retroelement 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 Roche 454 high-throughput DNA sequencing and a genome coverage of ~1%, Macas *et al*. (2007) were 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) and banana (Hribova *et al.*, 2010). These efforts have aided in the genome characterisation of species where a genome assembly is not available, or with the generation of repeat libraries with potential uses during subsequent genome assembly steps.

However, these studies have not focused on addressing the question of how repeat sequences respond to allopolyploidy, the principal objective of this chapter. Here I 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, at random, 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 the analysis of patterns of evolution subsequent to allopolyploidy for the most abundant repetitive sequences.

Materials and Methods

Plant material

Nicotiana sylvestris (ac. ITB626) and *Nicotiana tabacum* ac. SR1, Petit Havana, were obtained from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France. *Nicotiana tomentosiformis* (ac. NIC 479/84) was from the Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. 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), amplified fragment length polymorphisms (AFLPs, unpublished data) and gemini virus related DNA (GRD) insertions. 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 1/8th of a 70 x 75 picotiter plate was used. 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 were clipped of the first ten bases to remove any associated adapter sequences. The stand-alone BLAST program was used to screen 454 reads for similarity to the appropriate plastid genome. Reads with significant hits (e-value <1e10^{-6}) to plastid DNA were excluded from further analysis.

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 number of reads with significant hits (e-value <1e10^{-8}) was recorded for each sequence. All other BLAST parameters were set to default. 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). An *in silico N. tabacum* was created, consisting of a random set of 35,000 reads from each of the parental datasets reflecting 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.

Individual *N. tabacum* 454 reads 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. The resulting data were plotted in the R statistical package (R Development Core Team, 2010).

Genome-wide analysis of mean similarity of related sequences

The genomes of parental and progenitor species were analysed by comparing mean sequence similarity between related sequences. BLASTn analysis (e-value cut-off of <1e10^{-6}) was used to identify similarity, while custom Perl scripts and the R statistical package were used to extract and calculate the mean sequence similarity for each sequence and its corresponding hits, provided that high-scoring segment pair (HSP) was above 80 bp in length.

Clustering, contig assembly, graph visualisation and species-specific read mapping

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 MGBLAST analysis between sequences. A Fruchterman-Reingold algorithm is used to position each node where more similar sequences are placed closer together, whereas more distantly related reads are placed further apart. 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 for similarity to known repeat databases. 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. Assembly of reads within each cluster was performed using CAP3 requiring 80% similarity along 55% of the sequence read.

Using the CLC Genomics Workbench v. 3 reads derived from each species were independently mapped to the reference sequences derived from the clustering and assembly process. 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 each 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 database, using an e-value cut-off score of <1e10^{-6}. Additional annotation using the BLAST function on the Gypsy Database (GyDB) was required in order to establish the family to which Ty3/*Gypsy*-like elements belonged (Llorens *et al.,* 2008). The total GR and GP of a given repeat was calculated by summing all GR and GP estimates for clusters associated with that repeat type.

All scripts are available on request.

Results

454 high-throughput DNA sequencing

454 GS FLX Titanium sequencing of genomic DNA of *N sylvestris, N. tomentosiformis* and *N. tabacum* returned between 68,000 and 75,000 reads per species, with an average read length of 360-370 bp. Filtering for plastid contaminants and trimming of primer sequences resulted in 19-25 Mb of DNA sequence for each species. This amounts to ~0.9% coverage of the *N. sylvestris* (2636 Mbp/1C) genome, ~0.8% coverage of the *N. tomentosiformis* (2682 Mb/1C) genome and ~0.5% coverage for *N. tabacum* (5061 Mb/1C) (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, I conducted pair-wise 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 to each parent (Fig. 2.1 A and B). The output reflects the abundance of sequences in the *Nicotiana* genomes. One might 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 of one parental genome by the other parental genome.

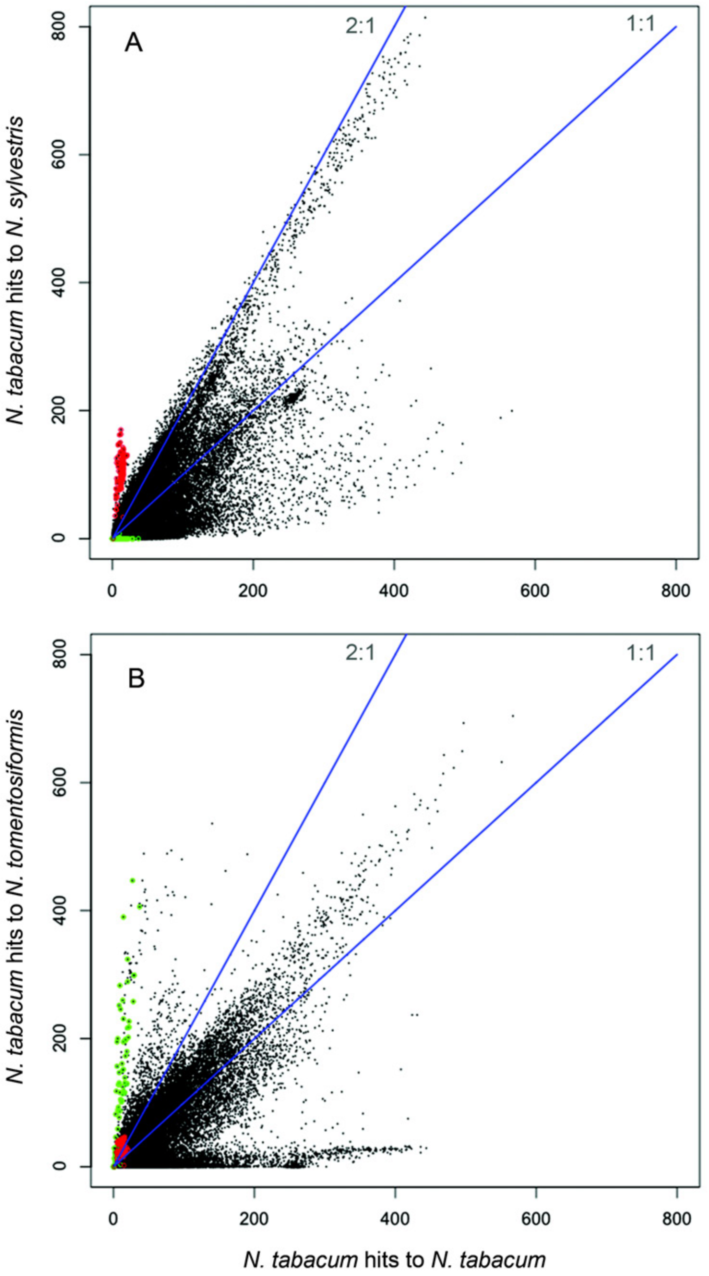

Figure 2.1 Genome comparisons using pair-wise similarity analysis of individual 454 reads.

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 cut-off of <1e10^{-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 and dilution (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 roughly similar copy numbers.

Those sequences falling above the 2:1 line are under-represented in *N. tabacum*. Similarly, sequences falling on a 1:1 line are expected to be of roughly similar abundance in both parents.

Fig. 2.1 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. 2.1 A, which shows the analysis of *N. tabacum* against *N. sylvestris*, there is a distinct clustering of reads on or close to the 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 annotated using similarity searches to RepBase or pfam domains. In comparison, when the same analysis is conducted using the *N. tomentosiformis* dataset, fewer sequences fall on the 2:1 (Fig. 2.1 B).

The analysis in Fig. 2.1 A 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*. These sequences are predominantly rDNA (highlighted red in Fig. 2.1 A and 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; see appendix Figure A.1).

In Fig. 2.1 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. 2.1 B). This pattern is absent in the *in silico N. tabacum* made from a mix of the two parental datasets (Figure A.1).

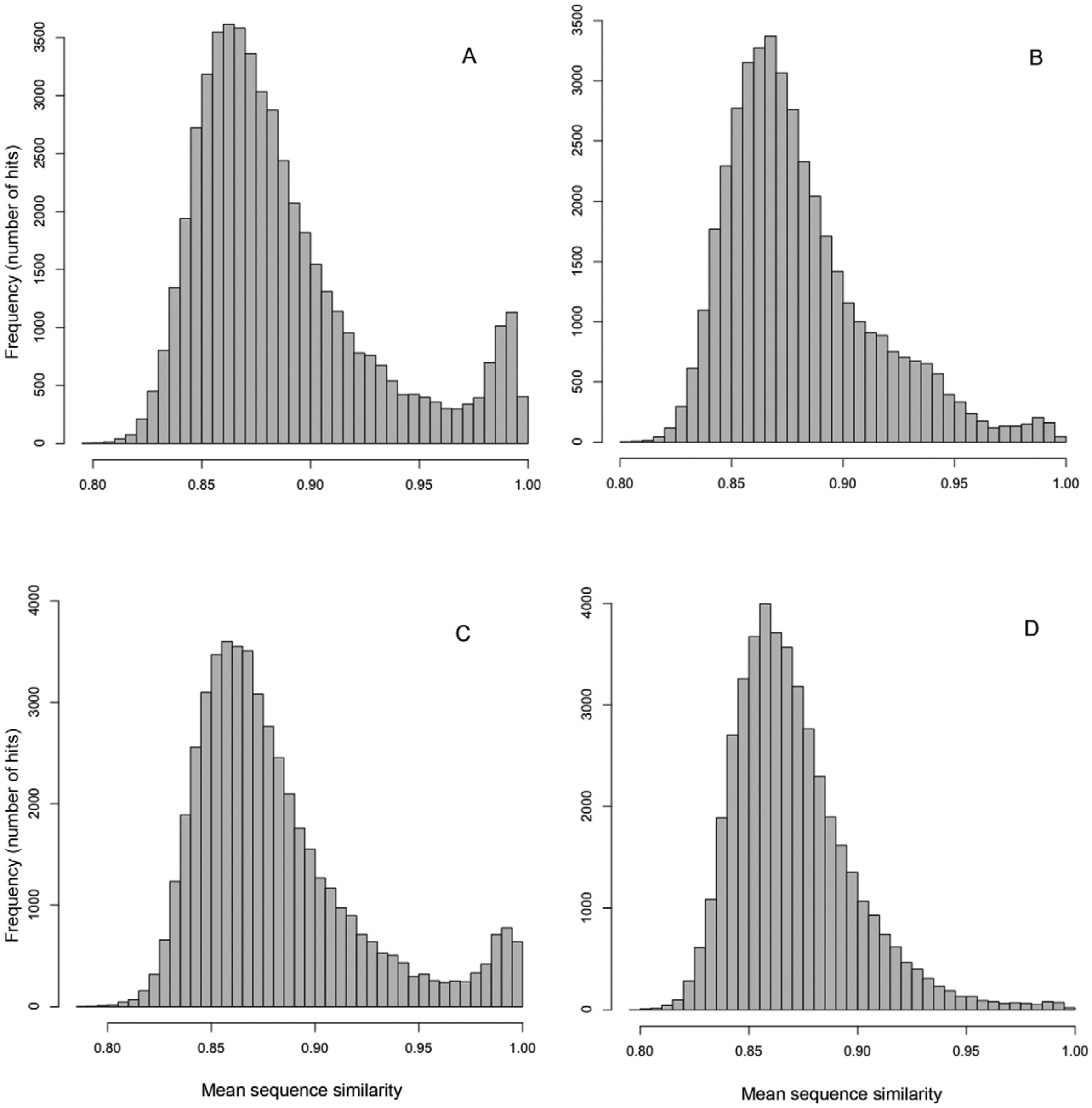

Figure 2.2 Sequence similarity between 454 reads

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).

For the reads above the 2:1 line (i.e. they are under-represented 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 and sum of residuals was only 9.6 and 56822 repesctively. Amongst the reads above the 2:1 line in Fig. 2.1 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 (RT) pfam domains.

Comparison of genome wide sequence similarity

For all reads in each of the three datasets I calculated the mean sequence similarity between that read and all related sequences within the same dataset. Figure 2.2 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. 2.2 A). These latter sequences are likely either highly constrained by selection or have experienced recent expansion/homogenisation. 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 and include 5S and 35S rDNA sequences (totalling 353 of the reads).

Table 2-1 Output of the clustering and assembly algorithm

The combined dataset (producing the reference sequences) and the species specific read mapping analyses. The repeat identification algorithm is described in detail in Novak *et al*. (2010).

		number of clusters in assembly	number of contigs in assembly	minimum/maximum contig length	% of reads mapped to contigs
combined assembly		16229	17443	107/9632	N/A
species specific read mapping	*N. tabacum*	8496	10464	109/5198	44
	N. sylvestris	9791	11446	107/5198	63
	N. tomentosiformis	6131	7378	108/9362	53

In addition there were a number of Ty3/*Gypsy*-like repeat sequences, although considerably less than rDNA. *Nicotiana tomentosiformis* (Fig. 2.2 B) lacks such an abundance of sequences with a high mean sequence similarity. The *in silico N. tabacum* (Fig. 2.2 C) exhibits a secondary peak of high mean sequence similarity as seen in *N. sylvestris* (Fig. 2.2 A), but the peak is absent in natural *N. tabacum* (Fig. 2.2 D).

Clustering and contig assembly

All 454 high-throughput DNA sequencing reads from the three *Nicotiana* species (>70 Mb of DNA sequence) were combined and subjected to a graph-based clustering repeat identification procedure described in detail by Novak *et al.* (2010). This leads to partitioning of sequencing data into groups of overlapping reads representing individual repeat families. After read-mapping, the average read-depth in each cluster reflects the genomic proportions of the corresponding repeat, and thus read-depth analysis across each cluster was used to estimate the repeat composition in the genomes of each species. Details of the repeat identification and assembly output are given in Table 2-1. The contribution of each species to the 30 largest clusters is shown in Figure 2.3. Reads were assembled on a cluster by cluster basis providing reference sequences that were used as a scaffold for the independent mapping of reads for each of the three *Nicotiana* species (Table 2-1 and Fig. 2.4). This allowed characterisation of the average read-depth along the length of the contig (RD), genome representation (GR, calculated as RD × contig length) and genome proportion (GP, calculated as (GR/total size of the dataset in base pairs) × 100) for each species. GP is 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. 2.5) and has sequence similarity to Ogre-like LTR retroelements (Macas & 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. 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. 2.5). 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. 2.5, respectively). Both repeat variants are present in *N. tabacum*.

To check the validity of the contigs developed *in silico* (as described above), I cloned and sequenced a region of cluster 3 contig 8 and found clones sharing between 92-96% identity with the consensus.

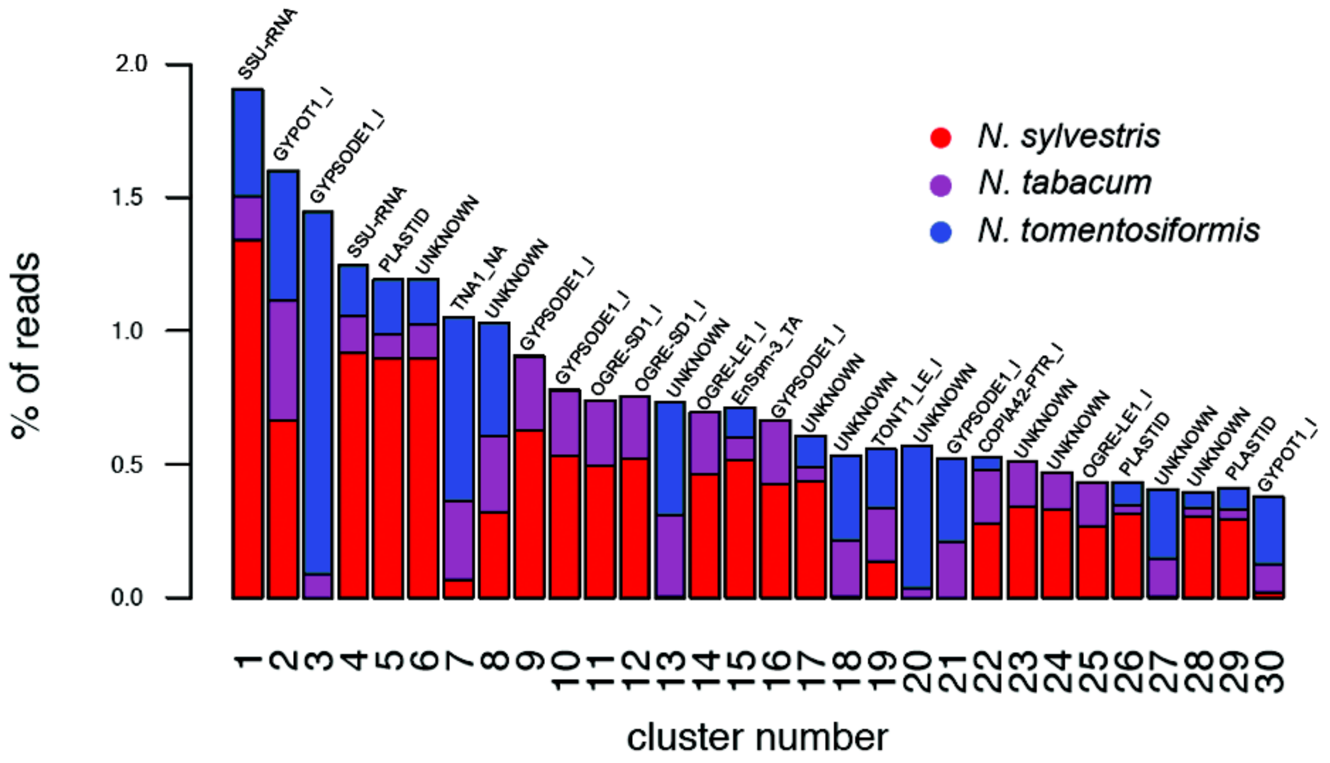

Figure 2.3 The contribution of each species in the 30 largest clusters

Histogram showing the contribution of 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.

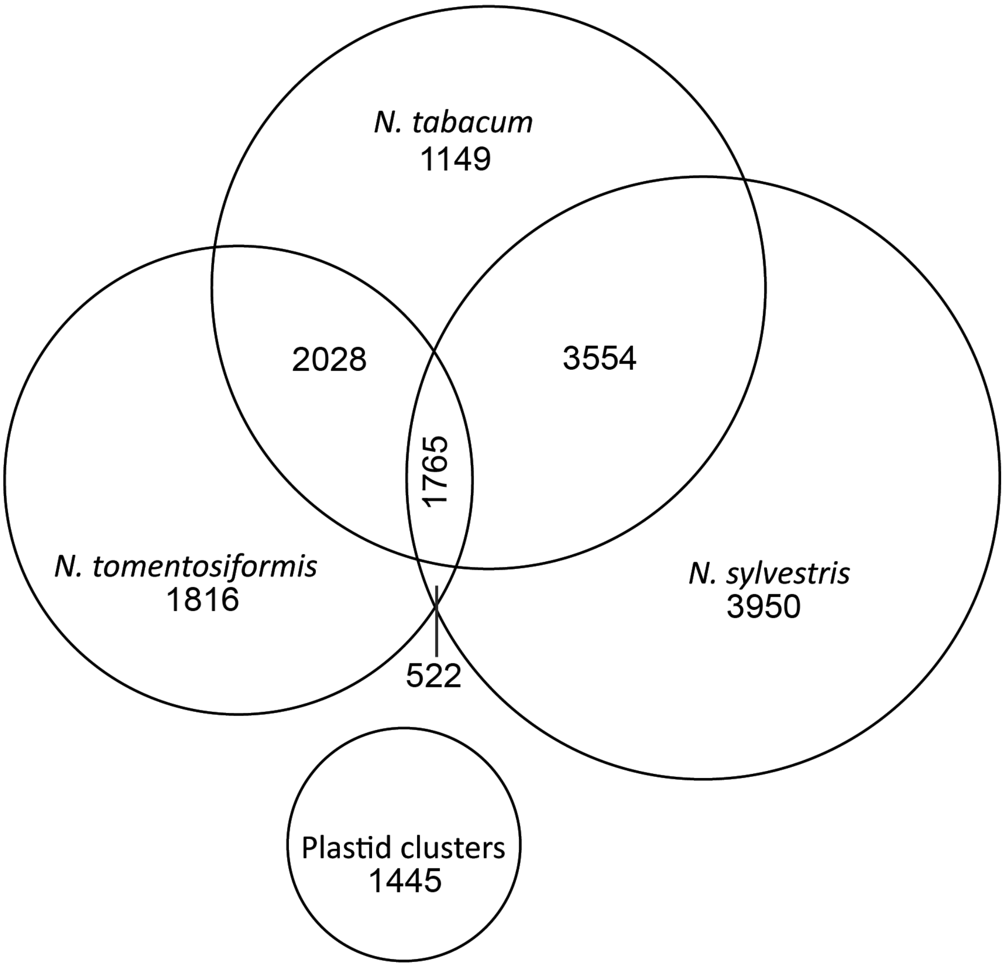

Figure 2.4 Cluster similarity in three *Nicotiana* species

Venn Diagram where the number in each circle (and the intersections) indicates the number of clusters that have reads mapped from each species.

Table 2-2 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 GP 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* GP is 19.8% lower than expected. Figure 2.6 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 lower GP in *N. tabacum* than would be expected based on the proportions observed in the diploid progenitors (Fig. 2.6 B), indicating that sequence loss may have occurred subsequent to allotetraploidy.

Estimates of abundance have shown 35S rDNA makes 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 0.17%, where GP is 84% lower than expected. I 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%; see also chapter

3). In addition pararetrovirus-like sequences are more abundant in the *N. tomentosiformis* genome (0.54%) than they are in both *N. sylvestris* (0.22%) and *N. tabacum* (0.25%), revealing a 34% reduction in GP from expectation (Table 2-2).

Comparing observed with expected genome proportions in N. tabacum

Linear regression was used to compare GP estimates of 14,634 repeat clusters in *N. tabacum* against those in the two progenitor species (Fig. 2.7). If the *N. tabacum* genome was an equal mix of the two progenitors the slope of the regression would have been 0.5 for each of the parental species. The actual estimate of the slope (Fig. 2.7) 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 Figure 2.7, notice that: [1] the fitted surface (pink) 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; [3] this discrepancy is greatest for the most common repeat elements in *N. tomentosiformis;* [4] this pattern is not observed for *N. sylvestris*.

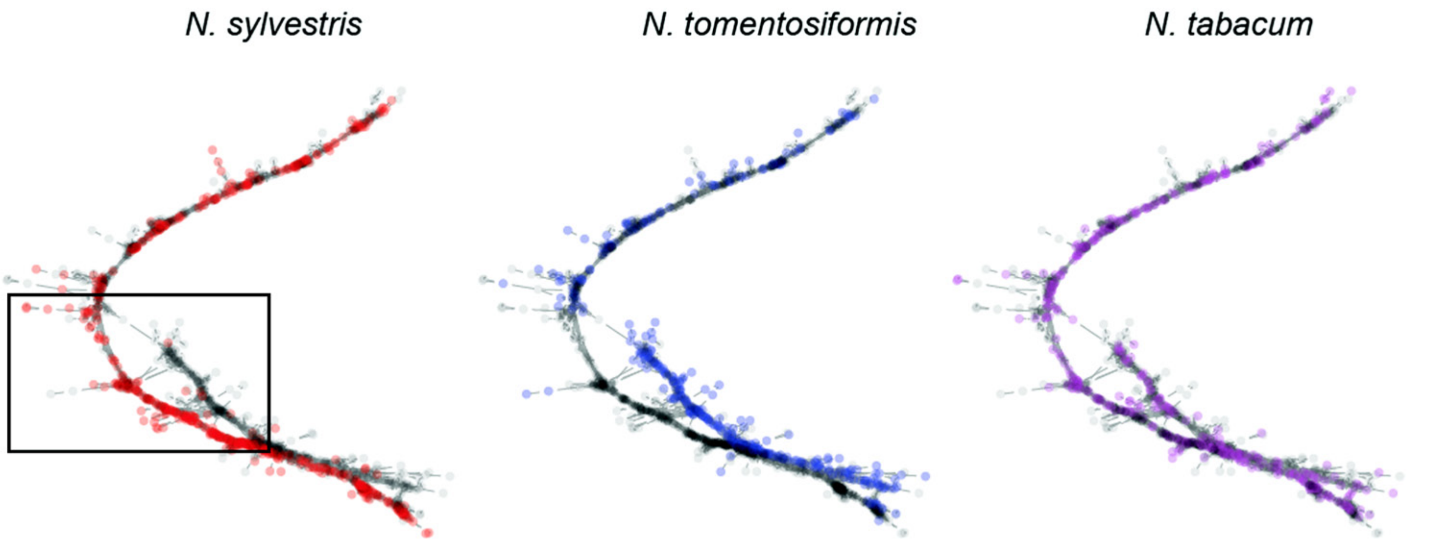

Figure 2.5 Three-dimensional graph networks of CL2

An example of the output of the clustering based repeat assembly algorithm (Novak *et al.*, 2010) showing 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.

Table 2-2 Repetitive DNA in *N. tabacum*, *N. sylvestris* and *N. tomentosiformis*

Genome proportion (GP) of major repeat classes within the *N. tomentosiformis*, *N. sylvestris* and *N. tabacum* genomes. GR = genome representation (number of bp), GP = genome proportion (% of the genome).

			N. sylvestris		*N. tomentosiformis*		*N. tabacum*			
Class	Order	Superfamily	GR	GP (%)	GR	GP (%)	GR	GP (%)	Deviation from parental average	Difference as % of expected GP
Retroelement			6870681	27.17	4340540	22.90	49879834	20.52	-4.49	-18.03
	LTR		6614318	26.16	4219779	22.26	4843776	19.93	-4.28	-17.68
		Ty3/*Gypsy*	5694116	22.52	3800409	20.05	4148861	17.07	-4.22	-19.8
		Ty1/*Copia*	864519	3.47	419370	2.26	668687	3.10	0.23	8.01
		Unknown	55682	0.22	21937	0.12	25477	0.10	<0.1	-41.18
	LINE		212097	0.84	74864	0.39	122512	0.50	-0.11	-18.7
		L1	100815	0.40	21044	0.11	29646	0.12	-0.14	-52.95
		RTE	111282	0.44	53820	0.28	77699	0.32	<0.1	-11.11
		Unknown	44266	0.18	N/A	N/A	15167	0.06	<0.1	-33.34
	SINE	TS/TS2	N/A	N/A	23960	0.13	21696	0.09	<0.1	38.46

Table 2-2 continued

			N. sylvestris		*N. tomentosiformis*		*N. tabacum*			
Class	Order	Superfamily	GR	GP (%)	GR	GP (%)	GR	GP(%)	Deviation from parental average	Difference as % of expected GP
DNA transposon			450969	1.78	327076	1.73	410725	1.69	<0.1	-3.7
	Helitron		129168	0.51	128361	0.68	145124	0.60	<0.1	0.84
	Ac		40235	0.20	38761	0.2	46353	0.19	<0.1	-5
	MuDR		132438	0.52	73586	0.39	95050	0.39	<0.1	-14.29
	EnSpm		84257	0.33	45352	0.24	63659	0.26	<0.1	-8.77
	Harbinger		231	<0.01	N/A	N/A	3075	0.01	<0.1	2551
	TIR	hAT	41797	0.17	26123	0.14	33887	0.14	<0.1	-9.68
	Unknown		12595	0.03	14895	0.08	23577	0.10	<0.1	81.82
35S rDNA[1]			430296	1.70	90945	0.48	42313	0.17	-0.92	-84.4
5S rDNA[1,2]			1575	0.01	1750	0.013	4200	0.031	<0.1	150
Satellite		NTRS	0	0	402601	0.21	~80	<0.0001	-0.15	-99
		SYL2[2]	~6750	0.03	0	0	~4730	0.02	<0.1	33

[1] The genic, but not intergenic regions were estimated here

[2] These estimates were based on BLAST read depth to cloned sequences

Discussion

454 Titanium sequencing to estimate repeat abundance

The major repeat composition of the genomes of three *Nicotiana* species has been characterised using Roche 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 (Macas *et al.*, 2007). 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 several angiosperms (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker *et al.*, 2009; Hribova *et al.*, 2010).

A relatively low abundance of TEs was observed in the *Nicotiana* genomes analysed (Table 2-2). *Hordeum vulgare* (1C = 5,439 Mb) with a genome size similar to *N. tabacum* has higher proportions of Ty1/*Copia* (~16%) and Ty3/*Gypsy* (~30%) retroelements (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%, with ~46% Ty3/*Gypsy* elements (Schnable *et al.*, 2009). However, *N. tabacum* has similar abundance of retroelements to *Pisum sativum* (4778 Mb/1C), 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

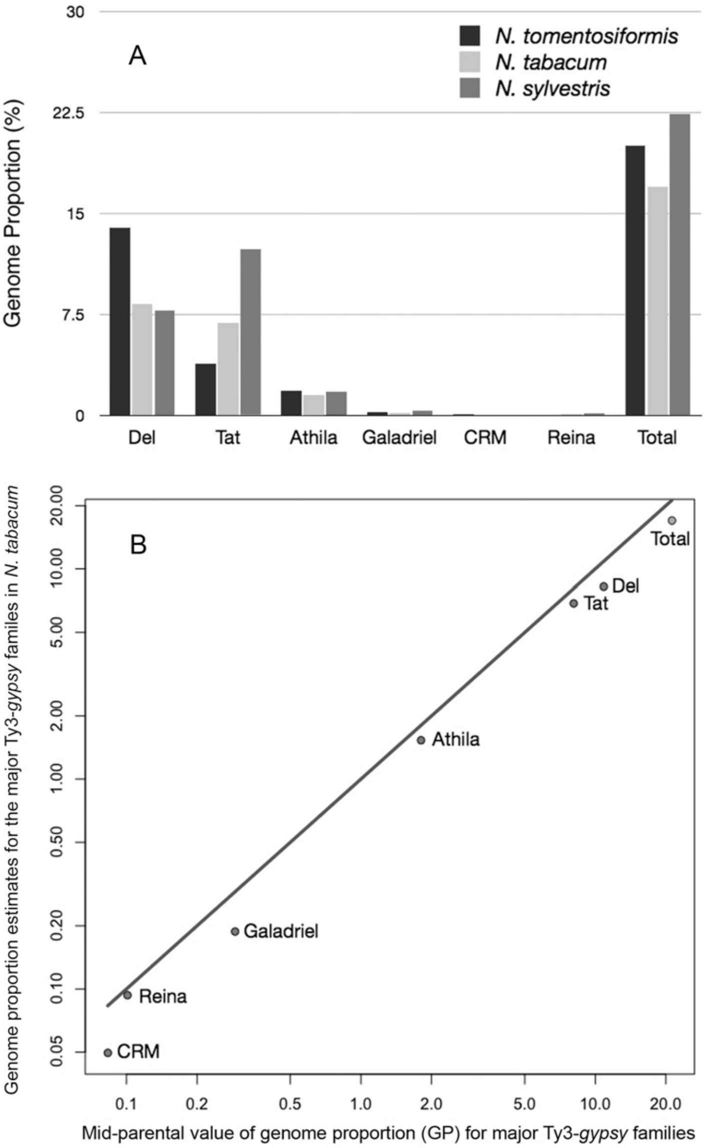

Figure 2.6 Ty3*/Gypsy* retroelements in three species of *Nicotiana*

(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, black line), plotted on a log scale. Genome proportion estimates are below the expected value for all Ty3/*Gypsy* families observed.

consists of Ty1/*Copia* elements and ~11% of Ty3/*Gypsy* (International Rice Genome Sequencing Project, 2005).

It is possible GP estimates of repeat sequences could be low because of an abundance of a diverse range of low copy repeats in the genome. Furthermore, RepBase has only 67 Solanaceae entries (October 2010), including complete and incomplete Ty1/*Copia* and Ty3/*Gypsy* elements, and this may make identification of repeats difficult by sequence similarity searches. However, these interpretations seem unlikely given conserved domains are normally detectable across wide phylogenetic distances. Importantly, on lowering the BLAST e-value threshold to <1e10^{-2} only a small (3-5%) increase in the total proportion of the genome that was annotated. Further support for the estimates of TE abundance provided here (~22% of genome in *N. tabacum*) is provided by re-association 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 & 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 (Figure A.2). Nevertheless, although the absolute numbers of tandem repeats may be underestimated in 454 datasets, the relative abundance of these repeats in the three species of *Nicotiana* are fully concordant.

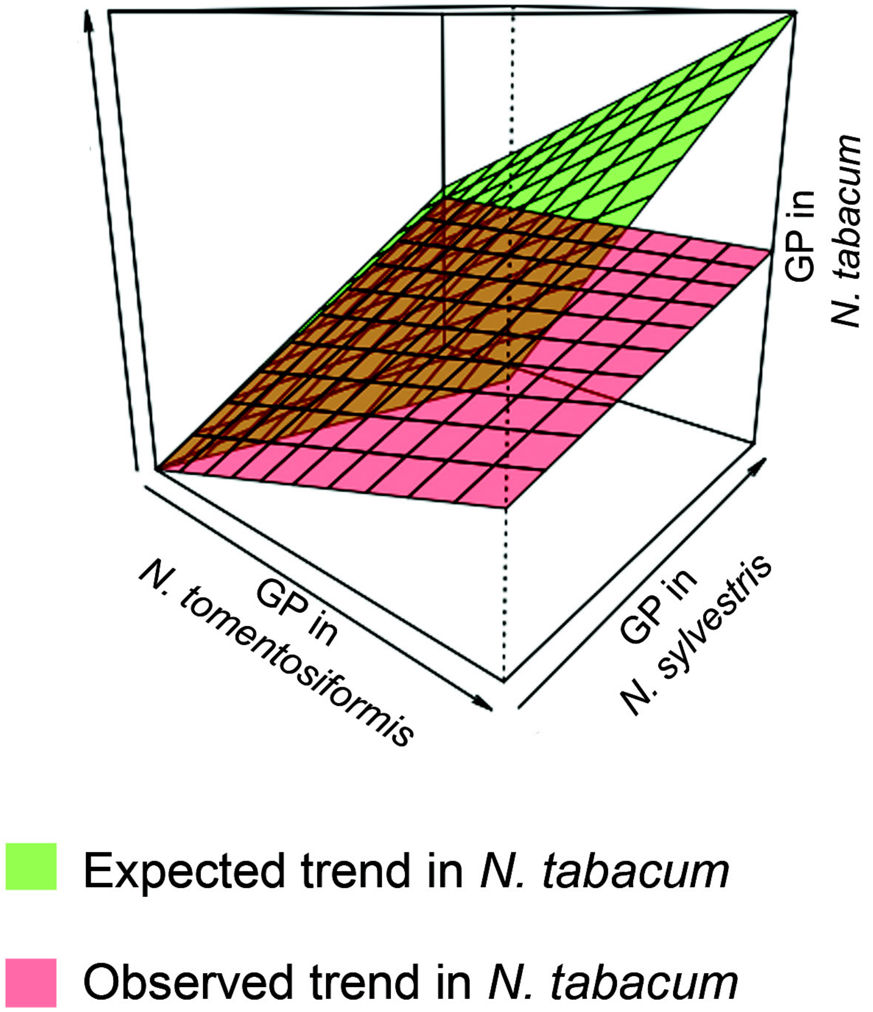

Figure 2.7 Elimination of paternally derived DNA

Linear regression analysis was used 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*. On each axis the data range is for clusters with a GP between 0 – 0.5% of the genome.

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 & Wendel, 2005; Leitch & Leitch, 2008; Feldman & Levy, 2009; Parisod *et al.*, 2009; 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 & 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. 2.2 A). Some sequences in *N. sylvestris* showing evidence of recent expansion belong to the Ty3/*Gypsy* superfamily, other TE groups and rDNA sequences. Ribosomal DNA 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. 2.2 B) and *N. tabacum* (Fig. 2.2 D). The low levels of these sequences in *N. tabacum* perhaps indicate 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 2-2), explaining why only a few copies are found.

Genome downsizing in polyploids

The 1C genome size (GS) of *N. tabacum* (5061 Mb) is ~ 3.7% less than would be expected by summing the sizes of the two parental genomes (*N. sylvestris* = 2636 Mb/1C, *N. tomentosiformis* = 2682 Mb/1C), which has been proposed as evidence of genome downsizing in this species (Leitch *et al.*, 2008; Bennett & Leitch, 2010). 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 & Bennett, 2004). Indeed several allotetraploids in *Nicotiana* have been shown to have a larger genome than predicted based on the sum of their progenitor genomes (see chapter four; 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 2-2). All major Ty3/*Gypsy* like families detected exhibit a lower GP than expected (Fig. 2.6 B), suggesting 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 2-2). [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. Thus independent methods suggest repetitive DNA is being lost from the genome of *N. tabacum* at a rapid rate. Indeed *Nicotiana tabacum* may have lost ~200 Mb of DNA in less than 200,000 years, much of which could be accounted for by the sequences described here. All these observations are in line with a hypothesis of genome downsizing and the disparity between the expected genome size of *N. tabacum* (5318 Mbp/1C) and that estimated by flow cytometry (5061 Mbp/1C; Leitch *et al.*, 2008).

Loss of repeats from paternally derived T-genome of N. tabacum

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 the satellite repeat HRS60, inherited from *N.*

sylvestris, does not deviate from expectation in *N. tabacum*. 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). I 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 from the T-genome of *N. tabacum* (see also chapter three).

On a genome-wide scale I 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. 2.1 B) showed many sequences in *N. tomentosiformis* 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. 2.1 A), they do not deviate as substantially from expectation; the sum of residuals is less than that observed for *N. tomentosiformis*. This suggests that repeats derived from *N. sylvestris* may not be affected to the same degree as those of the T-genome.

I 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 (Fig. 2.7). Linear regression analysis revealed unequal contribution from the progenitor species to the *N. tabacum* genome, with particular under-representation of repeats from *N. tomentosiformis*, 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. 2.7) indicating faithful inheritance from *N. sylvestris*.

The lack of evidence for repeat expansion in *N. tomentosiformis* (Fig. 2.2) indicates that the under-representation 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, these data are congruent with the loss of repeats derived from the T-genome of *N. tabacum* and the maintenance of repeats from the S-genome.

Instability of a particular genomic component of an allopolyploid has also been observed in crosses between *Brassica nigra* and *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 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 *N. sylvestris*–derived repeats, which were maternally inherited. The small RNAs may have acted to enhance the stability and reduce the frequency of DNA loss from the S-genome of *N. tabacum* in early generations after its formation.

The processes giving rise to repeat copy number changes may be stochastic or directed. Evidence for the latter may be the preferential loss of repeats from the

N. tomentosiformis derived T-genome of *N. tabacum* and the repeated loss of the same DNA families in natural and synthetic lines (see also chapter three). These data suggest that the three species studied have experienced distinctive patterns of genome evolution. *Nicotiana sylvestris*, the maternal progenitor of *N. tabacum*, appears to have many features indicating recent expansion of repeats. In contrast, this signature is much less apparent in *N. tomentosiformis,* suggesting that its genome has not experienced substantial rounds of 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. The same families of repeats, in *N. sylvestris*, *N. tomentosiformis* and *N. tabacum* are dynamic, stable and down-sizing respectively, indicating varying fates in different, yet closely related lineages.

Chapter 3 Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of *Nicotiana tabacum*

Publication information

This chapter is based on a manuscript published in *PLoS One*. Ales Kovarik performed most of the Southern blots in Figure 3.4 and Table 3-2, Michael Chester assisted me with the remainder of the Southern blots. All authors on this paper contributed to editing, proof reading and commenting on the published manuscript.

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

Summary

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. I examine the immediate and short-term (less than 0.2 million years) 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 analysis and fluorescence *in situ* hybridisation (FISH) I characterise two highly repetitive DNA sequences (*Nic*CL3 and *Nic*CL7/30). Two independent high-throughput DNA sequencing datasets indicate *Nic*CL3 forms 1.6-1.9% of the genome in *N. tomentosiformis*, sequences that occur in multiple, discontinuous tandem arrays (monomer length of 2.2 kb) scattered over several chromosomes. Abundance estimates, based on sequencing depth, indicate *Nic*CL3 is almost absent in *N. sylvestris* and has been dramatically reduced in copy number in the allopolyploid *N. tabacum*. Surprisingly elimination of *Nic*CL3 is repeated in some synthetic lines of *N. tabacum* by their forth generation. The retroelement *Nic*CL7/30, which occurs interspersed with *Nic*CL3, 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 *Nic*CL3 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 *Nic*CL3 units, was concurrent with rounds of sequence homogenisation.

Introduction

Polyploidy is often associated with interspecific hybridisation (allopolyploidy), where divergent genomes are unified within a single nucleus. It has been suggested that allopolyploidy can induce rapid, reproducible and directional changes to the progenitor-derived sub-genomes (Comai *et al.*, 2003; Liu & Wendel, 2003; Lim *et al.*, 2004a; Skalicka *et al.*, 2005 Chen & Ni, 2006; Matyasek *et al.*, 2007; Feldman & Levy, 2009).

Abundance estimates of repetitive DNA in the genomes of the allopolyploid *Nicotiana tabacum* and its diploid progenitors indicate the preferential elimination of paternally-derived DNA, contributing to genome downsizing in this species (Leitch *et al.*, 2008; Renny-Byfield *et al.*, 2011; see also chapter two). Analysis of wheat F_1 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 (Shaked *et al.*, 2001). Second-generation neo-tetraploids of a cross between *Aegilops tauschii* × *Triticum turgidum* have shown repeated elimination of a sequence derived from *A. tauschii* likely to have occurred during embryo development (Khasdan *et al.*, 2010). However we know relatively little regarding the scale, scope and repeatability of sub-genome specific DNA loss, a gap in our knowledge that is addressed in this chapter.

In this chapter I take advantage of synthetic allopolyploid lines (Th37) made in the 1970s. These synthetic lines were made to mimic natural *N. tabacum* by crossing *N. tomentosiformis* and *N. sylvestris*, the closest known relatives of the true progenitors of *N. tabacum* (Burk 1975; Lim *et al.*, 2006b; Skalicka *et al.*, 2005). An F1 hybrid was converted to a fertile polyploid by *in vitro* callus culture. From a single plant, Th37, several lines were established and have been characterised extensively using both traditional molecular biology techniques and fluorescence *in situ* hybridisation (Kovarik *et al.*, 2004; Lim *et al.*, 2006b; Skalicka *et al.*, 2005). Comparable patterns of evolution are observed in both natural *N. tabacum* and Th37 lines (Skalicka *et al.*, 2005; Volkov *et al.*, 1999), for example rapid homogenization of 35S rDNA units (Kovarik *et al.*, 2004; Volkov *et al.*, 1999). In addition some lines of Th37 show chromosomal translocations similar to that seen in natural *N. tabacum* (Skalicka *et al.*, 2005).

The emergence of high-throughput DNA sequencing (Margulies *et al.*, 2005) has allowed the analysis of 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 (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker *et al.*, 2009; Hribova *et al.*, 2010; Macas *et al.*, 2011; Renny-Byfield *et al.*, 2011). Here I examine *Nicotiana tabacum* and its progenitors *N. sylvestris* (maternal S-genome donor) and *N. tomentosiformis* (paternal T-genome donor) focusing on the genomic organisation and abundance of two newly described repeat families, *Nic*CL3 and *Nic*CL7/30. I also take advantage of synthetic (Th37) lines to examine the fate of *Nic*Cl3 and *Nic*CL7/30 after only four generations post-allopolyploidy. High-throughput DNA sequencing was used 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 accessions were used: [1] *Nicotiana tabacum* cv. SR1 Petit Havana and cv. 095-55 and [2] *Nicotiana sylvestris* ac. ITB626, both originating from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France. [3] *Nicotiana tomentosiformis* ac. NIC 479/84 (Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany), TW142 (USDA, 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 (1973) and characterised previously (Skalicka *et al*., 2003; Skalicka *et al*., 2005). [5] Synthetic *N. tabacum* TR1-A in generation S0, generated and characterised by Lim *et al.* (2006b). [6] Diploid species in section *Tomentosae* (Kelly *et al*., 2010) derived from USDA: *N. tomentosa* Ruiz and Pav.; *N. kawakamii* Y. Ohashi; *N. otophora* Griseb., *N. setchelli* Goodsp.; and *N. glutinosa* L. (section *Undulatae*, formerly section *Tomentosae* (Goodspeed, 1954; Knapp *et al*., 2004).

High-throughput sequencing of genomic DNA

I used Roche 454 FLX pyrosequencing (454 sequencing) as generated in Renny-Byfield *et al*. (2011). Sequence reads are deposited in the NCBI sequence read archive (SRA) under the study accession number SRA023759. Sequencing was performed here using the Illumina Genome Analyzer xII at The Genome Centre Queen Mary University of London, covering 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 line (details of the sequencing output can be found in Table 3-1; sequence reads were submitted to the NCBI SRA under the study accession number: SRA045794). The *N. tomentosiformis* accession NIC 479/84 was chosen because it most closely resembles the T-genome of *N. tabacum* (Murad *et al.*, 2002; Petit *et al.*, 2007), although there is no *N. sylvestris* accession that is considered to be more closely related to the *N. tabacum* S-genome than any other (Petit *et al.*, 2007).

Table 3-1 Dataset size and average read length for the four Illumina runs used in this analysis

species	average read length (bp)	dataset size (bp)
N. tomentosiformis	96.7	1,390,138,328
N. sylvestris	96.7	1,259,648,345
N. tabacum (SR1)	96.7	2,046,389,071
TR1-A (synthetic tobacco)	96.5	2,318,504,049

Clustering, contig assembly and sequence analysis

A graph-based clustering approach 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*. (2011). 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. 3.1 A and Fig 3.2 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 distal (described in detail in Novak *et al.*, 2010). Contig assembly is performed with reads from each cluster and contigs are named according to the number of the cluster from which they derive (X) and *Nic* designates *Nicotiana*, i.e. *Nic*CLX. Each cluster typically generates multiple contigs, each of which is designated a number (Y), giving a format *Nic*CLX contigY. All contigs assembled in this work are available to download 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.

I estimated the genomic abundance of contigs in each cluster using the "map reads to reference" function of CLC Genomics workbench v. 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 (Jurka *et al.*, 2005) in order to identify, where possible, the repeat type from which they derive.

Analysis of Nic*CL3 using paired-end reads*

Analyses of paired-end Illumina data from *N. tomentosiformis* and *N. tabacum* was used to assess the occurrence of sequences where one of the paired reads hits *Nic*CL3 and the partner read did not. Beforehand 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 55% of the sequence read) against *Nic*CL3 contig8. 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 *Nic*CL3 (termed solo HSPs), were subjected to further sequence similarity searches to all other clusters. Those that hit other contigs in the *Nic*CL3 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 *Nic*CL3 contig 8.

Sequence similarity in Nic*CL3*

I compared sequence similarity of *Nic*CL3 derived sequences in *N. tomentosiformis* and *N. tabacum* using 454 reads described in Renny-Byfield *et al*. (2011). Reads deriving from *N. tomentosiformis* and *N. tabacum* were analysed by BLASTn using the stand alone BLAST program (Altschul *et al*., 1990) 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 I analysed a mix of all the *Nic*CL3 derived reads from both of the progenitor species. Pair-wise similarity scores for *Nic*CL3 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 (R Development Core Team, 2010).

BLASTn was used to analyse the proportion of *Nic*CL3 reads that matched the consensus sequence (*Nic*CL3, contig 8) at any given nucleotide using custom BioPerl scripts. Sequences from *N. tomentosiformis* and *N. tabacum* were analysed seperaetly.

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 NH_4 Buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP and 0.2 µM of each primer pair.

(i) Primer pair 1 (forward: 5′-GGTAGAGTAGTGATGAGG-3′ reverse: 5′-TGGTGGATTAAGGATTGG-3′, Fig. 3.1 B, filled arrows). PCR primers were designed from *Nic*CL3 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) Primer pair 2 (5′-TAAAACTCCCAACATCCG-3′ and reverse 5′-TGGGTATAGTGAAGACGA-3′, Fig. 3.1 B, open arrows). PCR primers were designed against a second region of *Nic*CL3 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′-TGTGTTGGGCTGTTTTGT′ and reverse 5′-CTTGCTGCTCTCTTGACT-3′). PCR primers were designed against *Nic*CL7 contig 7. PCR reaction followed that described for primer pair 1.

Cloning and sequencing

PCR products of *Nic*CL3 and *Nic*CL7 were cleaned using the Qiagen PCR purification kit and cloned using a TOPO® cloning kit with the pCR®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 fluorescence *in situ* hybridisation (FISH) and Southern blot hybridisation.

Probes for FISH

(1) Probes were prepared from clone number 9 (NCBI accession JQ899200) of *Nic*CL3 using primer pair 1 and from clone number 1 (NCBI accession JQ899201) of *Nic*CL7 using primer pair 3. PCR amplification used the conditions described above 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* (Chester *et al.*, 2010), 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′-CGGAGAATTAGGGTTCGATTC-3′ and AB101R 5′-ACGAATTCATGGTCCGGTGAAGTGTTCG-3′, the latter modified from Sun *et al.* (1994).

(3) Total genomic DNA for genomic *in situ* hybridisation (GISH) from *N. tomentosiformis* (ac. NIC 479/84) and *N. sylvestris* was labelled with biotin-16-dUTP and digoxigenin-11-dUTP, respectively, using the Roche Nick Translation Kit according to the manufacturer's 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℃ in 100% ethanol. Root-tip metaphases were spread onto glass slides after enzyme digestion as described in Lim *et al.* (1998), and checked for quality using phase contrast microscopy.

FISH followed the protocol described in Lim *et al*. (2006a). Briefly, probe DNA (delivering 50 ng of cloned, or 100 ng of genomic probe per slide) was added to the 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 Biochemicals 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 theoretical stringency of 110% (60% v/v formamide, 0.1x SSC at 42°C). Slides were checked to ensure no signal could be visualised and 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*. (2000), 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 ^{32}P-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 ^{o}C (Sambrook & Russell, 2001). 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 ^{o}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 *Nic*CL3 used in the FISH experiments. All materials, scripts 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*. (2010). Briefly sequence reads are subjected to pair-wise sequence similarity analysis where related sequences are grouped 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 contigs allows estimation of genomic proportion (GP) of the corresponding repeat. Moreover *N. tabacum* is a *symmetrical* hybrid since both progenitors have roughly the same genome size (~2,650 Mbp/1C; Bennett & Leitch 2010). Therefore for a uni-parentally inherited repeat, the expected genome proportion (GP) in *N. tabacum* is 0.5 of the parental GP.

Read-depth analysis revealed two clusters (*Nic*CL3 and *Nic*CL7) to be highly abundant in the genome of *N. tomentosiformis*. Illumina sequencing read-depth across *Nic*CL3 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 that would be predicted given additivity of the parents (0.80%/0.95%; Table 3-2).

A graphical representation of sequence relationships in the cluster containing *Nic*CL3 is shown in Figure 3.1 A. Reads form a circle-like pattern indicative of direct terminal or tandem repeats (*Nic*CL3 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. *Nic*CL3 does not have this pattern due to its length (2.2kb). Not all of the ~360bp reads that make up the graph share sequence similarity (i.e. not all reads overlap as with a short monomer) and not all nodes are connected. 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 to a wheel like pattern. Copy number estimates along the most abundant contig (8) in the cluster *Nic*CL3 are shown in Figure 3.1 B. A MGBLAST search was conducted using the consensus *Nic*CL3 monomer as a query to *N. tabacum* genome survey sequences (GSSs; e-value < 1e10^{-15}). This produced 741 hits along the whole length of the monomer, with 381 hits showing 95% to 100% similarity, supporting the restriction digest, sequencing and clustering/assembly data.

Sequence similarity searchers of *Nic*CL3 to RepBase returned a small region (positions 717-953 with 40% amino acid identity) with similarity to GYPSODE1_I a Ty3/*Gypsy*-like retroelement identified in *Solanum demissum* (Jurka *et al.*, 2005) while searches against the pfam conserved protein domain database returned no matches. Regions with similarity to *Nicotiana tomentosiformis* endogenous pararetrovirus (NtoEPRV) insertion sites (Matzke *et al.*, 2004) were identified and indicated in black (Fig. 3.1 B). The *Nic*CL7 cluster is closely related to cluster 30 (*Nic*CL30) and they are likely derived from the same repeat family. Therefore, they were merged into a single network, hereafter called *Nic*CL7/30, shown graphically to be circle-like (Fig. 3.2 A-C).

Protein BLAST searches indicate that reads within *Nic*CL7/30 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). *Nic*CL7/CL30 is likely to be a chromoviruse-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 *Nic*CL3 and *Nic*CL7/30 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.

I analysed Illumina paired-end data to assess the proportion of paired end sequences where one read hits *Nic*CL3 and the other member of the pair did not (solo HSPs). In *N. tomentosiformis* and *N. tabacum*, 3.16% and 8.80% of paired reads had only one match (solo HSPs) to the *Nic*CL3 respectively. In *N. tomentosiformis* there were 95 instances where one sequence of a pair matched *Nic*CL3 while the other matched *Nic*CL7/30. In *N. tabacum* comparisons of the distribution of solo HSPs along the length of *Nic*CL3 revealed regions of the sequence with high proportions of solo HSPs (Fig. 3.1 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 the *Nic*CL3 monomer corresponds closely with the distribution of solo HSPs (compare Fig. 3.1 B with Fig. 3.1 C).

Table 3-2 *NicCL3* and *Nic*CL7/30 in the genomes of *N. tabacum*, *N. sylvestris* and *N. tomentosiformis*

Estimated abundance of two families of repetitive DNA sequences in the genomes of the allotetraploid *N. tabacum* and progenitor species *N. sylvestris* and *N. tomentosiformis*.

		Abundance of cluster: % of genome (454 / Illumina estimation)				
cluster name	most abundant contig (length in bp)	*N. tomentosiformis*	*N. sylvestris*	parental additivity	*N. tabacum*	TR1-A (S0 synthetic tobacco)
*Nic*CL3	*Nic*CL3 contig 8(2926)	1.91 / 1.60	<0.01 / <0.01	0.95 / 0.80	0.10 / 0.09	NA / 0.77
*Nic*CL7/30	*Nic*CL7 contig 7(4759)	1.40/1.27	0.15/0.20	0.78/0.74	0.52 / 0.56	NA/0.71

Cloning regions of Nic*CL3 and* Nic*CL7*

PCR using primer pair 1 (thick black arrows in Fig. 3.1 B) against the consensus of *Nic*CL3 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 *Nic*CL7 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 *Nic*CL3 and clone 1 for *Nic*CL7 were chosen for further analysis.

FISH

FISH using the *Nic*CL3 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. 3.3 A, C and Table 3-3). 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*. 2000b) lacks any detectable signal. In contrast there is *Nic*CL3 signal at an interstitial locus on the orthologous 18S rDNA-bearing chromosome of the diploid relative *N. kawakamii* (Fig. 3.3 I). *Nic*CL3 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. 3.3 E). All *Nic*CL3 loci in *N. tabacum* are noticeably smaller than those in the progenitor *N. tomentosiformis* and the diploid *N. kawakamii*.

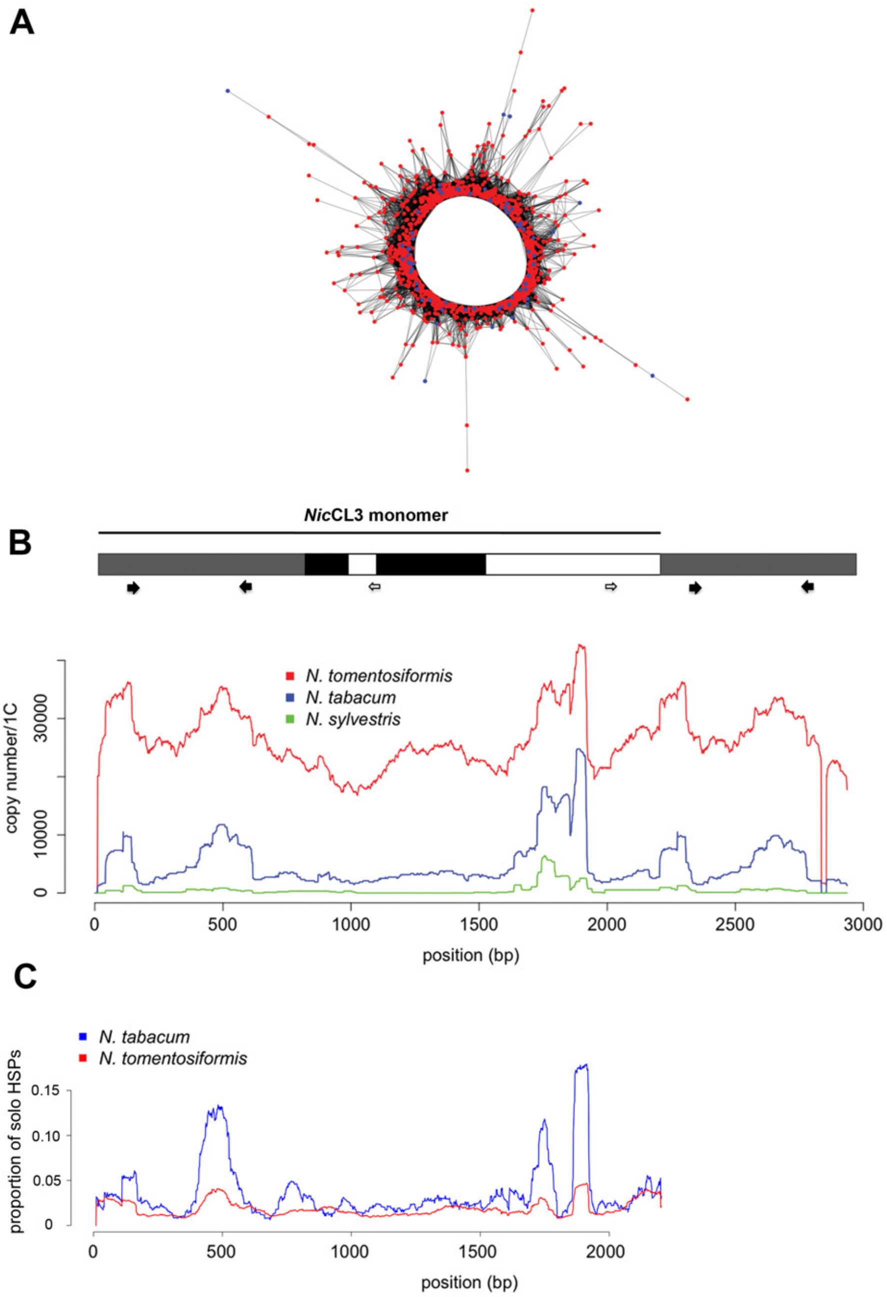

Figure 3.1 Structure and copy number of *Nic*CL3

(A) Graphical 2D projection of a three dimensional network where each node represents a single 454 sequence within *Nic*CL3. 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 originating from *N. tomentosiformis* and blue are reads originating from *N. tabacum*.

(B) A diagrammatic representation of the consensus sequence of the most abundant contig (contig 8) of CL3, here called *Nic*CL3. The line (top) indicates the *Nic*CL3 monomer, the greyed regions represents those regions of the contig that are repeated because it contains part of a second monomer. Copy-number estimates (estimated by 454 read-depth) for allopolyploid *N. tabacum* and the progenitor diploids are shown. The approximate positions of primer set 1 (black arrows) and primer set 2 (open arrows) are shown (see Experimental Procedures). Regions in *Nic*CL3 matching the d and j-locus found flanking a endogenous pararetrovirus (NtoEPRV) described in Matzke *et al*. (2004) are highlighted in black.

(C) Paired-end reads were used to determine the occurrence of dispersed *Nic*CL3 sequence and/or insertion of other sequences within *Nic*CL3. The proportion of solo HSPs (*Nic*CL3 sequences whose paired read does not match *Nic*CL3) is shown mapped along the monomer of *Nic*CL3 contig 8 for *N. tabacum* and *N. tomentosiformis*. Note there are regions along the monomer that are more likely to be associated with sequences other than *Nic*CL3 (solo HSPs) and that the proportion of solo HSPs is considerably higher in *N. tabacum*.

Metaphase chromosomes of several synthetic *N. tabacum* lines (Th37-3, -7 and -14) reveal only two *Nic*CL3 signals, on a single pair of large submetacentric chromosomes (Fig. 3.3 B, D and Table 3-3). 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. 3.3 F, G). The S0 generation synthetic *N. tabacum* TR1-A has eight *Nic*CL3 signals, as expected (Fig. 3.3 J). *Nic*CL7 has a dispersed signal on all *N. tomentosiformis* chromosomes (Fig. 3.3 A), although some regions bind the probe more efficiently producing a band-like pattern on large submetacentric chromosome pairs. This is particularly evident on the 18S rDNA-bearing chromosome. *Nic*CL7 signal is associated with all *Nic*CL3 signals in *N. tomentosiformis*, Th37 and TR1-A. Th37-3 has *Nic*CL7 signal on 24 of the 48 chromosomes (Fig. 3.3 J); it is likely these derive from *N. tomentosiformis*. *Nic*CL3 and *Nic*CL7 signals were not detected in *N. sylvestris*.

Southern blot hybridisation

Southern blot hybridisation was carried out using *Nic*CL3 clone 9 as a probe. For each species 1-2 μg of genomic DNA was digested with *Bam*HI and *Spe*I enzymes (Fig. 3.4, Table 3-3), which have a single restriction site within *Nic*CL3. A ladder pattern of bands was evident in *N. kawakamii*, *N. tomentosiformis* (TW142 and NIC 479/84), natural *N. tabacum* (095-55 and SR1) and synthetic *N. tabacum* (Th37-3, 5, 6, 7 and 8). The bands are indicative of tandemly arranged satellite repeats in head to tail orientation. The fastest migrating band corresponded to the satellite monomer (2.2 kb) contained within the 2.9 kb *in silico* reconstruction (Fig. 3.1 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 3-3).

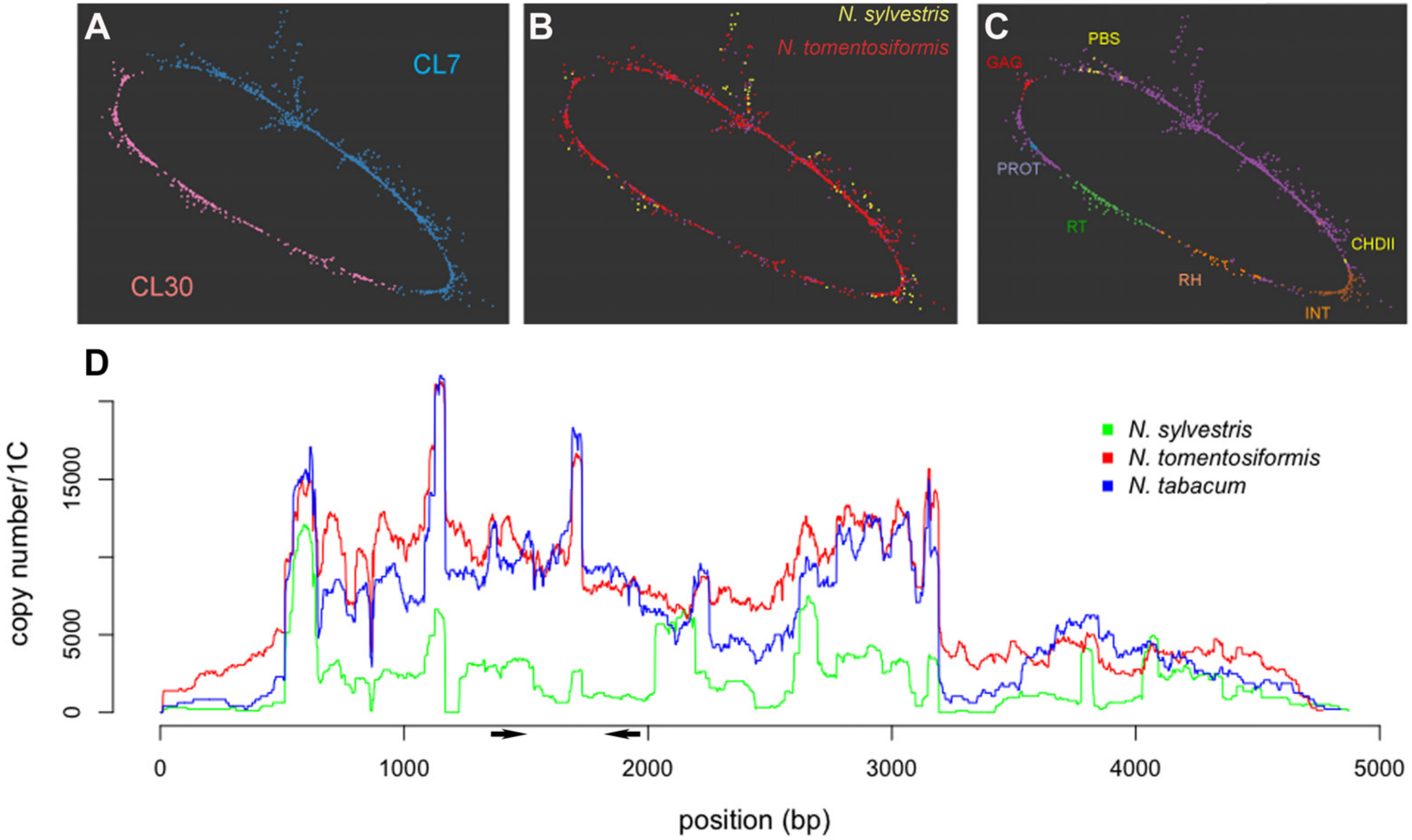

Figure 3.2 The cluster *Nic*CL7/30

(A) Cluster *Nic*CL7/30 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 *Nic*CL7/30 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 *Nic*CL7/30. 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.

In natural *N. tabacum*, Th37 and *N. tomentosiformis* digestion of the *Nic*CL3 unit is inhibited when the methylation sensitive restriction enzyme *Hae*II is used (with one restriction site in the monomer), indicating cytosine methylation of *Nic*CL3 at the restriction site in these species (Fig. 3.4).

The *in silico* consensus of *Nic*CL3 sequence includes terminal repeats (Fig. 3.1 B) and to confirm that these arise because the consensus includes a whole monomer and part of a second monomer in the tandem array, I designed PCR primer pair 2 (open arrows in Fig. 3.1 B). PCR analysis generated a product of ~1400 bp, consistent with a monomer length of 2.2 kb (Figure A.3). Sanger sequencing of a clone of this PCR product confirmed the expected arrangement of a 2.2 kb monomer (Fig. 3.1 B).

Sequence similarity in NicCL3

Comparisons of sequence similarity between *Nic*CL3 derived 454 reads in *N. tomentosiformis*, *N. tabacum* and *N. sylvestris* was used in order to detect evidence for rounds of amplification and/or homogenisation. Reads deriving from *N. tomentosiformis* and *N. tabacum* were analysed separately. In addition, I 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 (Figure A.4). Pair-wise similarity scores for *Nic*CL3 sequences from *N. tabacum* and *N. tomentosiformis* were plotted as frequency distributions and kernel density estimates (Fig. 3.5). This analysis revealed a peak of identical

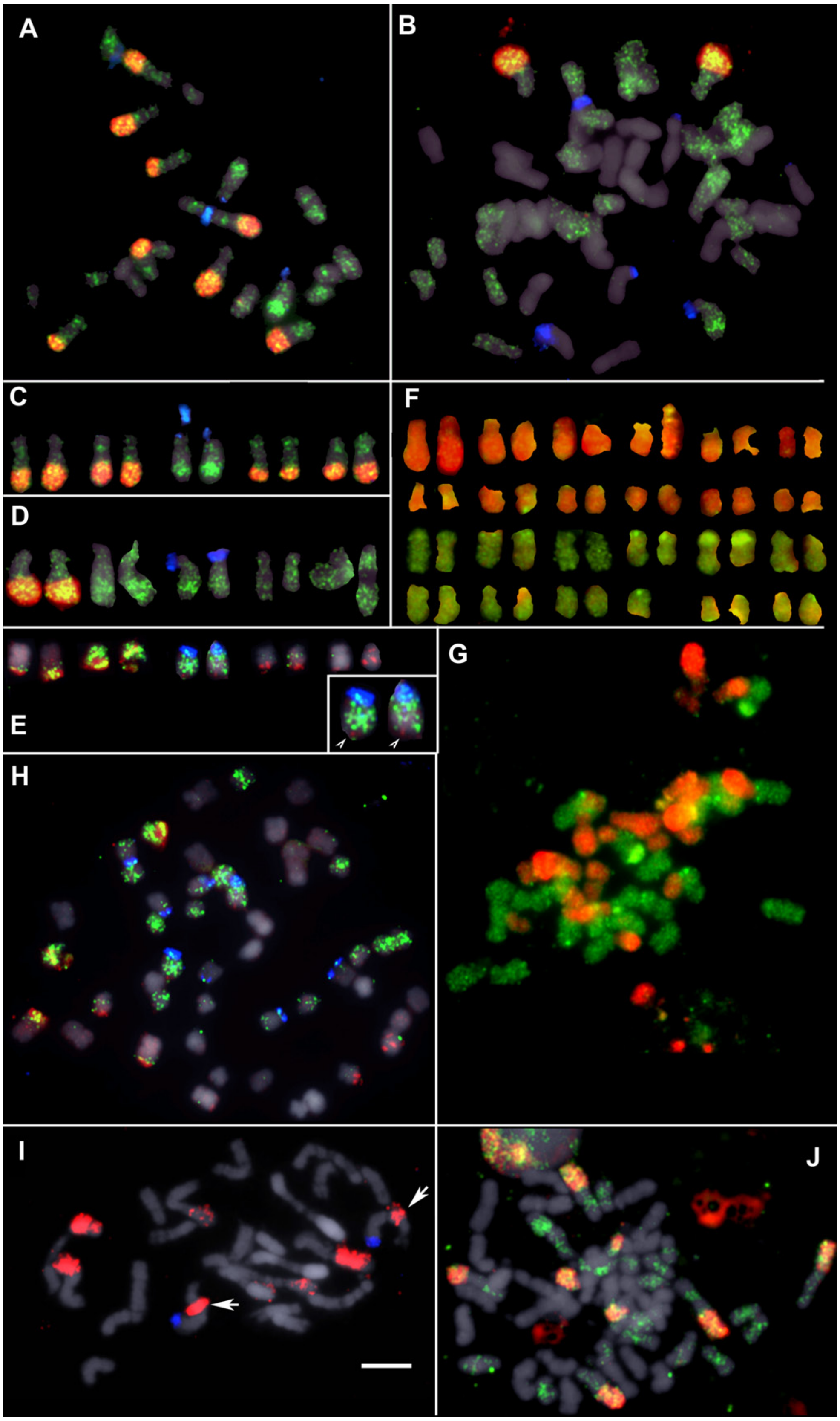

Figure 3.3 FISH of *Nic*CL3 and *Nic*CL7/30

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), *Nic*CL7 (green) and *Nic*CL3 probes (red) counter stained with DAPI (grey). Inset (E) shows enlarged chromosome T3 with *Nic*CL3 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 *Nic*CL3 signal proximal to the centromere (arrows). (J) TR1-A an S0 synthetic *N. tabacum* with the expected number of *Nic*CL3 (red) signals and highly localised *Nic*CL7 (green) signals. Scale bar is 5 μm.

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. 3.6). A two-sample Wilcoxon test revealed a significant difference ($p < 0.00001$) between mean sequence similarity of *Nic*CL3 derived-sequences in *N. tomentosiformis* (0.93) and *N. tabacum* (0.90).

I also examined the proportion of sequence reads from *N. tabacum* or *N. tomentosiformis* matching the consensus (*Nic*CL3, contig 8) for each nucleotide along its length (Fig. 3.6 A), plotting the average proportion of bases identical to the consensus over consecutive 20 bp windows (Fig 3.6 B). The data indicate that a similar proportion of bases match the consensus along the length of the unit in both species, with the exception of a region towards the end of *Nic*CL3, where the reads are more divergent.

Table 3-3 *Nic*CL3 in the genomes of *Nicotiana tabacum*, *N. tomentosiformis* and members of the section *Tomentosae*. The meaning of symbols within the table are as follows: + Indicates a ladder like pattern following restriction digestion and Southern blot analysis, n.s indicates the accession was not screened, – indicates no signal detected. [a]Groups of Th37 plants as described in Skalicka *et al*. (2003).

		Southern hybridisation	number of FISH signals
N. tomentosiformis (TW142)		+	n.s
N. tomentosiformis (NIC 479/84)		+ (Fig. 3.4)	8 (Fig. 3.3 A,C)
N. tomentosiformis (Nee et al. 51771)		n.s	8
N. tabacum (SR1)		+ (Fig. 3.4)	8
N. tabacum (095-55)		+	8 (Fig. 3.3 H,E)
Th37[a]	1	–	n.s
	3	+	2 (Fig. 3.3 B,D)
	5	+ (Fig. 3.4)	n.s
	6	+	n.s
	7	+	2
	8	+	n.s
	9	n.s	–
	14	n.s	2
TR1-A		n.s	8 (Fig. 3.3 J)
N. sylvestris		–	–
N. kawakamii		+	8 (Fig. 3.3 I)

Table 3-3 continued

	Southern hybridisation	number of FISH signals
N. otophora	–	–
N. tomentosa	trace amounts	–
N. setchellii	trace amounts	n.s
N. glutinosa	–	n.s

Discussion

Nic*CL3, a highly 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. I show one of the most abundant repeats in the *N. tomentosiformis* genome, *Nic*CL3 (Table 3-2), is predominantly arranged in tandem (Fig. 3.1 C, Fig. 3.4), has a unit length of ~2.2 kb (Fig. 3.4, Table 3-3) and is highly localised in *N. tomentosiformis*, *N. kawakamii*, several synthetic *N. tabacum* lines and natural *N. tabacum* (Fig. 3.3).

However the sequence is not a typical tandem repeat, such as the *Nicotiana* satellites belonging to the HRS60 family (Koukalova *et al.*, 2010), for the following reasons. (1) Typically tandem repeat monomers in angiosperms are ~180 bp in length (Heslop-Harrison & Schwarzacher, 2011). Even the long monomer pSc250 in *Secale cereale* is only 550 bp (Vershinin *et al.*, 1995) and so a monomer length of 2.2kb is unusually long.

(2) Satellite blocks usually occur in long arrays of similar units. However *Nic*CL3 also includes a proportion of units that are dispersed (c. 3% in *N. tomentosiformis* and 9% in *N. tabacum* and Fig. 3.1 C), some of which are associated with *Nic*CL7/30. In *N. tomentosiformis*, Th37 and *N. tabacum Nic*CL3 digestion is almost entirely inhibited when using a methylation sensitive restriction enzyme (Fig. 3.4).

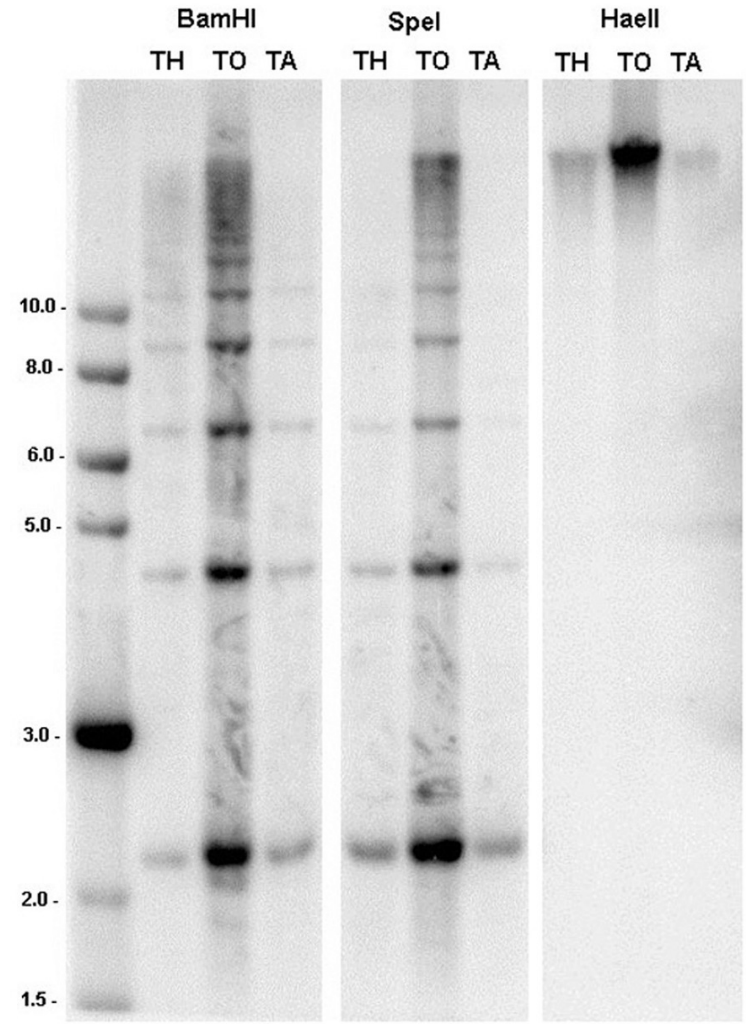

Figure 3.4 Tandem arrangement of *Nic*CL3

Southern blot hybridisation of genomic DNA from (TH) synthetic tobacco Th37-5, (TO) *N. tomentosiformis* ac. NIC 479/84 and (TA) *N. tabacum* ac. SR1 digested with *Spe*I, *Bam*HI and *Hae*II (a methylation sensitive isoschizomer of *Bam*HI) and probed with *Nic*CL3. Size indicators on the left are in kb. Digestion with *Bam*HI and *Spe*I results in a ladder like pattern, typical of tandem repeats. Digestion is inhibited when using *Hae*II, indicating extensive CG methylation of tandem units.

These findings indicate that *Nic*CL3 loci are likely to be heavily methylated. However there were reads derived from *Nic*CL3 in *N. tabacum* GSS sequences (obtained after methylation filtration of genomic DNA), although the number of hits was much lower than expected based on 454 abundance estimates. Since the *Nic*CL3 is highly methylated (Fig. 3.4) it follows that most units were lost by methylation filtration. Still rare hits are interesting since they may originate from euchromatic, potentially transcribed parts of the array.

*Nic*CL3 shares sequence similarity with regions previously found flanking NtoEPRV (endogenous pararetrovirus) insertions (Matzke *et al.*, 2004) (Fig. 3.1 B). There is a complex relationship between NtoEPRV, *Nic*CL3 and Ty3/*Gypsy* sequences, which is not well understood at this stage. However the unusually long tandem sequence (2.2 kb) and similarity to retroelement sequences might indicate that *Nic*CL3 is derived from an ancient Ty3/*Gypsy* retroelement, that now occurs predominantly in tandem array. Similar compound satellites with long monomers that include sections of retroelement sequences have been described in *Solanum tuberosum* (Tek *et al.*, 2005) and *Secale cereale* (Langdon *et al.*, 2000).

Elimination of Nic*CL3 in synthetic and natural* N. tabacum

Next generation sequence (using both Illumina and Roche 454) and FISH analysis have revealed the genome of *N. tabacum* to have a much lower abundance of *Nic*CL3 than expected given its abundance in *N. tomentosiformis*, suggesting large-scale losses (Table 3-2). A reduction in copy number of

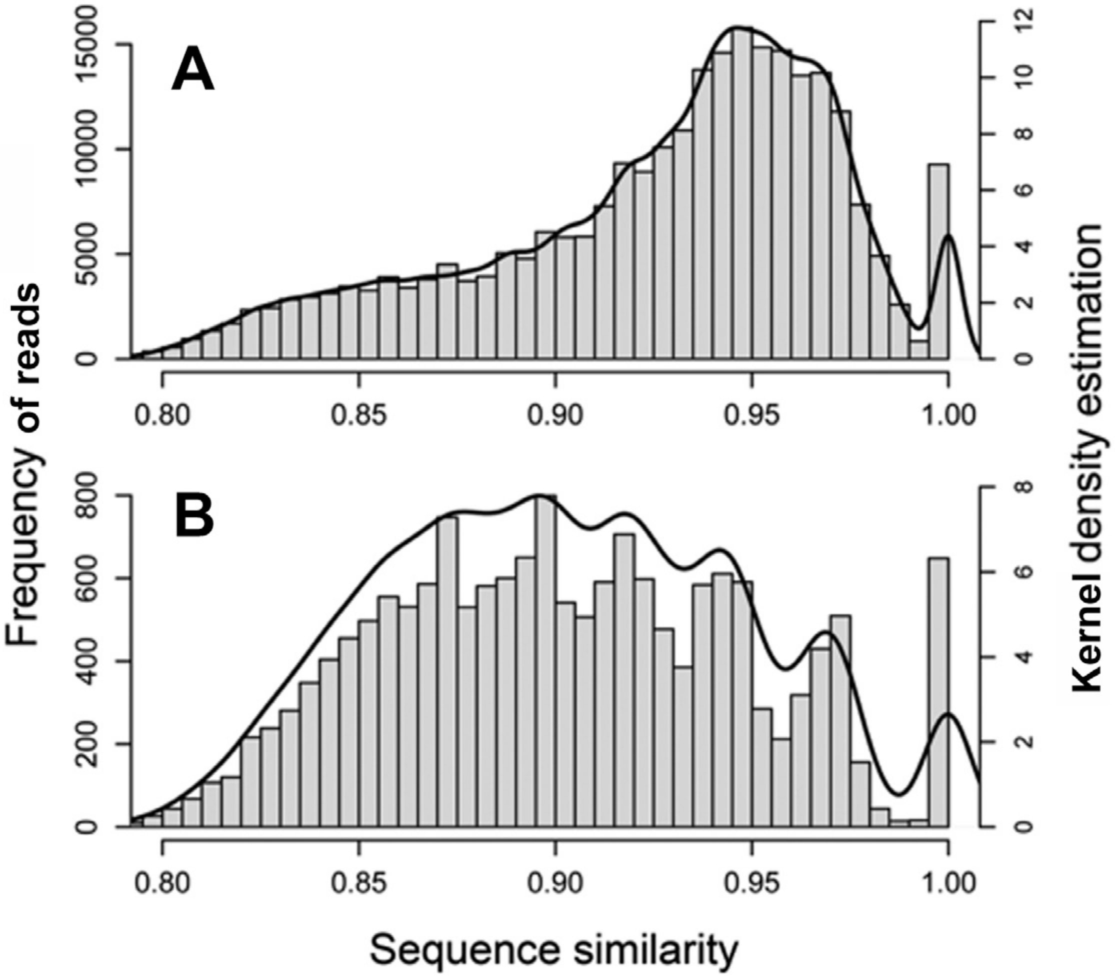

Figure 3.5 Sequence similarity of *Nic*CL3

Histogram of sequence similarity of *Nic*CL3 derived reads in *N. tomentosiformis* (a) and *N. tabacum* (b). Kernel density estimations are also shown (black line). Note that both species have evidence of sequence amplification and/or homogenisation (peak at sequence similarity of one). There are 6 peaks in *N. tabacum*, perhaps indicative of several independent rounds of ancient amplification and/or homogenisation. There are a relatively high proportion of low similarity sequences in *N. tabacum* compared to *N. tomentosiformis.*

*Nic*CL3, amounting to thousands of units has also been observed in fourth generation synthetic *N. tabacum* (Th37). The supposition that *Nic*CL3 has experienced dramatic loss in natural and synthetic lines is evidenced by:

(1) One of the *N. tomentosiformis* accessions analysed here is the closest known diploid relative of the T-genome of *N. tabacum* and the actual paternal progenitor lineage of Th37 (ac. NIC 479/84; reference Murad *et al*. 2002). This accession of *N. tometosiformis* has *Nic*CL3 in high abundance (Table 3-2 and Fig. 3.3).

(2) A total of three *N. tomentosiformis* accessions all show similar and strong *Nic*CL3 hybridisation patterns in either FISH (Nee *et al*. 51771 and NIC 479/84; Fig. 3.3 and Table 3-3) and/or Southern blot hybridization (TW142 and NIC 479/84; Fig. 3.4 and Table 3-3).

(3) *N. kawakamii* and *N. tomentosiformis* are sister taxa in phylogenetic analysis (Kelly *et al*., 2010) and both have strong *Nic*CL3 probe binding in FISH and Southern blot analysis (Fig. 3.3, 3.4 and Table 3-3).

Together (1), (2) and (3) indicate *Nic*CL3 was probably abundant in the common ancestor of *N. kawakamii* and *N. tomentosiformis* as well as the true paternal ancestor of *N. tabacum*. Therefore the discrepancy between the expected GP and observed GP in *N. tabacum*, as well as the loss of *Nic*CL3 loci in Th37, is likely to be due to sequence reduction in the allopolyploids rather than expansion in the progenitor post allopolyploidy.

I have shown that, in synthetic *N. tabacum* lines Th37-3, -7 and -14 the number of large blocks of *Nic*CL3 signal is reduced from eight signals to two (Fig. 3.3 B, D). Several lines of Th37 (3,5,6,7 and 8) show low, but detectable levels of *Nic*CL3 following Southern blot analysis.

It is clear that whole loci carrying many thousands of *Nic*CL3 units have been lost from synthetic lines. In addition two synthetic lines (Th37-1, 9, Table 3-3) lack any detectable *Nic*CL3 signal both in Southern blot and FISH analysis indicating that this sequence has been completely (or near completely) eliminated very rapidly indeed – within the first four generations of selfing. This amounts to the removal of nearly 1% of the Th37 genome in only four generations.

Directional loss of parental sequences has been observed in several synthetic Th37 lines (Skalicka *et al.*, 2005; Petit *et al.*, 2010), as well as in natural *N. tabacum* (Volkov *et al.*, 1999; Petit *et al.*, 2007), where there is a trend for repeats derived from *N. tomentosiformis* to be under-represented (Renny-Byfield *et al.*, 2011). In this paper I have shown that *Nic*CL3 is eliminated or reduced in copy number in synthetic *N. tabacum* lines and is much reduced in copy number in natural *N. tabacum*, suggesting directed mechanisms of removal.

Mechanisms of Nic*CL3 loss*

The loss of *Nic*CL3 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 of Th37 accessions show the expected number of *N. tomentosiformis*-derived chromosomes (Fig. 3.3 and Skalicka *et al*. 2005). Repeats arranged in tandem, rDNA for example, are thought to alter their copy number via unequal crossing-over, although the exact mechanisms are still obscure

(Eickbush & Eickbush, 2007; Ganley & Kobayashi, 2007). Homeologous chromosome pairing has been proposed as a mechanism of sequence and chromosome loss (Jones & Hegarty, 2009), and compelling evidence exists for such chromosomal rearrangements in synthetic *Brassica* hybrids (Gaeta *et al.*, 2007; Szadkowski *et al.*, 2010) and recently formed *Tragopogon* allopolyploids (Kovarik *et al.*, 2005; Lim *et al.*, 2008; Chester *et al.*, 2012). However Salina *et al.* (2004) suggested that changes in copy number of tandem repeats Spelt1 and Spelt52 in synthetic wheat, were not a consequence of intergenomic recombination during meiosis, as wildtype and mutant plants at the *Ph1* locus show similar patterns of copy number change (*Ph1* mutants have increased frequency of homeologous pairing). Similarly in *Nicotiana* there is no evidence for extensive homeologous pairing (Goodspeed, 1954; Lim *et al.*, 2000a; Lim *et al.*, 2004a), 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*. 2004a). Hence, it is possible that recombination between large blocks at homologous and non-homologous loci carrying *Nic*CL3 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 *Nic*CL3 in the allopolyploid (Fig. 3.1 C).

The outcome of such processes would be the generation of chromosomes with either extremely large arrays and/or chromosomes with large deletions. Indeed if small deletions within the unit were responsible for lowering the genome proportion of *Nic*CL3 in *N. tabacum*, then one might expect to see a smear towards smaller molecular weight fragments in Southern blot analysis (Fig. 3.4), 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 *Nic*CL3 in *N. tabacum*.

Recombination mechanisms are thought to be responsible for homogenisation of sequences arranged in tandem and there is evidence this has occurred in *Nic*CL3 (Fig. 3.5). 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, perhaps indicative of more ancient rounds of homogenisation. It is possible that these events coincide with *Nic*CL3 unit loss.

The series of peaks in Fig. 3.5 could be explained by several regions of *Nic*CL3 being more variable than others. 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. 3.6). Hence a hypothesis of repeated rounds of sequence homogenisation seems a better explanation for the pattern in Figure 3.5.

This 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 tandem arrays of *Nic*CL3 in *N. tabacum.* The maintenance of the dispersed units of *Nic*CL3 suggests they are more stable than those in tandem array and I hypothesise that the loss of *Nic*CL3 is most likely the result of multiple unequal recombination events between tandem components of *Nic*CL3.

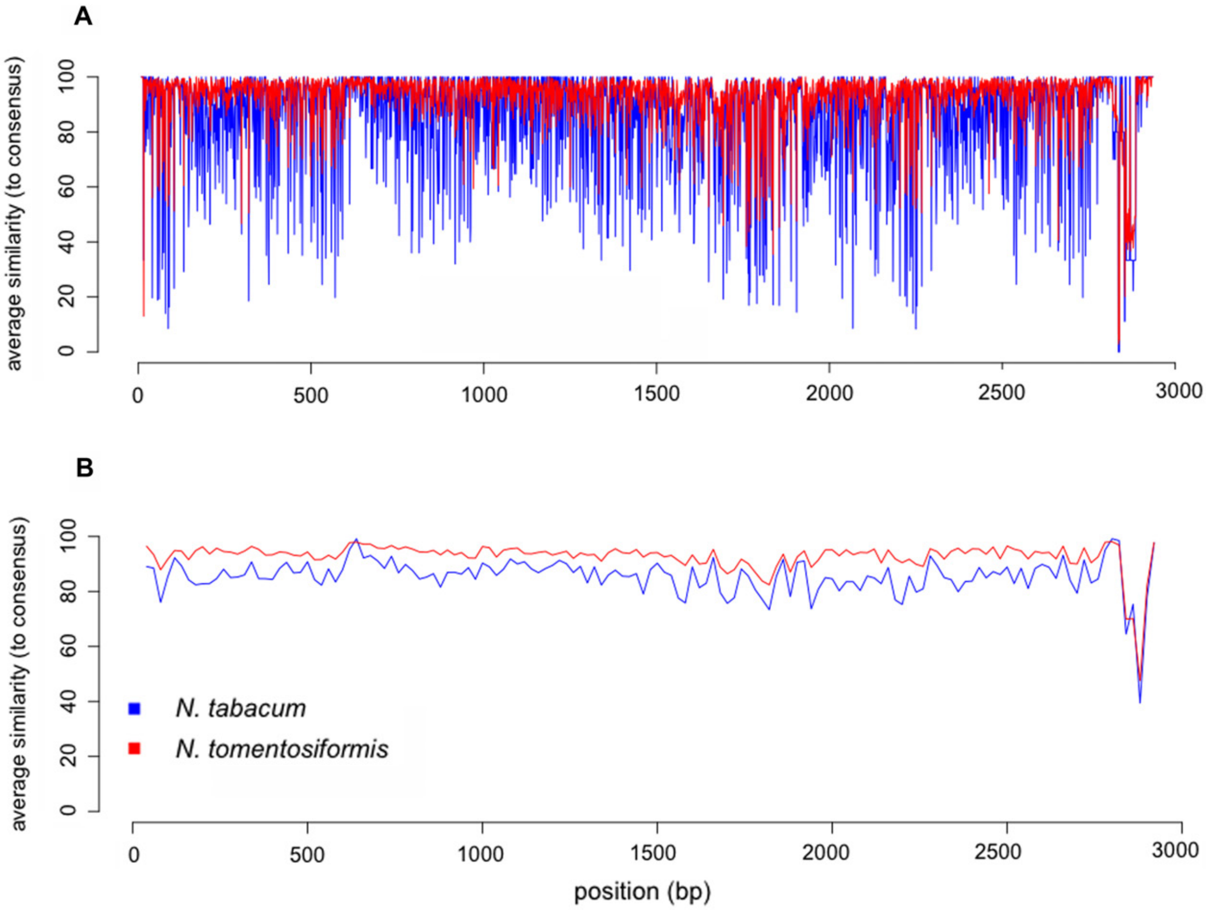

Figure 3.6 Sequence similarity along the *Nic*CL3 unit

Sequence similarity of BLASTn hits to the consensus of *Nic*CL3 (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 20bp windows.

Chapter 4 Diploidisation and genome size change in polyploids is associated with differential dynamics of low and high copy sequences

Summary

Recent advances have highlighted the ubiquity of whole genome duplication (polyploidy) in seed plants and angiosperms, although subsequent genome size change and diploidisation (returning to a diploid-like condition) are poorly understood. *Nicotiana* section *Repandae* arose via a single allopolyploid event (c. 5 mya) involving relatives of extant *N. sylvestris* and *N. obtusifolia*, and provides an excellent model system for dissecting these processes. Speciation in *Repandae* has resulted in allotetraploids with divergent genome size, including *N. repanda* (5320 Mbp/1C) and *N. nudicaulis* (3477 Mbp/1C), which I examine here. Graph-based clustering of next-generation sequence data enabled assessment of the global genome composition of the allotetraploids and their diploid progenitors. Interestingly the majority of sequence clusters (>90%) were at lower abundance than expected in both allotetraploids. Moreover under-representation affected predominantly the low copy-number fraction of both genomes. In *N. nudicualis* this accounted for a 9.8% reduction in genome size compared to expectation. In contrast, *N. repanda* shows expansion of clusters mostly derived from already abundant Ty3/*Gyspy* retroelements, which counteracts loss of lower copy-number repeats. This has led to a 26.0% increase in genome size. These phenomena were associated with the failure to distinguish progenitor genomes by genomic *in situ* hybridization in both allotetraploids. Thus diploidisation processes in these allopolyploids involves erosion of low copy-number nuclear DNA and genome divergence manifests through the lineage specific amplification of Ty3/*Gypsy* retroelements.

Introduction

Most angiosperms have experienced several rounds of whole genome duplication (WGD) in their ancestry (Vision *et al.*, 2000; Bowers *et al.*, 2003; Jaillon *et al.*, 2007; Barker *et al.*, 2009; Jiao *et al.*, 2011). Despite the frequency of WGD the majority of angiosperm species have relatively small genome size (GS; Bennett and Leitch 2010).

Large-scale analyses of GS in angiosperms, reveals a trend towards GS reduction in polyploid lineages (Leitch & Bennett, 2004; Leitch *et al.*, 2008), likely due to elimination of redundant sequences and repetitive DNA (Hawkins *et al.*, 2009; Renny-Byfield *et al.*, 2011; Renny-Byfield *et al.*, 2012). The process of sequence losses and gains, coupled with DNA turnover, is associated with diploidisation (Lim *et al.*, 2007), where the genome of a polyploid returns to a more diploid like state. However the processes and mechanisms governing GS change and diploidisation in allopolyploids are poorly understood, and detailed descriptions of the DNA sequences involved are lacking.

Evolution of repetitive DNA can be analysed via 'genome skimming', in which short read next generation sequencing (NGS) data are used to reconstruct and quantify the repetitive fraction of the genome (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker *et al.*, 2009; Hribova *et al.*, 2010; Macas *et al.*, 2011; Renny-Byfield *et al.*, 2011). By using low coverage datasets the reads are largely composed of sequences in multiple copies, whereas low copy-number sequences are rare. These datasets allow the reconstruction of repeats and sequencing read-depth of a given repeat is typically proportional to its abundance within the genome, allowing quantification and comparisons between species.

Recent studies have used these methods to better understand genome evolution following allopolyploidy. Studies on the young allotetraploid (<0.2 mya; Clarkson *et al*., (2005)) *Nicotiana tabacum* with its diploid progenitors identified a bias towards removal of paternally derived repeats in conjunction with a reduction in the repetitive fraction of the genome (Renny-Byfield, *et al*. 2011). Patterns of sequence loss observed in this naturally occurring allotetraploid are repeated in synthetic lines after only four generations (Renny-Byfield, *et al*. 2012). Similarly, reproducibility of DNA loss has been reported in wheat (Ozkan *et al*. 2001, Salina *et al*. 2004).

Nicotiana section *Repandae* provides an ideal model group for dissecting GS change, particularly in the context of allopolyploidy. A single hybridisation event between ancestors of extant *N. sylvestris* (2636 Mbp/1C) and *N. obtusifolia* (1511 Mbp/1C), followed by speciation, has produced four allopolyploids (Chase *et al*., 2003; Clarkson *et al*., 2004; Clarkson *et al*., 2005; Clarkson *et al*., 2010). Importantly these deviate from their expected GS (sum of GS for the closest extant relatives of the progenitor diploids: 4147 Mbp/1C). For example there is a ca. 29% genome upsizing in *N. repanda* (5320 Mbp/1C) and a ca. 14% genome downsizing in *N. nudicaulis* (3477 Mbp/1C; Leitch *et al*., 2008) compared with expected GS. In addition, previous studies using GISH (Lim *et al*., 2007) have failed to discriminate the two progenitor chromosome complements, although this failure is qualitatively different among allotetraploids of section *Repandae*.

Here I use GISH, ‘genome skimming’ and a graph-based clustering pipeline to identify, quantify and compare the repetitive DNA in these two allopolyploids and relatives of their diploid progenitors. Thus I provide an in-depth analysis of repeat expansions and contractions producing GS change and diploidisation in *N. repanda* and *N. nudicaulis*.

Materials and Methods

Plant material

I used [1] *Nicotiana obtusifolia* (ac. 8947501/176) and [2] *N. nudicaulis* (ac. 964750051) both from the Botanical and Experimental Garden, Radboud, University of Nijmegen, Netherlands. [3] *N. sylvestris* (ac. ITB626) from the Tobacco Institute, Imperial Tobacco Group, Bergerac, France, and [4] *N. repanda* (ac. TW18) from USDA, North Carolina State University, NC, USA.

DNA sequencing

DNA extractions were performed according to Fojtova *et al.* (2003) and DNA integrity was checked using gel electrophoresis. A random sample of DNA from the genomes of *Nicotiana sylvestris*, *N. obtusifolia, N. repanda* and *N. nudicaulis* with sequenced using the Illumina Genome Analyzer xII, at The Genome Centre, Queen Mary University of London, generating read lengths of 108 bp. Raw sequence reads are deposited at the Sequence Read Archive at NCBI under the study accession numbers SRA045794 and SRA051392. Resulting sequence reads were then screened for quality and removed if they contained more than 5 unidentified nucleotides (Ns) or were shorter than 95 bp in length. All sequences passing quality checks were trimmed to 95 bp and screened against plastid genomes and adapter sequence databases; those reads with significant similarity were removed from further analysis.

Clustering and repeat identification

A random sample of 2% of each genome was combined into a single dataset and subjected to a graph-based clustering procedure described in Novak *et al*. (2010). Details of the data used in this analysis are provided in Table 4-1. This approach identifies repetitive DNA families using a *community* approach by grouping high-throughputput sequencing reads into clusters based on shared sequence similarity. Each sequence read was compared to all other reads in a pair-wise analysis using MGBLAST (Altschul *et al*., 1990), requiring at least 90% sequence identity along 55% of the sequence read. Graph-based clustering was performed in the R programming language, detecting sets of reads that are more densely connected among each other than to other reads. These groups are termed *clusters* and correspond to families of repetitive DNA that were characterised further.

Table 4-1 Description of the data used in the clustering analysis

	database size (bp) /number of reads	genome size (bp)	% genome sampled
N. nudicaulis	69,540,000/732,000	3,477,000,000	2
N. repanda	106,400,000/1,120,000	5,320,000,000	2
N. obtusifolia	30,219,975/318,105	1,511,000,000	2
N. sylvestris	52,719,870/554,946	2,636,000,000	2

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

Given that all sequences are the same length it follows that the number of reads in each cluster is a measure of its abundance within the original dataset. Moreover each species sample amounts to 2% of the genome and so a count of the number of sequence reads from each species within a cluster gives a quantitative measure of abundance in the genome of each species. Thus counting the number of reads allows the genome proportion (GP: a percentage of the genome) for each cluster for all four species to be calculated.

After graph-based clustering, reads were assembled using the TGICL (Pertea *et al.,* 2003) version of CAP3 on a cluster by cluster basis, requiring 80% sequence similarity along a 30 bp length. Reads within clusters (consisting of at least 10 reads) were assessed for sequence similarity to a database of known repetitive sequences (RepBase 16.03; Jurka *et al.*, 2005) using RepeatMasker (with the –s option that invokes slower and more sensitive searches). To avoid spurious labelling of clusters only those annotations encompassing at least 10% of the reads, or totalling 100 hits were considered. The number of reads in clusters with the same annotation were summed in order to calculate GP for each repeat type.

Comparing deviation of repeat abundance in the allotetraploids

Analyses of clusters were restricted to those that included ten or more reads derived from the parents. All analyses were performed using custom R, perl and shell scripts, which are available to download at http://webspace .qmul.ac.uk/sbyfield/Simon_Renny-Byfield/Research_Projects.html.

Genomic in situ *hybridisation (GISH)*

Genomic DNA was extracted from fresh leaf material of *N. obtusifolia* and *N. sylvestris* using the Qiagen DNeasy kit according to the manufacturer's instructions. Following extraction 1 µg of genomic DNA was labelled with either biotin-14-dUTP or digoxigenin-11-d-UTP using the Roche Nick translation kit (Roche) according to the manufacturer's instructions.

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

GISH followed the protocol described in Lim *et al.* (2006a). Briefly, probe DNA (~100 ng of each 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)). Approximately 50 µl of the probe mixture was added to each slide and the material denatured with a Dyad slide heating-block (MJ Research) at 70 °C for 2 min. After hybridisation at 37 °C overnight, slides were washed in 20% (v/v) formamide in 0.1x SSC at 42 °C for ten minutes, giving an estimated hybridisation stringency of 85%. Probe hybridisation was detected with 20 µg ml^{-1} FITC conjugated anti-digoxigenin IgG (Roche Biochemicals 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. Subsequently images were processed uniformly with Improvision Openlab® and Adobe Photoshop CS2 software.

Results

Clustering

Graph-based clustering was used to characterise, quantify and compare repetitive DNA sequences in the genomes of the diploid species *N. sylvestris* and *N. obtusifolia*, and their derived allotetraploids *N. repanda* and *N. nudicaulis*. Clustering of 2,725,051 Illumina reads, each 95 bp long (2% coverage for each of the genomes), produced 218,995 clusters, consisting of as few as two reads to 183,322 reads. Such clusters correspond to families of repetitive DNA and can be assessed for their abundance in each genome as well as similarity to known repeats. Examples of the resulting 3D networks are shown in Fig. 4.1 and a more extensive set is available at http://webspace.qmul.ac.uk/ sbyfield/Simon_Renny-Byfield/Data.html.

Genome Characterisation

To investigate global genome composition in *N. sylvestris*, *N. obtusifolia*, *N. repanda* and *N. nudicaulis*, repeat clusters were annotated using similarity to known repetitive DNA (using a RepBase library and RepeatMasker). The majority of repeat clusters had no similarity to known repeats (Table 4-2). Of those clusters that could be annotated, the majority were retroelements, contributing between 15.1% and 33.8% of the genome depending on the species (Table 4-2). The majority of LTR retroelements (between 13.1% and 31.2% of the genome) were Ty3/*Gypsy*-like, with Ty1/*Copia*-like elements contribute a smaller fraction of the genome, between 1.6% and 2.6% (Table 4-2). These figures are in broad agreement with values reported in chapter two. The smallest genome analysed (*N. obtusifolia*) contained the lowest proportion of retroelements, while *N. repanda* (the largest genome) contained the highest.

Although the repetitive fraction of all four *Nicotiana* genomes is dominated by retroelements, there are low levels of several other repeat types. DNA transposons are estimated to contribute between 0.3% and 0.8% of the genome, with *N. sylvestris* having the smallest genomic fraction and *N. obtusifolia* the largest. I identified a number of SINEs, LINEs, low complexity and satellite repeat families in the dataset, but these were in low abundance in all four genomes (Table 4-2).

Comparing observed with expected values in the allotetraploids

The expected abundance of repeat clusters was compared with those observed in the two allotetraploids. For 3052 clusters (where the expected number of reads was ≥10), 2779 and 2762 were found to have fewer reads than expected in *N. repanda* and *N. nudicaulis* respectively. Many clusters (2663) were under-represented in both allotetraploids. For 1446 of these under-representation was greater in *N. repanda*, whereas only 326 were most under-represented in *N. nudicaulis*. The remaining 891 clusters were equally under-represented in both allotetraploids.

In contrast 266 and 264 clusters were over-represented in *N. repanda* and *N. nudicaulis* respectively. Indeed across *all* clusters *N. repanda* had a higher than expected repeat abundance, accounting for a 26.0% increase in expected GS (sum of the progenitors; 4147 Mbp/1C). However, in *N. nudicaulis* all clusters had a combined abundance lower than expected, revealing, overall, that repeats were under-represented in this species. The sum of deviation from expectation accounted for a 9.8% decrease in the expected GS of *N. nudicaulis*.

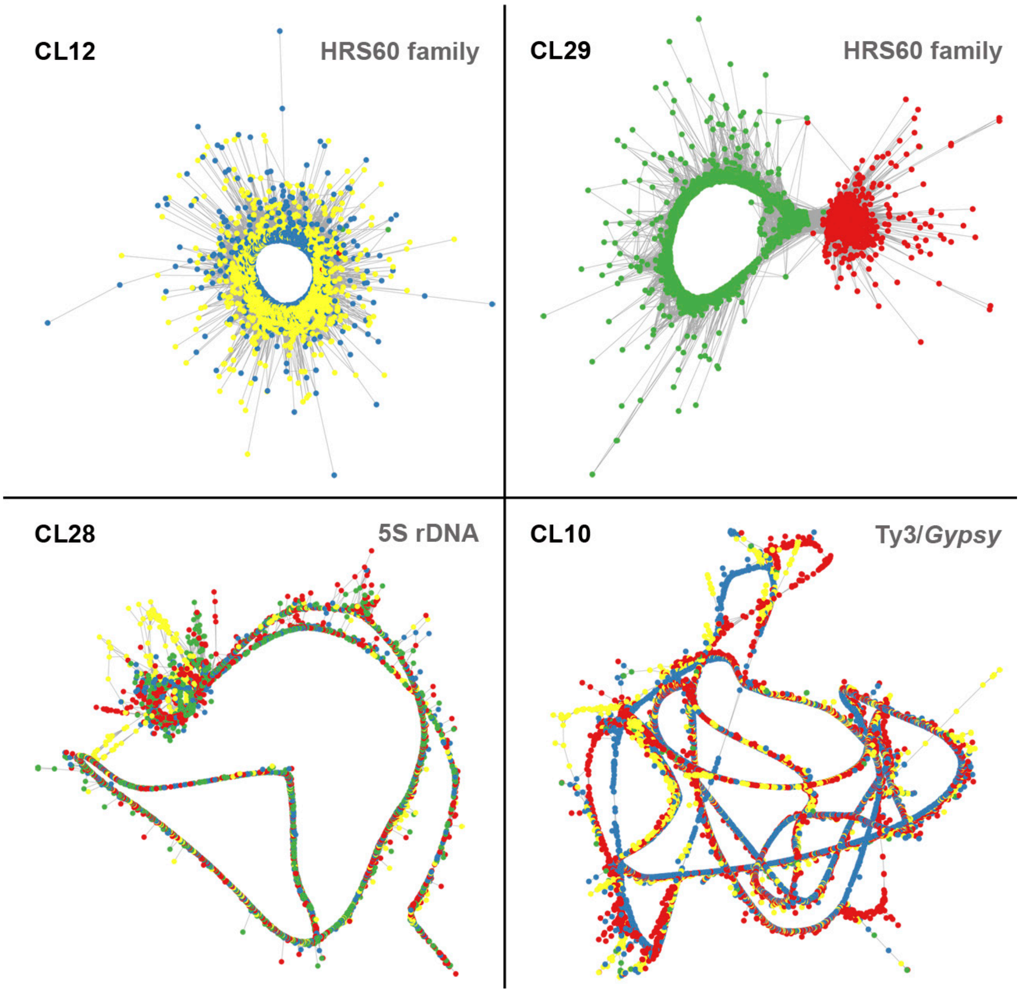

Figure 4.1 Three dimensional graph networks in *Repandae*

Two-dimensional projections of 3D networks representing families of repetitive DNA within the genomes of four *Nicotiana* species. Each node corresponds to a single Illumina read, where the position of each node relative to others is calculated using a Fruchterman-Reingold algorithm and proximity is based on sequence similarity between reads. Nodes are colour coded according to the species from which they were derived, *N. obtusifolia* (yellow), *N. sylvestris* (green), *N. nudicaulis* (blue) and *N. repanda* (red). A RepBase annotation for each cluster is indicated in grey.

In order to determine which clusters are associated with genome upsizing in *N. repanda* and genome downsizing in *N. nudicaulis* I plotted the cumulative deviation from expectation in cluster size in the two allopolyploids (Fig. 4.2 A). This revealed that clusters at low abundance (<100 reads from the progenitors) are under-represented in both *N. repanda* and *N. nudicaulis* (Fig. 4.2 A, from the origin until position i), where both allopolyploids follow a similar pattern. Thereafter, for clusters with higher expected values, the trend in the allotetraploids follow different trajectories, leading to an over-representation of repeats in *N. repanda*, whilst in *N. nudicaulis* repeat copy-numbers remain close to expectation. Figure 4.2 A also reveals that a cluster with an expected value of 616 reads is greatly over-represented in the polyploids (23,182 and 15,549 reads in *N. repanda* and *N. nudicaulis* respectively; Fig. 4.2 A ii).

In order to remove the effect of cluster size the data were re-analysed to give cumulative proportional deviation scores. This score considers the observed number of reads divided by the expected number of reads for each cluster (see also Fig. 4.2 B legend). A negative score indicates cluster size is smaller than expected, while a positive score indicates larger cluster size. Plotting this score, in a cumulative manner from the smallest clusters (by expected size) to the largest, indicates for small clusters there are similar proportional losses in both allopolyploids. However there are exceptions, a few of these low copy-number clusters are over-represented in both species (asterisks in Fig. 4.2 B). For larger clusters only *N. repanda* shows evidence of increases in cluster size.

To assess the impact of cluster abundance on the genome size discrepancy between *N. repanda* (5320 Mbp/1C) and *N. nudicaulis* (3477 Mbp/1C) I compared the abundance of each cluster in the allopolyploids (Fig. 4.2 C). Most clusters have minimal or no effect on genome size differentiation between *N. repanda* and *N. nudicaulis*. However a few clusters, inherited in high copy-numbers, have a marked effect. These clusters account for the majority of the ~29% genome size upsizing reported for *N. repanda*.

To compare the abundance of each cluster in the allotetraploid species relative to expectation (the sum of the parents; Fig. 4.3 A, B) I used heatmap analysis, implemented in R. When considering the 200 clusters expected to have the highest number of reads in the allopolyploids the majority of clusters are under-represented (Fig 4.3 A). However a minority of clusters (blue Fig. 4.3 A) are over-represented. In many cases over-representation is apparent in both allotetrapoids, although more pronounced in *N. repanda*. For the 3052 clusters with expected values higher than ten (this includes those clusters in Fig. 4.3 A), the majority are under-represented in the allopolyploids, with only a few clusters occurring with values close to expectation (Fig 4.3 B). As in Fig. 4.3 A, some clusters show evidence of being over-represented, and again these clusters are often the same in both allopolyploids. Furthermore for those clusters that are over-represented deviation from expectation is often greater in *N. repanda* compared to *N. nudicaulis*. The dendrograms in both Fig. 4.3 A and B group species based on the similarity of cluster abundance. However the two allotetraploids differ, *N. nudicaulis* is closer to 'expected' and both progenitors whereas *N. repanda* is more distant and is an 'outlier'.

Table 4-2 Genome characterisation in *Repandae*

The major repetitive DNA component of the genome in two allopolyploids of section *Repandae* (*N. repanda* and *N. nudicaulis*) and close relatives of their diploid progenitors (*N. sylvestris* and *N. obtusifolia*). GR = genome representation (number of bp attributed to each repeat), GP = genome proportion. [a]Unclassified cluster where similarity searches return no matches. This category includes clusters that have a minimum sum total of ten reads derived from the progenitors. [b] As in (a) but including smaller clusters. * indicates values that are above a threshold of deviation from expectation according to Kenan-Eichler *et al*. (2011).

	N. sylvestris		*N. obtusifolia*		expected		*N. repanda*		*N. nudicaulis*	
	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)
Retroelements	11531575	21.87	4561330	15.09	16092905	19.40	36002055*	33.84	18004305	25.89
LTR/*Gypsy*	10391860	19.71	3952570	13.08	14344430	17.29	33212000*	31.21	15691245	22.56
LTR/*Copia*	918460	1.74	531620	1.76	1450080	1.75	1713325	1.61	1810035*	2.60
LTR/*Caulimovirus*	5605	0.01	15960	0.05	21565	0.03	9975*	0.01	12635*	0.02
LINE/L1	95	<0.01	2565	0.01	2660	<0.01	190*	<0.01	665*	<0.01
LINE/Penelope	51775	0.10	17005	0.06	68780	0.08	54910*	0.05	44840	0.06
LINE/RTE-BovB	160265	0.30	34675	0.11	194940	0.24	1008235*	0.95	439375*	0.63
SINE/tRNA	3515	0.01	6935	0.03	10450	0.01	3420*	<0.01	5510*	0.01
DNA transposons	134995	0.26	247285	0.82	382280	0.46	633555	0.60	521740*	0.75
DNA/CMC-EnSpm	30400	0.06	145825	0.48	176225	0.21	234555	0.22	240445*	0.35

Table 4-2 continued

	N. sylvestris		*N. obtusifolia*		expected		*N. repanda*		*N. nudicaulis*	
	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)	GR	GP(%)
DNA/hAT-Ac	91105	0.17	95095	0.31	186200	0.22	390355*	0.37	272175*	0.39
DNA/hAT-Tip100	1045	<0.01	0	0	1045	<0.01	285*	<0.01	285*	<0.01
DNA/MULE-MuDR	2090	<0.01	190	<0.01	2280	<0.01	2185	<0.01	2375	<0.01
DNA/TcMar-Pogo	0	0	1330	<0.01	1330	<0.01	1330	<0.01	760*	<0.01
DNA/TcMar-Stowaway	9975	0.02	3705	0.01	13680	0.02	4655*	<0.01	5320*	0.01
Other										
RC/Helitron	380	<0.01	1140	<0.01	1520	<0.01	190*	<0.01	380*	<0.01
rRNA	715255	1.36	197790	0.65	913045	1.10	641250*	0.60	452580*	0.65
Low complexity	248615	0.47	65360	0.22	313975	0.38	2271070*	2.13	356250	0.51
Satellite	360240	0.68	621775	2.06	982015	1.18	2609840*	2.45	2290355*	3.30
Simple repeat	3895	0.01	22325	0.07	26220	0.03	17385*	0.02	15675	0.02
Unknown[a]	9958090	18.89	3423325	11.33	13381415	16.13	12914300	12.14	8264620	11.88
small clusters [b]	13300190	25.23	8451675	27.97	21751865	26.23	20328575	19.11	15821965*	22.75
singletons	16467015	31.24	12629135	41.79	29096280	35.08	30981970	29.12	23812510	34.24
total	52719870	100.00	30221140	100.00	82941520	100.00	106400190	100.00	69540380	100.00

The repeats causing GS change

For both allopolyploids I summed the abundance (number of bp) of each repeat type (as identified by RepeatMasker, see Genome Characterisation section above and Table 4-2) and compared this to progenitor additivity. I show the deviation as a percentage of the expected GS (sum of the progenitors, 4147 Mbp/1C; Fig. 4.4). For the majority of repeat types deviation from additivity is minimal e.g. DNA transposons and rDNA sequences. For both *N. repanda* and *N. nudicaulis* repeats of unknown origin (no similarity to know repetitive sequences) are under-represented and account for ~0.6% and ~6% reduction in GS respectively.

However in *N. repanda* there is an over-abundance of Ty3/*Gypsy*-like retroelements compared to parental additivity, accounting for around 22.7% increase in GS. The Ty1/*Copia*-like retroelements, LINEs (RTE-BovB), low complexity and satellite repeats have also made a positive contribution to GS expansion in *N. repanda*, although to a lesser degree. In contrast for *N. nudicaulis* there is a much less pronounced increase in Ty3/*Gypsy*-like retroelements (~1.6% increase in GS) and low complexity sequences.

GISH

GISH to *N. nudicaulis* (Fig. 4.5 A) reveals many sites of probe hybridisation, including *N. obtusifolia* probe signal at sub-telomeric regions (red, arrowed in Fig. 4.5 A) and more uniform binding of *N. sylvestris* probe (green in Fig. 4.5 A). However discrimination of probes is difficult and it is not possible to resolve progenitor chromosome sets with any degree of certainty, although a few chromosomes are distinguishable as red or green.

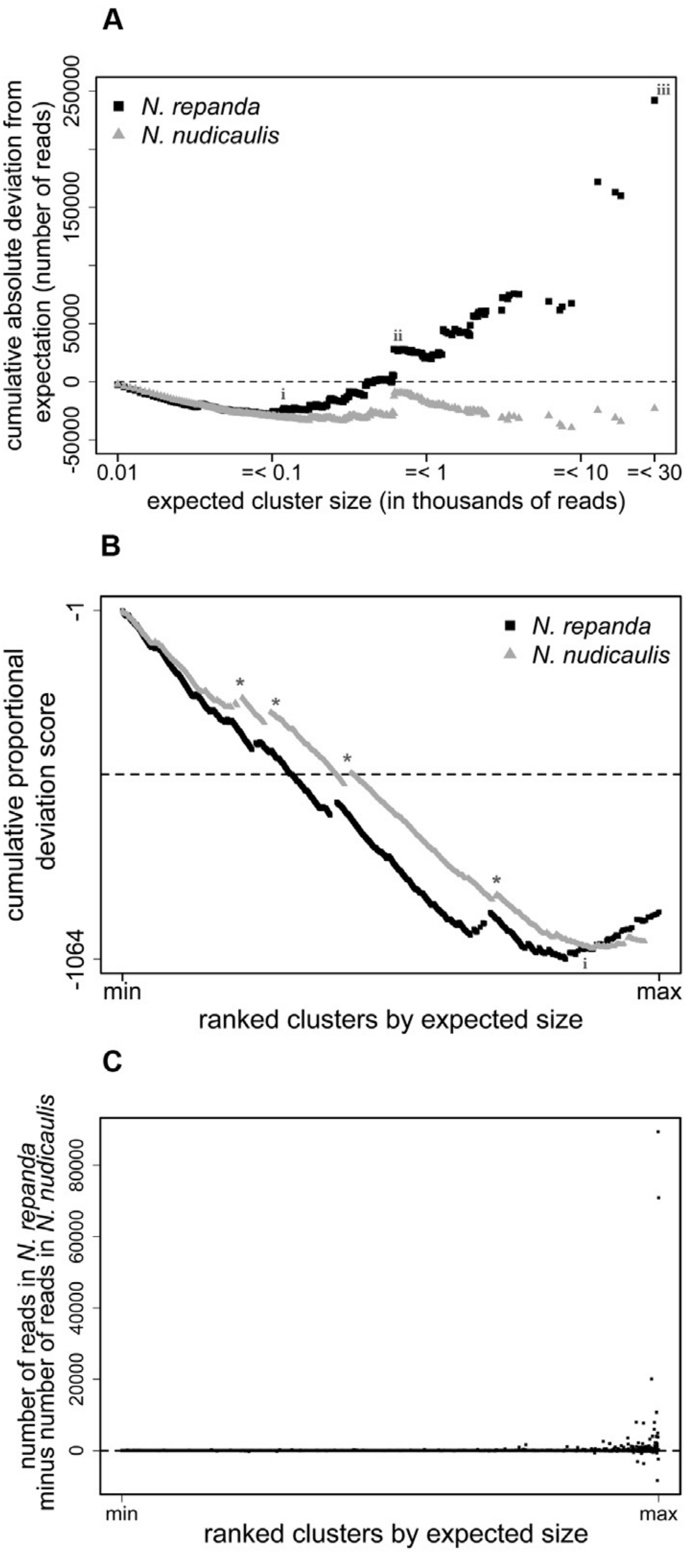

Figure 4.2 Deviation from expectation of clusters in *Repandae*

(A) Graph showing how clusters in the two allopolyploids deviate from their expected size (cumulative change in expectation over the range of cluster sizes). [i] Small clusters (up to an expected size of 100 reads) are under-represented in both *N. nudicaulis* and *N. repanda*. Note that between (i) and (iii) the trajectory of each allopolyploid differs. At (ii) a cluster is over-represented in both allopolyploids, resulting in a step change in cumulative deviation from expectation. (B) Plot showing the cumulative proportional deviation score along the range of cluster sizes. This score reflects proportional change in cluster size from expectation (calculated as observed/expected rotated through 45 º along a Cartesian plane). This score removes any effects of cluster size and is rotated through 45 º so a negative score indicates a reduction in cluster size relative to expectation. A negative slope reflects a trend of reduction in cluster size and a positive slope the reverse. Sections indicated by * show a shared expansion of clusters in both allopolyploids. Note at (i) the gradient of the slope becomes positive in *N. repanda*. (C) Scatter plot indicating the difference in read numbers between *N. repanda* and *N. nudicaulis* within each cluster. Clusters are ranked by size. A data point below zero indicates higher abundance in *N. nudicaulis*, whereas a data point above zero indicate a higher abundance in *N. repanda*. Note the x-axis in (A) is a log scale of expected cluster size whilst in (B) and (C) it is clusters ranked by size.

In *N. repanda* GISH using the same probes produced weak binding of the *N. sylvestris* probe along most chromosomes. Nevertheless there was stronger signal at sub-telomeric regions (arrowed, green in Fig. 4.5 B). However probe binding was weak compared to *N. nudicaulis,* particularly for the *N. obtusifolia* probe where little or no signal was detected. As in *N. nudicaulis*, it was not possible to resolve progenitor chromosome sets.

Discussion

Nicotiana section *Repandae* is an ideal model system for studying genome divergence subsequent to allopolyploidy. This section is thought to have formed from a single allopolyploid event ca. 5 mya (Clarkson *et al.*, 2005; Clarkson *et al.*, 2010). Therefore variation in GS within *Repandae* is likely to have arisen subsequent to allopolyploidy. Indeed previous studies have shown a 33% difference in GS between two allotetraploids (*N. repanda* and *N. nudicaulis*) in section *Repandae* (Leitch *et al.,* 2008). This difference is likely due to genome upsizing in the lineage leading to *N. repanda* (GS is 5320 Mbp/1C, i.e. larger than the sum of the progenitors, 4147Mbp/1C) and genome downsizing in the lineage leading to *N. nudicaulis* (GS is 3477 Mbp/1C, smaller than the sum of the progenitors). Here I suggest why there are such differences in GS.

Low to middle copy-number repeats

Most angiosperms fall within a narrow range of GS (50% of species have a genome size less than 2500 Mbp/1C (Leitch & Leitch, 2012), despite multiple WGDs in their ancestry (Jiao *et al.,* 2011). Indeed global analyses of GS in angiosperms indicate that polyploid genomes tend to decrease in size subsequent to formation (Leitch & Bennett, 2004). For example, genome downsizing has been proposed in allotetraploid *N. tabacum* (Leitch & Bennett, 2004), and the mechanism leading to this loss could have acted rapidly, since some repeats are lost even in synthetics just a few generations old (Skalicka *et al.*, 2003; Renny-Byfield *et al.*, 2012).

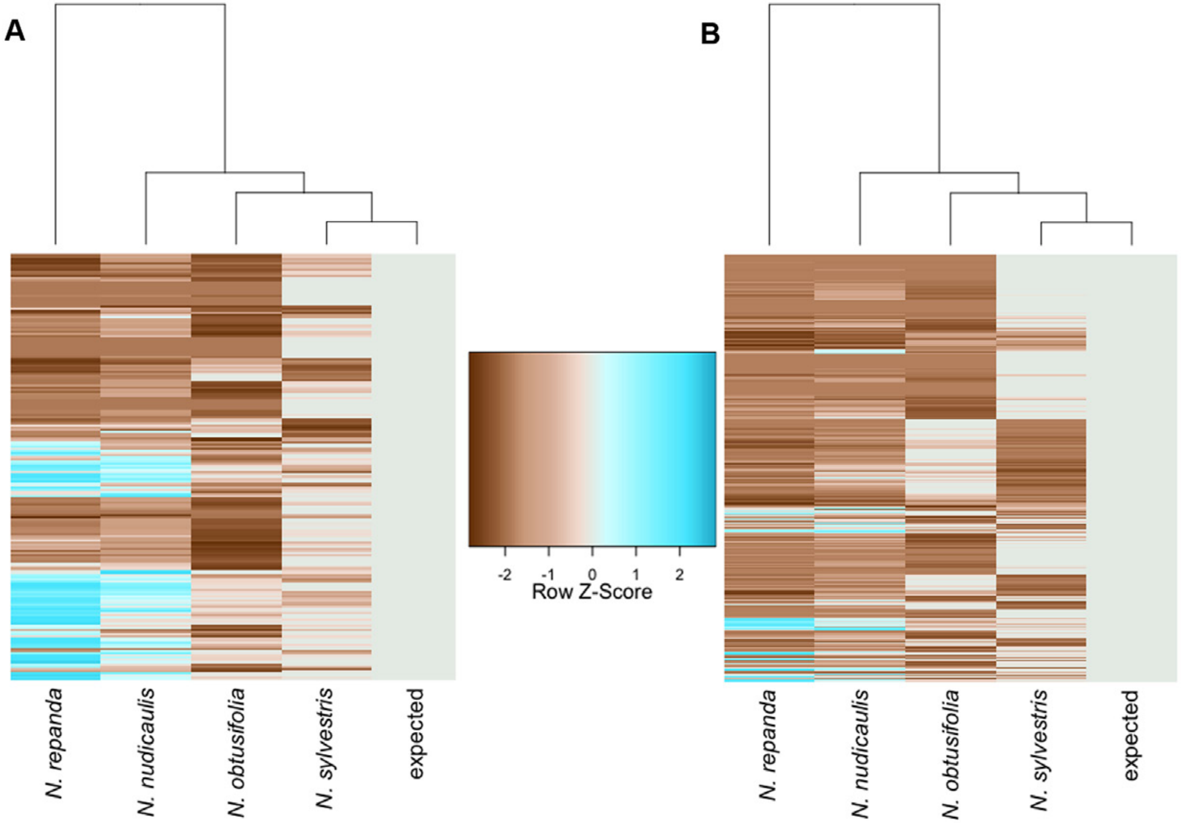

Figure 4.3 Heatmap analysis showing normalised deviation from expectation

Heatmap analysis showing deviation from expectation for the 200 (A) and 3052 (B) clusters with the highest expected values (progenitor additivity). Deviation is normalised across clusters and represented by a Z-score (the number of standard deviations of the mean of all species away from expected). The majority of clusters in both allopolyploids are under-represented (brown), and only a few are overrepresented, particularly in *N. repanda* (blue). Species are grouped by dendrograms (upper side of each panel).

I analysed differences in repetitive DNA content between the two allotetraploids and compared them with the extant diploids (*N. sylvestris* and *N. obtusifolia*) most closely related to the true progenitors. Although there is a substantial difference in GS between *N. repanda* and *N. nudicaulis*, an analysis of low abundance repeats (Fig. 4.2 A, B) shows that they are predominantly under-represented in both species. Repeat reduction in the allopolyploids is also evident in the heatmap analysis, where the majority of repeat clusters (>90%) are under-represented (Fig. 4.3, brown). The under-representation of these low abundance repeats could be a product of; (1) losses in the common allopolyploid ancestor prior to species divergence; (2) common mechanism of loss in the two allopolyploids subsequent to speciation, or (3) increased numbers of low copy repeats in the progenitors subsequent to the formation of the allotetraploid lineage.

The mechanisms resulting in DNA loss are poorly understood, although various recombination-based processes that generate deletions have been proposed (reviewed in Grover and Wendel 2010 and Kejnovsky *et al.*, 2009). For example there is a ca. 80 Mbp difference in GS between *Arabidopsis thaliana* and *A. lyrata*, thought to be the result of increased rates of deletion in the former. These deletions were often small, but numerous and common in non-coding and repetitive regions, including within TEs (Hu *et al.,* 2011). In addition, unequal intra-strand homologous recombination and illegitimate recombination have been identified as mechanisms that remove transposable element (TE) insertions and thus contribute to GS reduction (Devos *et al.*, 2002; Kellogg & Bennetzen, 2004). Moreover in rice the removal of TEs has accounts for the loss of ~190 Mbp of DNA, equivalent to a 38% change in GS over five million years (Ma *et al.*, 2004). It is possible that low abundance repeats are being lost in section *Repandae* via similar mechanisms.

Our clustering analysis has accounted for 9.8% of the 14% GS reduction inferred in *N. nudicaulis.* Repeats that fail to resolve into clusters (perhaps due to insufficient frequency) may also contribute to genome size reduction.

Repeat expansion

Despite an overall reduction in low copy-number repeats in both allotetraploids, there is evidence of substantial expansion of a small number of repeats, including those in low and high copy-number. Indeed the over-representation of these repeats accounts for a 26.0% increase in GS in *N. repanda* (Fig. 4.2 A B and C), almost all of the 29% GS change predicted by Leitch *et al.* (2008). Furthermore these are likely inherited from the diploid progenitors in high copy-number (Fig. 4.2 A, B and C).

There are examples of a few low copy-number repeats that are overrepresented in both *N. nudicaulis* and *N. repanda*, perhaps suggesting amplification in the common ancestor of the allopolyploids or reduction in the progenitors. In contrast *N. nudicaulis* does not exhibit extensive over-representation of high copy-number repeats, as seen in *N. repanda*. Transposable elements (TEs) are often major contributors to angiosperm genomes (Kumar & Bennetzen, 1999), and Ty3/*Gypsy*-like retroelements are particularly prevalent (Macas *et al.*, 2007; Macas *et al.*, 2011) as they are in all four species examined here (Table 4-2; and see chapter 2).

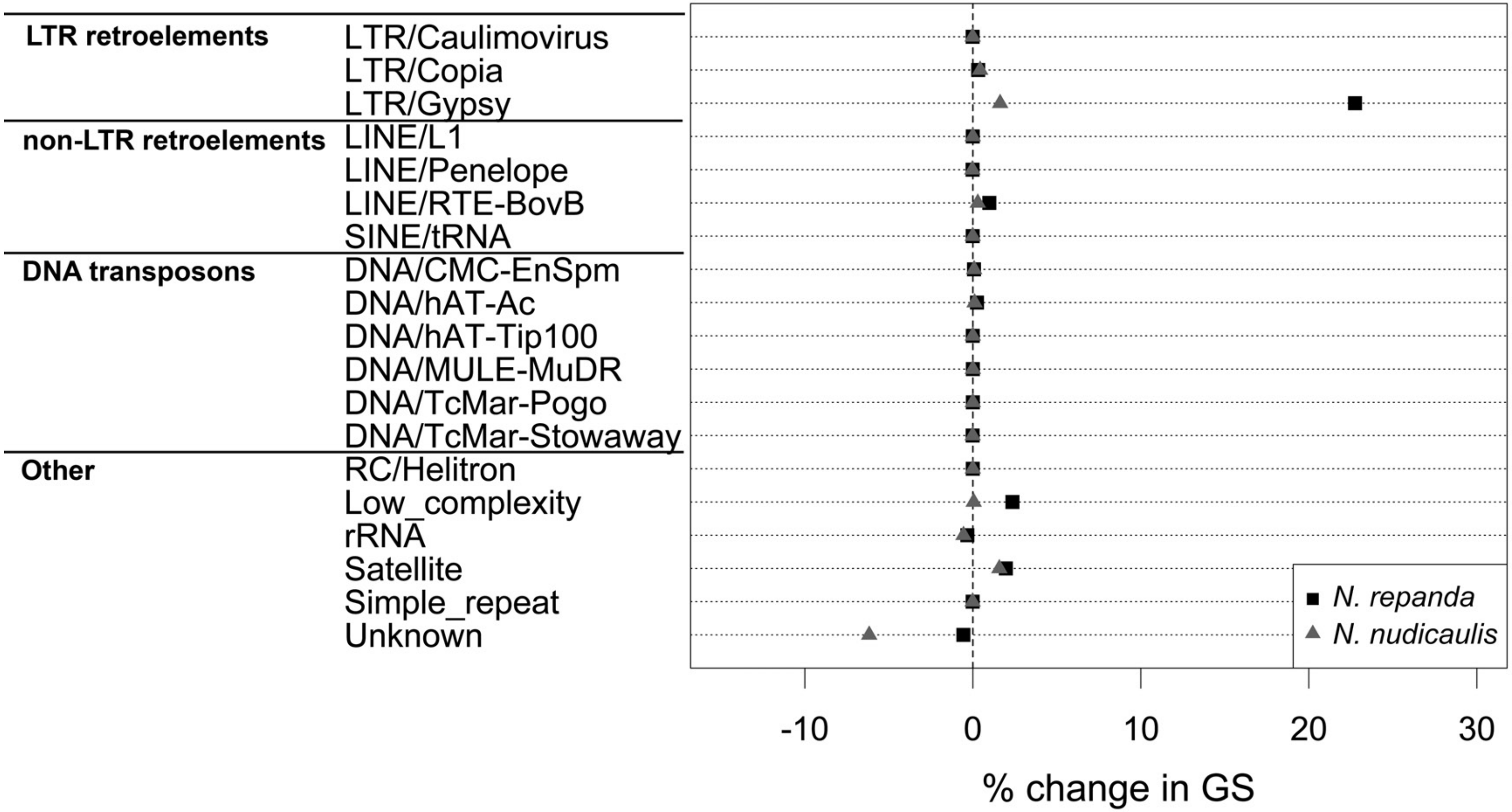

Figure 4.4 Dot-chart showing the contribution of each repeat type to GS change in the allopolyploids *N. nudicualis* and *N. repanda*.

Repeat types are identified by RepeatMasker and the corresponding abundance (number of bp) of each type is compared to the sum of the progenitor diploids. Any deviation from expectation is indicated by a percentage change over the expected GS of 4147 Mbp/1C. The vertical dashed line indicates zero deviation.

The expansion of Ty3/*Gypsy*-like retroelements in *N. repanda* has likely caused a 23% increase in GS (Fig. 4.4). Similarly, there is a three-fold variation in GS observed in diploid *Gossypium,* due to the accumulation of Ty3/*Gyspy*-like *Gorge3* TEs (Hawkins *et al.,* 2006). Furthermore in allopolyploids activation and integration of TEs can occur after only a few generations (Petit *et al.,* 2010). Together these observations suggest that TE dynamics play an important role in governing GS in both diploids and allopolyploids.

Genome Turnover

The dual processes of DNA loss and gain can lead to "genome turnover" (Lim *et al.*, 2007) and analyses of retroelements (Ramakrishna *et al.*, 2002; Ma *et al.*, 2004; Bennetzen, 2005) and nuclear integrants from the plastid genome (Matsuo *et al.*, 2005) have suggested half-lives of only one to a few million years. I show here that the loss of low copy-number, and the expansion of high copy-number repeats, contributes to genome turnover and diploidisation processes. This results in the loss of GISH signal previously reported in *N. nesophila* section *Repandae* (Clarkson *et al.*, 2005; Lim *et al.*, 2007) and shown here, particularly in *N. repanda* (Fig. 4.5). Since GISH is more effective in *N. nudicaulis* one might suggest that the gain of high copy-number repeats in *N. repanda* is more antoganistic to a successful GISH experiment. It is also noteworthy that in heatmap analyses the change in repeat numbers has resulted in *N. repanda* having the most divergent genome, whereas *N. nudicaulis* is closer to the expected value. Moreover analysis of graph networks revealed, in many cases, allotetraploid-specific regions associated with individual clusters (Fig. 4.1). These regions may indicate the evolution of novel variants arising in the common ancestor of section *Repandae*. These observations correspond well with the efficacy of GISH in these two allotetraploids (Fig. 4.5).

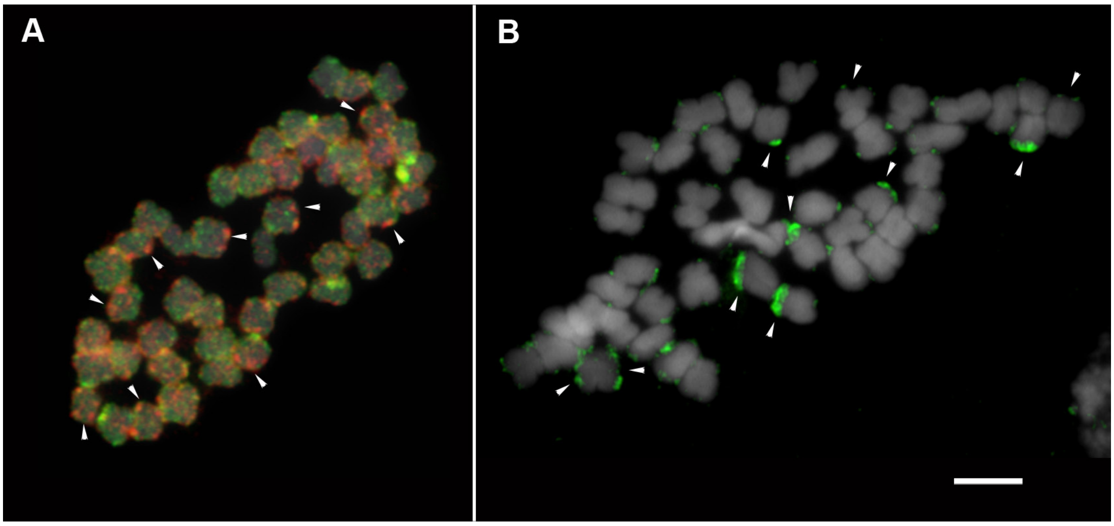

Figure 4.5 Genomic *in situ* hybridisation to *Repandae*

Genomic *in situ* hybridisation to metaphase chromosomes of (A) *N. nudicaulis* and (B) *N. repanda* using genomic DNA probes of the progenitor species *N. obtusifolia* (red) and *N. sylvestris* (green). Chromosomes are counterstained with DAPI (grey) and examples of signal at sub-telomeric regions are indicated with arrows. Scale bar is 5 μm.

Conclusion

The difference in GS between *N. repanda* and *N. nudicaulis* is likely to be a consequence of expansion of a few repeats in the former, particularly Ty3/*Gypsy*-like retroelements (Fig. 4.4). Moreover expansion of DNA in *N. repanda* is primarily the result of amplification of repeat families inherited from the progenitors in high copy-number (Fig. 4.2 A, B and C) and can involve the proliferation of novel variants within a family (Fig. 4.1). In contrast both allotetraploids appear to have experienced erosion of low copy-number repeats

(Fig. 4.2 and 4.3).

Diploidisation of allopolyploids is associated with the loss of duplicate genes, repetitive DNA and a reduction in chromosome number. I propose the loss of differentiation between the progenitor sub-genomes of an allopolyploid (i.e. homogenisation of the genome), through the processes described here, can also be considered a part of the diploidisation process.

Chapter 5 Conclusion

The power of next generation sequencing

The emergence of next generation sequencing (NGS) provides a novel opportunity to revolutionise biology, in similar ways to the advent of molecular biology techniques many decades before (Mardis, 2008; Shendure & Ji, 2008; Metzker, 2010). The power, volume and rapidity of data delivery available with NGS was inconceivable just a few years ago and biologists now have the possibility of designing experiments that would have otherwise been unthinkable. Indeed data analysis and storage are now major considerations when designing research projects and have become a significant bottle-neck when delivering research objectives. However many recent projects, both large and small have set out to use NGS to increase our understanding of genomics, gene expression (Buggs *et al.*, 2010a), metagenomics (Qin *et al.*, 2010) and human disease (Wheeler *et al.*, 2008) while re-sequencing of multiple individuals of the same species, has allowed detailed evolutionary analyses at the population level (Eckert *et al.*, 2009).

Next generation sequencing in allopolyploid plants

Although the use of NGS technology has been adopted by researchers in multiple disciplines, the use of such data in the examination of allopolyploid plants has hitherto been restricted to gene expression analysis (Buggs *et al.*, 2010a; Buggs *et al.*, 2010c; Buggs *et al.*, 2011) and analysis of small RNAs (sRNAs; Kenan-Eichler *et al.,* 2011). This thesis expands upon such studies and is the first example of genome analysis of allopolyploid plants using NGS data, as far as I am aware.

A theme of the thesis is the use of a newly developed graph-based clustering pipeline (Novak *et al.,* 2010). Graph-based clustering allows rapid characterisation of a genome at relatively low cost and has been used to survey plant genomes (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Wicker *et al.*, 2009; Novak *et al.*, 2010; Macas *et al.*, 2011). Previous efforts at genome surveys using NGS reads have relied on binning reads into groups based on similarity to known repeats (Swaminathan *et al.*, 2007; Wicker *et al.*, 2009). However graph-based clustering provides a powerful way to group reads into clusters (repetitive DNA families), which are subsequently assessed for similarity to known repeats. The advantage of such an approach is two-fold: [1] Entire clusters can be annotated via sequence similarity, even when some reads therein do not contain conserved repeat regions and alone cannot be classified as repetitive. With simple binning such reads would not be placed into any category despite being repetitive in origin. [2] Several species can be combined into a single graph-based clustering run and the abundance of each cluster compared between them. This allows detailed evolutionary comparisons between species. In this thesis I have taken advantage of both points [1] and [2] and used NGS, coupled with graph-based clustering, to analyse repetitive DNA and GS change in several diploid and allopolyploid species of *Nicotiana*.

The contribution of this thesis to the field of allopolyploid genomics

In chapter two I analysed the allotetraploid *N. tabacum* together with close relatives of its diploid progenitors *N. slyvestris* and *N. tomentosiformis* (Renny-Byfield *et al.*, 2011) where repetitive DNA characterisation was based on Roche 454 sequence data at low coverage (~0.5-1% of the genome). The reconstruction and quantification of repetitive DNA coupled with sequence similarity searches to known repeat families (RepBase; Jurka *et al.*, 2005) provided the first broad classification of repetitive DNA in the genus. By comparing the allopolyploid with diploid progenitors I demonstrated transposable elements (TEs), particularly retroelements, are major constituents of the genome. Furthermore many repeat clusters were under-represented in the allopolyploid *N. tabacum* and comparisons with abundance in the progenitors provided evidence for preferential elimination of paternally-derived DNA. I argued this elimination was responsible for the apparent genome downsizing seen in *N. tabacum*.

Since Song *et al.* (1995) published RFLP analysis of synthetic allopolyploids in *Brassica* it has been known that rapid genomic changes can occur in allopolyploid genomes. However as this study, and others that followed, used RFLP, AFLP and SSAP analysis the scale and quantification of DNA loss or gain has been lacking. In the third chapter (Renny-Byfield *et al.*, 2012) I continued work on *N. tabacum* and described the tandem repeat *Nic*CL3, unusual for its high abundance (~1.9%) in the genome of the progenitor *N. tomentosiformis* and its large monomer size (2.2 kb). I demonstrated the presence of this repeat in a number of accessions of *N. tomentosiformis* as well as some closely related species in section *Tomentosae*. Interestingly, read-depth analysis, using two independent NGS data sets (Roche 454 and Illumina), indicated *Nic*CL3 had been almost completely lost from the genome of *N. tabacum*. Fluorescence *in situ* hybridisation and Southern blot analysis revealed *Nic*CL3 was also absent in some fourth generation synthetics (Th37 lines). These observations demonstrated that in allopolyploids considerable volumes of DNA (around 1% of the genome) are lost in only four generations. Additionally the loss of *Nic*CL3 in natural and synthetic lines of *N. tabacum* indicated repeatability of DNA loss subsequent to allopolyploidy. Sequence similarity estimates between reads derived from *Nic*CL3 suggested that removal of *Nic*CL3 in natural *N. tabacum* may be the result of homogenisation coupled with unequal recombination.

The loss of DNA in allopolyploids is reasonably well documented and it has long been noted that the genomes of polyploid plants tend to downsize subsequent to their formation (Leitch & Bennett, 2004). Such downsizing is thought to be the result of repetitive DNA loss (Leitch & Bennett, 2004; Lim *et al.*, 2007). However detailed descriptions of GS change in allopolyploids have been lacking and the precise sequence types involved and their effects are poorly understood. In the final data chapter (chapter four) I examined GS change in two allopolyploids (*N. repanda* and *N. nudicaulis*) in *Nicotiana* section *Repandae* to investigate this phenomenon further.

Genome size in section *Repandae* varies by 33%, as well as departing from progenitor additivity. Importantly there has been apparent genome downsizing in the lineage leading to *N. nudicaulis* (14% smaller GS) and upsizing (29% larger GS) in *N. repanda* (Leitch & Leitch, 2008). Using Illumina data I quantified and compared clusters of repetitive DNA in the genomes of the diploids (*N.*

sylvestris and *N. obtusifolia*) and the two derived allopolyploids. This analysis revealed three factors effecting GS change: [1] Most (>90%) of repeat clusters were under-represented in both allopolyploids. [2] Overall clusters at low abundance were under-represented in *N. repanda* and *N. nudicaulis*. [3] However a minority of repeat clusters, inherited from the progenitors in high abundance, were over-represented, more evident in the genome of *N. repanda*. These observations revealed an erosion of low copy-number repeats in both allopolyploids, but that erosion was counteracted by expansion specific to the most abundant repeats in the genome of *N. repanda*. The majority of this expansion could be explained by an excess of Ty3/*Gypsy* like retroelements, equivalent to 23% increase in GS. These processes are concurrent with the failure to distinguish progenitor genomes by genomic *in situ* hybridisation and I propose that together these phenomena can be considered part of the diploidisation process.

Genome divergence in allopolyploids has been of considerable interest to researchers in plant genetics for many years. This thesis contributes to our knowledge of this subject by using a recently developed bioinformatics pipeline, covering a large number of repeat families, quantifying DNA loss and linking non-additivity in derived allopolyploids with GS change subsequent to their formation.

New questions and future research

Although genome alterations have been described in many allopolyploids little is known about the underlying causes of such change. Why do some allopolyploids experience relatively little perturbation (Liu *et al.*, 2001; Baumel *et al.*, 2002), while others demonstrate significant change in only a few generations (Song *et al.*, 1995; Ozkan *et al.*, 2001; Kovarik *et al.*, 2005; Gaeta *et al.*, 2007; Lim *et al.*, 2008; Feldman & Levy, 2009; Malinska *et al.*, 2010). Perhaps there are a variety of factors involved, such as control of TE proliferation, alterations to gene expression and epigenetic control mechanisms.

Epigenetic mechanisms have been well examined in diploids and include the action of small RNAs (sRNA), known to be associated with TEs (Cantu *et al.*, 2010) and which play a key role in limiting transcription through chromatin remodelling (Matzke *et al.*, 2009). Additionally DNA methylation, itself directed by sRNA pathways, is responsible for inactivation of the DNA transposon Activator (Ac) in maize (Chomet *et al.*, 1987) while in *Arabidopsis* mutants of the decreased in DNA methylation1 (*ddm1*) pathway show evidence of increased copy-number in several TE families (Tsukahara *et al.*, 2009).

To date we have only limited understanding of how epigenetic mechanisms are perturbed in allopolyploids. However we do know that in *Spartina* allopolyploids cytosine methylation patterns are altered near TE insertions compared to their diploid progenitors (Parisod *et al.*, 2009). Epigenetic repatterning also occurs in allotetraploid orchids (Paun *et al.*, 2010) and in wheat the abundance of siRNAs, which are key in the repression of TEs, decreased with increasing ploidy-level (Kenan-Eichler *et al.*, 2011). Kenan-Eichler and colleagues hypothesised that siRNA abundance may be involved in the relaxation of TE suppression in early generations of synthetic allopolyploid wheat and that these mechanisms could act across the whole spectrum of allopolyploids.

It is possible to use graph-based clustering to assess the impact of multiple epigenetic factors on repeat abundance in allopolyploids. Bisulphite treatment of DNA, which converts methylated cytosine to uracil allowing the identification of methylation patterns can be sequenced with NGS technology and combined with untreated genomic DNA. This would allow the joint clustering of both datasets and the assessment of methylation across a repeat cluster. In turn one could add small RNAs and ChIP-seq data to a clustering run. All these datasets combined would give information about change in abundance, methylation status, the population of sRNA associated with a cluster, as well as histone modifications, all in a single clustering run. Using clustering combined with mutants in various epigenetic pathways may elucidate deep and novel links between epigenetic states and sequence losses and gains across multiple families of repetitive DNA.

While much effort has been expended on analysing patterns of TE insertion, activation and removal little research time has been invested on TE dynamics at a population level and the processes and factors affecting the spread and fixation of TE insertions are poorly understood (reviewed in Tenaillon *et al.* 2010). The power of NGS allows the mapping of TE polymorphisms at a population level in both diploids, natural allopolyploids and synthetic lines. The clustering analysis used in this thesis can shed light on processes at the population level as many individuals of a species can be examined for repetitive DNA content relatively cheaply. Hopefully such approaches will allow us to untangle the impacts of effective population size, selection and drift on TE distribution and abundance.

Efforts at sequencing the genome of the allopolyploid *Brassica napus* will result in the publication of a draft genome in the coming year. This presents a wonderful opportunity for the study of allopolyploid genomics. A reference genome, together with the potential of re-sequencing of multiple individuals in several populations, will provide valuable tools to help understand genome divergence following allopolyploidy in this species.

Final remarks

Recent work, both presented here and in the literature, has uncovered dramatic, rapid and substantial changes to some allopolyploid genomes. Although these advances have been enlightening, we still have little knowledge of the underlying mechanisms, in part due to a lack of focus on epigenetics, few studies at the population level and a focus on lineage specific, rather than allopolyploid-induced, patterns of divergence. With research in these areas over the coming years we will uncover the processes that are most important in governing genome divergence following allopolyploidy.

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 4649-4654.

Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168**: 2217-2226.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**: 135-141.

Ainouche ML, Baumel A, Salmon A. 2004. *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. *Biological Journal of the Linnean Society* **82**: 475-484.

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Ashton PA, Abbott RJ. 1992. Multiple origins and genetic diversity in the newly arisen allopolyploid species, *Senecio cambrensis* (Compositae). *Heredity* **68**: 25-32.

Auger DL, Gray AD, Ream TS, Kato A, Coe EH, Birchler JA. 2005. Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics* **169**: 389-397.

Baack, E.J., Whitney, K.D. and Rieseberg, L.H. (2005) Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in Helianthus homoploid hybrid species. *New Phytologist* **167**, 623-630.

Barabaschi D, Guerra D, Lacrima K, Laino P, Michelotti V, Urso S, Vale G, Cattivelli L. 2012. Emerging knowledge from genome sequencing of crop species. *Molecular biotechnology* **50**: 250-266.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**: 2445-2455.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* **1**: 391-399.

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

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

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

Bennett MD, Leitch IJ 2010. Angiosperm DNA C-values database.Royal Botanic Gardnes Kew.http://www.kew.org/cvalues/

Bennett MD, Leitch IJ, Price HJ, Johnston JS. 2003. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25 % larger than the *Arabidopsis* genome initiative estimate of ~125 Mb. *Annals of Botany* **91**: 547-557.

Bennett MD, Smith JB. 1976. Nuclear DNA amounts in angiosperms. *Philisophical Transactions of the Royal Society of London B* **274**: 227-274.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics & Development* **15**: 621-627.

Bennetzen JL, SanMiguel P, Chen MS, Tikhonov A, Francki M, Avramova Z. 1998. Grass genomes. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 1975-1978.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, ColladoVides J, Glasner JD, Rode CK, Mayhew GF, *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453-1462.

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433-438.

Buerkle, C.A. and Rieseberg, L.H. (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266-275.

Buggs RJ, Chamala S, Wu W, Gao L, May GD, Schnable PS, Soltis DE, Soltis PS, Barbazuk WB. 2010a. Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology* **19**: 132-146.

Buggs RJ, Chamala S, Wu W, Tate J, Schnable P, Soltis D, Soltis P, Barbazuk WB. 2011. Rapid, repeated, and clustered loss of duplicated genes in allopolyploid *Tragopogon* populations of independent origin. *Curent Biology*: 996-1000.

Buggs RJ, Elliott NM, Zhang LJ, Koh J, Viccini LF, Soltis DE, Soltis PS. 2010b. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytologist* **186**: 175-183.

Buggs RJ, Renny-Byfield S, Chester M, Jordan-Thaden IE, Viccini LF, Chamala S, Leitch AR, Schnable PS, Barbazuk WB, Soltis PS, *et al.* 2012. Next-generation sequencing and genome evolution in allopolyploids. *American Journal of Botany*: 372-382.

Buggs RJ, Zhang L, Miles N, Tate J, Gao L, Wei W, Schnable P, Barbazuk W, Soltis P, Soltis D. 2010c. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Current Biology* **21**: 551-556.

Burk LG. 1973. Partial self-fertility in theoretical amphiploid progenitor of *N. tabacum Journal of Heredity* **64**: 348-350.

Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, Michelmore RW, Dubcovsky J. 2010. Small RNAs, DNA methylation and transposable elements in wheat. *BMC Genomics* **11**: 15.

Chague V, Just J, Mestiri I, Balzergue S, Tanguy AM, Huneau C, Huteau V, Belcram H, Coriton O, Jahier J, *et al.* 2010. Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. *New Phytologist* **187**: 1181-1194.

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

Chen ZJ, Ni ZF. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* **28**: 240-252.

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America* **109**: 1176-1181.

Chester M, Sykorova E, Fajkus J, Leitch AR. 2010. Single integration and spread of a copia-like sequence nested in rDNA intergenic spacers of *Allium cernuum* (Alliaceae). *Cytogenetic and Genome Research* **129**: 35-46.

Chomet PS, Wessler S, Dellaporta SL. 1987. Activation of the maize transposable element Activator (AC) is associateed with its DNA modification. *Embo Journal* **6**: 295-302.

Clarkson JJ, Kelly LJ, Leitch AR, Knapp S, Chase MW. 2010. Nuclear glutamine synthetase evolution in *Nicotiana*: Phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Molecular Phylogenetics and Evolution* **55**: 99-112.

Clarkson JJ, Knapp S, Garcia VF, Olmstead RG, Leitch AR, Chase MW. 2004. Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Molecular Phylogenetics and Evolution* **33**: 75-90.

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

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537.

Comai L, Madlung A, Josefsson C, Tyagi A. 2003. Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **358**: 1149-1155.

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. 2000. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* **12**: 1551-1567.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research* **12**: 1075-1079.

Dolezel J, Greilhuber J, Lucretti S, Meister A, Lysak MA, Nardi L, Obermayer R. 1998. Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Annals of Botany* **82**: 17-26.

Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB. 2009. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics & Genomes* **5**: 225-234.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics* **175**: 477-485.

Feldman M, Levy AA. 2009. Genome evolution in allopolyploid wheat-a revolutionary reprogramming followed by gradual changes. *Journal of Genetics and Genomics* **36**: 511-518.

Feldman M, Liu B, Segal G, Abbo S, Levy AA, Vega JM. 1997. Rapid elimination of low-copy DNA sequences in polyploid wheat: A possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**: 1381-1387.

Flagel L, Udall J, Nettleton D, Wendel J. 2008. Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biology* **6**: 16.

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

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531-1545.

Freeling M 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*, 433-453.

Fulnecek J, Lim KY, Leitch AR, Kovarik A, Matyasek R. 2002. Evolution and structure of 5S rDNA loci in allotetraploid *Nicotiana tabacum* and its putative parental species. *Heredity* **88**: 19-25.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* **19**: 3403-3417.

Ganley ARD, Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**: 184-191.

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

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**: 759-769.

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* **296**: 92-100.

Goodspeed T. 1954. The genus *Nicotiana*. *Chron Bot* **16**: 1-536.

Greilhuber J, Borsch T, Mueller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size. *Plant Biology* **8**: 770-777.

Grover CE, Wendel JF. 2010. Recent insights into mechanisms of genome size change in plants. *Journal of Botany* **2010**: 382732.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research* **16**: 1252-1261.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 17811-17816.

Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. 2006. Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current Biology* **16**: 1652-1659.

Hegarty MJ, Jones JM, Wilson ID, Barker GL, Coghill JA, Sanchez-Baracaldo P, Liu GQ, Buggs RJA, Abbott RJ, Edwards KJ, *et al.* 2005. Development of anonymous cDNA microarrays to study changes to the *Senecio* floral transcriptome during hybrid speciation. *Molecular Ecology* **14**: 2493-2510.

Heslop-Harrison JS, Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. *Plant Journal* **66**: 18-33.

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

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

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, *et al.* 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* **43**: 476-481.

Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal *Xenopus Laevis*. *Molecular Biology and Evolution* **10**: 1360-1369.

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

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463-467.

Jiang BA, Lou QF, Wu ZM, Zhang WP, Wang D, Mbira KG, Weng YQ, Chen JF. 2011. Retrotransposon- and microsatellite sequence-associated genomic changes in early generations of a newly synthesized allotetraploid *Cucumis* x *hytivus* Chen & Kirkbride. *Plant Molecular Biology* **77**: 225-233.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97-100.

Jones RN, Hegarty M. 2009. Order out of chaos in the hybrid plant nucleus. *Cytogenetic and Genome Research* **126**: 376-389.

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

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**: 462-467.

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

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, *et al.* 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Kazazian HH. 2004. Mobile elements: Drivers of genome evolution. *Science* **303**: 1626-1632.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution* **24**: 572-582.

Kellogg EA, Bennetzen JL. 2004. The evolution of nuclear genome structure in seed plants. *American Journal of Botany* **91**: 1709-1725.

Kelly LJ, Leitch AR, Clarkson JJ, Hunter RB, Knapp S, Chase MW. 2010. Intragenic recombination events and evidence for hybrid speciation in *Nicotiana* (*Solanaceae*). *Molecular Biology and Evolution* **27**: 781-799.

Kelly LJ, Leitch IJ. 2011. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research* **19**: 939-953.

Kenan-Eichler M, Leshkowitz D, Tal L, Noor E, Melamed-Bessudo C, Feldman M, Levy AA. 2011. Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics* **188**: 263-U259.

Kenton A, Parokonny AS, Gleba YY, Bennett MD. 1993. Characterisation of the *Nicotiana tabacum* -L genome by molecular cytogenetics. *Molecular & General Genetics* **240**: 159-169.

Khasdan V, Yaakov B, Kraitshtein Z, Kashkush K. 2010. Developmental timing of DNA elimination following allopolyploidization in wheat. *Genetics* **185**: 387-390.

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

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

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

Koukalova B, Reich J, Matyasek R, Kuhrova V, Bezdek M. 1989. A BamHI family of highly repeated DNA sequences of *Nicotiana tabacum*. *Theoretical and Applied Genetics* **78**: 77-80.

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

Kovarik A, Koukalova B, Lim KY, Matyasek R, Lichtenstein CP, Leitch AR, Bezdek M. 2000. Comparative analysis of DNA methylation in tobacco heterochromatic sequences. *Chromosome Research* **8**: 527-541.

Kovarik A, Matyasek R, Lim KY, Skalicka K, Koukalova B, Knapp S, Chase M, Leitch AR. 2004. Concerted evolution of 18-5.8-26S rDNA repeats in *Nicotiana* allotetraploids. *Biological Journal of the Linnean Society* **82**: 615-625.

Kovarik A, Pires JC, Leitch AR, Lim KY, Sherwood AM, Matyasek R, Rocca J, Soltis DE, Soltis PS. 2005. Rapid concerted evolution of nuclear ribosomal DNA in two *Tragopogon* allopolyploids of recent and recurrent origin. *Genetics* **169**: 931-944.

Kovarik A, Renny-Byfield S, Leitch AR 2011. Evolutionary implications of genome and karyotype restructuring in *Nicotiana tabacum*. L. In: P. S. SoltisD. E. Soltis eds. *Polyploidy and Genome Evolution* New York: Springer.

Kumar A, Bennetzen JL. 1999. Plant retrotransposons. *Annual Review of Genetics* **33**: 479-532.

Lai, Z., Gross, B.L., Zou, Y., Andrews, J. and Rieseberg, L.H. (2006) Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology* **15**, 1213-1227.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langdon T, Seago C, Jones RN, Ougham H, Thomas H, Forster JW, Jenkins G. 2000. De novo evolution of satellite DNA on the rye B chromosome. *Genetics* **154**: 869-884.

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

Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist* **194**: 629-646.

Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* **82**: 651-663.

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

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

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

Lim KY, Kovarik A, Matyasek R, Chase MW, Knapp S, McCarthy E, Clarkson JJ, Leitch AR. 2006a. Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section *Alatae*. *Plant Journal* **48**: 907-919.

Lim KY, Leitch IJ, Leitch AR. 1998. Genomic characterisation and the detection of raspberry chromatin in polyploid *Rubus*. *Theoretical and Applied Genetics* **97**: 1027-1033.

Lim KY, Matyasek R, Kovarik A, Leitch AR. 2004a. Genome evolution in allotetraploid *Nicotiana*. *Biological Journal of the Linnean Society* **82**: 599-606.

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

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

Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, Srubarova H, Kovarik A, Pires JC, Xiong Z, Leitch AR. 2008. Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). *PLoS One* **3**: e3353.

Lim KY, Souckova-Skalicka K, Sarasan V, Clarkson JJ, Chase MW, Kovarik A, Leitch AR. 2006b. A genetic appraisal of a new synthetic *Nicotiana tabacum* (Solanaceae) and the Kostoff synthetic tobacco. *American Journal of Botany* **93**: 875-883.

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

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

Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Molecular Phylogenetics and Evolution* **29**: 365-379.

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

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459-473.

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

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

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

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

Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. 2002. Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic *Arabidopsis* allotetraploids. *Plant Physiology* **129**: 733-746.

Malinska H, Tate JA, Matyasek R, Leitch AR, Soltis DE, Soltis PS, Kovarik A. 2010. Similar patterns of rDNA evolution in synthetic and recently formed natural populations of *Tragopogon* (Asteraceae) allotetraploids. *BMC Evolutionary Biology* **10**: 17.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**: 133-141.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

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

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

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

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

Matyasek R, Tate JA, Lim YK, Srubarova H, Koh J, Leitch AR, Soltis DE, Soltis PS, Kovarik A. 2007. Concerted evolution of rDNA in recently formed *Tragopogon* allotetraploids is typically associated with an inverse correlation between gene copy number and expression. *Genetics* **176**: 2509-2519.

Matzke M, Gregor W, Mette MF, Aufsatz W, Kanno T, Jakowitsch J, Matzke AJM. 2004. Endogenous pararetroviruses of allotetraploid *Nicotiana tabacum* and its diploid progenitors, *N. sylvestris* and *N. tomentosiformis*. *Biological Journal of the Linnean Society* **82**: 627-638.

Matzke M, Kanno T, Claxinger L, Huettel B, Matzke AJM. 2009. RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology* **21**: 367-376.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333**: 1257-1257.

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

Melayah D, Lim KY, Bonnivard E, Chalhoub B, De Borne FD, Mhiri C, Leitch AR, Grandbastien MA. 2004. Distribution of the Tnt1 retrotransposon family in the amphidiploid tobacco (*Nicotiana tabacum*) and its wild *Nicotiana* relatives. *Biological Journal of the Linnean Society* **82**: 639-649.

Metzker ML. 2010. Applications of next generation sequencing technologies - the next generation. *Nature Reviews Genetics* **11**: 31-46.

Murad L, Bielawski JP, Matyasek R, Kovarik A, Nichols RA, Leitch AR, Lichtenstein CP. 2004. The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. *Heredity* **92**: 352-358.

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

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

Ohno S. 1970. *Evolution by Gene Duplication*. New York: Spiringer-Verlag.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, *et al.* 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33**: 5691-5702.

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

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

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

Paun O, Bateman RM, Fay MF, Hedrén M, Civeyrel L, Chase MW. 2010. Stable epigenetic effects impact adaptation in allopolyploid orchids (*Dactylorhiza*: Orchidaceae). *Molecular Biology and Evolution* **27**: 2465-2473.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164**: 10-15.

Pertea G, Huang XQ, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, *et al.* 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**: 651-652.

Petit M, Guidat C, Daniel J, Denis E, Montoriol E, Bui QT, Lim KY, Kovarik A, Leitch AR, Grandbastien MA, *et al.* 2010. Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist* **186**: 135-147.

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

Pikaard CS. 2000. Nucleolar dominance: uniparental gene silencing on a multi-megabase scale in genetic hybrids. *Plant Molecular Biology* **43**: 163-177.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, *et al.* 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59-65.

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

Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162**: 1389-1400.

Rapp RA, Udall JA, Wendel JF. 2009. Genomic expression dominance in allopolyploids. *BMC Biology* **7**: 18.

Renny-Byfield S, Chester M, Kovařík A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novák P, W. Chase M, *et al.* 2011.

Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28**: 2843-2854.

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

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T. 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Molecular Phylogenetics and Evolution* **53**: 826-834.

Salina EA, Numerova OM, Ozkan H, Feldman M. 2004. Alterations in subtelomeric tandem repeats during early stages of allopolyploidy in wheat. *Genome* **47**: 860-867.

Salmon A, Ainouche ML, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Molecular Ecology* **14**: 1163-1175.

Sambrook J, Russell D. 2001. *Molecular cloning: a laboratory manual*. New York: Cold Srping Harbour Labaratory Press.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5463-5467.

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Functional & integrative genomics* **2**: 70-80.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**: 1749-1759.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135-1145.

Skalicka K, Lim KY, Matyasek R, Koukalova B, Leitch AR, Kovarik A. 2003. Rapid evolution of parental rDNA in a synthetic tobacco allotetraploid line. *American Journal of Botany* **90**: 988-996.

Skalicka K, Lim KY, Matyasek R, Matzke M, Leitch AR, Kovarik A. 2005. Preferential elimination of repeated DNA sequences from the paternal, *Nicotiana tomentosiformis* genome donor of a synthetic, allotetraploid tobacco. *New Phytologist* **166**: 291-303.

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

Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E. 2004. Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society* **82**: 485-501.

Song KM, Lu P, Tang KL, Osborn TC. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution.

Proceedings of the National Academy of Sciences of the United States of America **92**: 7719-7723.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C. and Liston, A. (2012) Navingating the tip of the genomic iceberg: Next generation sequencing for plant systematics *American Journal of Botany*, **99**, 349-364.

Stebbins GL. 1950. *Variation and Evolution in Plants*. New York: Columbia.

Steimer A, Amedeo P, Afsar K, Fransz P, Scheid OM, Paszkowski J. 2000. Endogenous targets of transcriptional gene silencing in *Arabidopsis*. *Plant Cell* **12**: 1165-1178.

Stephens SG. 1951. Possible significance of duplication in evolution. *Advanced Genetics* **4**: 247-265.

Sun Y, Skinner DZ, Liang GH, Hulbert SH. 1994. Phylogenetic analysis of *Sorghum* and related taxa using internal transcribed spacers of nuclear ribosomal DNA *Theoretical and Applied Genetics* **89**: 26-32.

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

Szadkowski E, Eber F, Huteau V, Lode M, Huneau C, Belcram H, Coriton O, Manzanares-Dauleux MJ, Delourme R, King GJ, *et al.* 2010. The first meiosis of resynthesized *Brassica napus*, a genome blender. *New Phytologist* **186**: 102-112.

Ungerer, M.C., Strakosh, S.C. and Zhen, Y. (2006) Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current Biology* **16**, R872-R873.

Tek AL, Song JQ, Macas J, Jiang JM. 2005. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. *Genetics* **170**: 1231-1238.

Tenaillon MI, Hollister JD, Gaut BS. 2010. A triptych of the evolution of plant transposable elements. *Trends in Plant Science* **15**: 471-478.

Trick M, Long Y, Meng J, Bancroft I. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal* **7**: 334-346.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**: 423-426.

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Research* **31**: 4401-4409.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Vershinin AV, Schwarzacher T, Heslop-Harrison JS. 1995. The large-scale genomic organization of repetitive DNA families at the telomeres of rye chromsomes. *Plant Cell* **7**: 1823-1833.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290**: 2114-2117.

Volkov RA, Borisjuk NV, Panchuk, II, Schweizer D, Hemleben V. 1999. Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. *Molecular Biology and Evolution* **16**: 311-320.

Wang JL, Tian L, Lee HS, Wei NE, Jiang HM, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, *et al.* 2006. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172**: 507-517.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Molecular Biology* **42**: 225-249.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-876.

Wicker T, Keller B. 2007. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Research* **17**: 1072-1081.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al.* 2007. A unified classification system for eukaryotic transposable elements. *Nature Review Genetics* **8**: 973-982.

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

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

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 13875-13879.

Yassour M, Kapian T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, *et al.* 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 3264-3269.

Zhang W, Dolan ME. 2010. Impact of the 1000 Genomes Project on the next wave of pharmacogenomic discovery. *Pharmacogenomics* **11**: 249-256.

Zhang Z, Belcram H, Gornicki P, Charles M, Just J, Huneau C, Magdelenat G, Couloux A, Samain S, Gill BS, *et al.* 2011. Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 18737-18742.

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

Appendix

Published articles:

Buggs RJ, **Renny-Byfield S**, Chester M, Jordan-Thaden IE, Viccini LF, Chamala S, Leitch AR, Schnable PS, Barbazuk WB, Soltis PS*, et al.* **2012.** Next-generation sequencing and genome evolution in allopolyploids. *American Journal of Botany* **99**: 372-382.

Kelly LJ, Leitch AR, Fay MF, **Renny-Byfield S**, Pellicer J, Macas J, Leitch IJ. **2012.** Why size *really* matters when sequencing plant genomes. *Plant Ecology and Diversity* in press.

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

Kovarik A, **Renny-Byfield S**, Leitch AR **2011.** Evolutionary implications of genome and karyotype restructuring in *Nicotiana tabacum*. L. In: P. S. Soltis, D. E. Soltis eds. *Polyploidy and Genome Evolution* New York: Springer.

Renny-Byfield S, Ainouche M, Leitch IJ, Lim KY, Le Comber SC, Leitch AR. **2010.** Flow cytometry and GISH reveal mixed ploidy populations and *Spartina* nonaploids with genomes of *S. alterniflora* and *S. maritima* origin. *Annals of Botany* **105**: 527-533.

Renny-Byfield S, Chester M, Kovařík A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novák P, W. Chase M*, et al.* **2011.** Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28**: 2843-2854.

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

Available PDFs of these papers are supplied with the attached DVD.

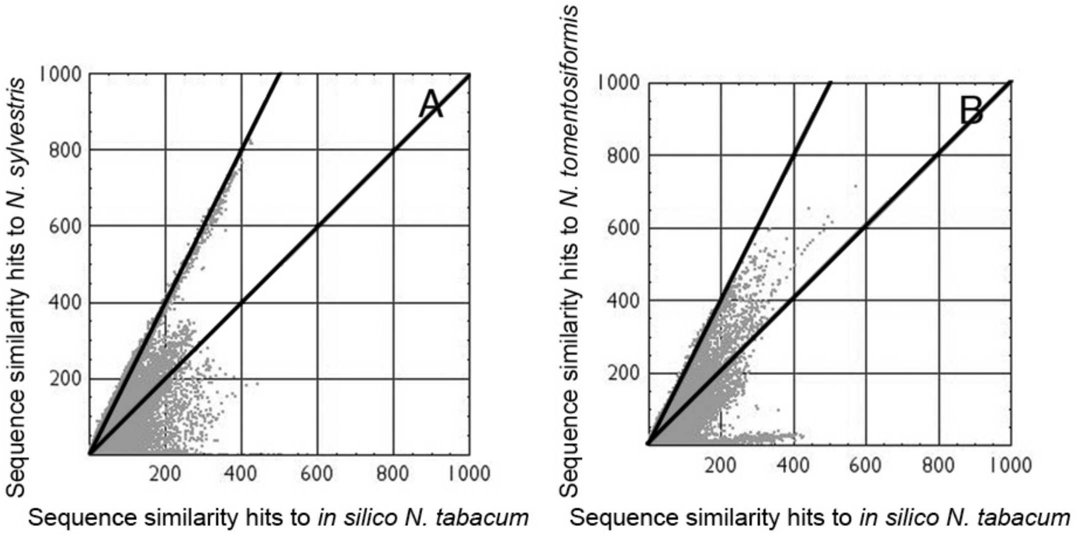

Figure A.1

Sequence similarity hits to the *in silico N. tabacum* made from an equal mix of 35,000 454 reads from each of the progenitor species. The number of hits in the *in silico N. tabacum* is compared to the number of hits in (A) *N. sylvestris* and (B) *N. tomentosiformis*.

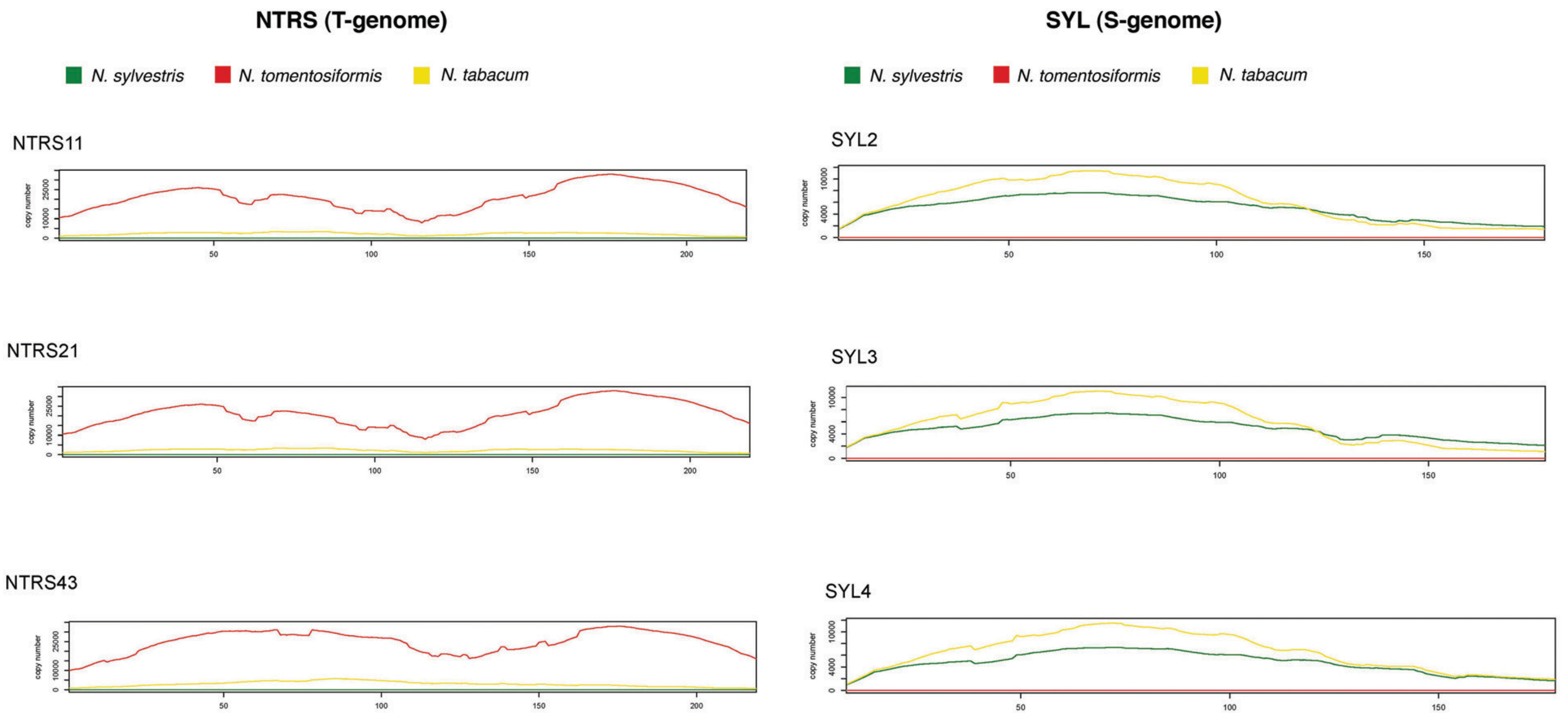

Figure A.2

Copy-number estimates along monomers of two satellite repeat families (NTRS and SYL) as estimated by BLAST analysis. The copy-number in *N. tabacum*, *N. tomentosiformis* and *N. sylvestris* is shown.

Figure A.3

A cloned sequence (1379 bp) bridging two monomers of *Nic*CL3, the length confirms the monomer size of 2.2 kb.

```
>CL3_tandem_clone_10

GGCGTCTGCTACGAGCTCGGACACTAGTAACGGCCGCCAGTGTGCTGGAATTCGCCCTTTAAA
ACTCCCAACATCCGGTGATACCAAGTCCCAAGCTCTACAACAGTATACTAAATATCCCCATACAA
CATTATCTATAAAAAGGGAAAATGAAATAGAACTAGGTAGAGGGTGACTCCGAGGCCCGCAGAT
GCCGGCAGGTATACCTTGAAGTCTCCATATCTAAACTCTACTCACCGGTGCCTAGGCTGGTAAG
GCTACCTGGATCTGCATAAAAAGATGTGCAAAAGTGTAGAATGAGTAGACTACAATGGTACCCA
ATAAGTGCTAACCTCGATAGAGTAGTGAAGAGGTCAGGTCAACACCCTACTAGAATAAATAAGA
AAATTGAACAGGTATATGATAGTGTGATAATGATAAACAACAAAAATGAAACAATGAAGAAAATA
TACCAAGAATAGATACACTGAATTAAGGCAATTAAGGCATCACGAAAGTAAACAATGAATAATAT
GAACATGGAAACAAAGAAACAAATACACAAGTGATATAATTGAATGGTATTAGCAAGCATCACTA
CCGAGGTACTGCCTCATAGTCTCATATCACAAAACAAATCATAATCATTCCTTATATCACCGCGG
GAGCCTTGCATTTAGTTTTGAAAATCATTTTTTCGAAATAGCTTCCCGTGTTTTAGCCCACCTTTT
CACACCGCGTGGCTTCAAGTAGTTCCTCTACTACCAACAAGCATATCAAGACCACCTTATCTCA
CCATATGCATTTCAACCCCAATCCTTAATCCACCACATGCGTATCAATGTCACAACATATAACTTT
CTGGTGGACACCACTATATCCACCAGAATCTCGAGGACTCTTGGTCTCTCTGTCATGTCTCCTG
GCCTTGCAAACCCTTCAATCTCCATGTGATTTCTACCACCTGCTGGTATGGAGTATCAGTCTGCA
ACTATCGAGCCATGCTGAATCTAATACCATAGTTGAGCCTCTTGATAAATCGACGGACTCGATCT
CTAATTGTGTAAACCAAGGCAGGTGCATATCTGAACAACTCACTAAACTTCAGAACATACTCTGA
TACGATCATAGAACCCTGGCACAACTGCTCAAACTGCTCAATCTATGTATCCCAAAGGCTTCGG
GGAACAAAATCTCAATAACATCTCTAAATATTGAGCCCATGTGAGTGAAGCTTCCTCGGTTGGG
CTACCCTCCTCCCAAGCCCGCCAAGACTAGTATGCTAGTCCCGATAGCTAGAAAGAAGTGAAA
GAACCCCCGCTCGTCTTCACTATACCCAAAGGGCGAATTCTGCAGATATCCATCACACTGGCG
GCCGCTCGAGCATGCATCAAGCGAT
```

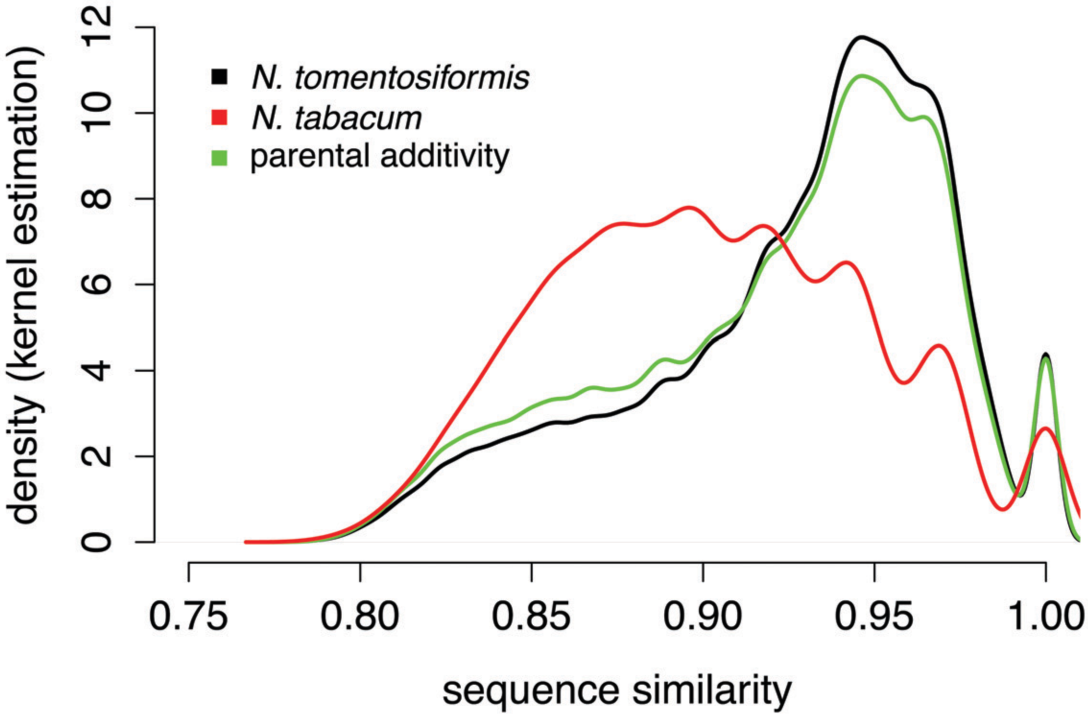

Figure A.4

Sequence similarity among *Nic*CL3 derived reads in *N. tomentosiformis*, *N. tabacum* and an *in silico* mix of reads from both progenitor species (*N. tomentosiformis* and *N. sylvestris*).