Chapter 5

Employing next generation sequencing to explore the repeat landscape of the plant genome

Hanna Weiss-Schneeweiss,[1] Andrew R. Leitch,[2] Jamie McCann,[1] Tae-Soo Jang[1] & Jiři Macas[3]

1 Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria
2 School of Biological and Chemical Sciences, Queen Mary College, University of London, London E1 4NS, U.K.
3 Biology Centre of the Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic
Author for correspondence: H. Weiss-Schneeweiss, hanna.schneeweiss@univie.ac.at

Abstract Plant genomes are dominated by repetitive DNA sequences of many distinct types which, due to their variability and high copy numbers, are difficult to analyse. This sequence complexity has recently become accessible to detailed investigations with the advent of next-generation sequencing technologies and novel analytical approaches utilizing low-pass genome sequence data. Here we explore the insights these methods have provided to our understanding of plant genomes, and review bioinformatic tools used in their analyses. We show that the repeats are diverse and fast diverging even between closely related lineages. We explore how the method has been used to study repeats in large and small plant genomes, and in association with polyploidy, distinctive chromosomes (sex chromosomes, B chromosomes, holocentric chromosomes), chromosome domains (e.g., rDNA loci and centromeres) and species groups of interest. It is apparent from these works that the dynamics of repeat evolution are complex, and divergence among repeats is shaping genome evolution, function and activity, the significance of which we are only starting to understand.

Keywords evolution; NGS; plant genome; repetitive DNA

Hörandl, E. & Appelhans, M.S. (eds.), Next-Generation Sequencing in Plant Systematics.
Chapter 5, http://dx.doi.org/10.14630/000006

Repetitive DNA in plant genomes

A substantial proportion of plant genomes is composed of repetitive DNA. Whilst gene numbers can vary to some extent, especially in relation to de novo polyploidy, it is the repetitive DNA which is largely responsible for the C-value paradox that refers to the enormous genome size variation among higher plants. Indeed, the 1C value, the amount of DNA in the haploid unreplicated nucleus, ranges 2,500 fold from 1C = 0.06 pg in *Genlisea margaretae* (Lentibulariaceae; Greilhuber & al., 2006) to 1C = 152 pg in *Paris japonica* (Melanthiaceae; Pellicer & al., 2010).

The repetitive DNA fraction of the genome is also diverse and fast evolving and can be involved in chromosomal and genomic rearrangements (Zhang J. & al., 2013; Knoll & al., 2014). Typically two major types of repetitive DNAs are distinguished, tandem repeats and dispersed repeats (Heslop-Harrison & Schmidt, 1998; Weiss-Schneeweiss & Schneeweiss, 2013). Within each of these classes further families or types of elements exist. The tandem repeats encompass satellite DNA, microsatellites and 5S and 35S ribosomal DNAs (rDNA). Whilst rDNAs are ubiquitous and the ribosomal RNA (rRNA) genes they encode are conserved, associated intergenic spacer sequences (IGSs) as well as satellite DNAs are fast diverging and typically species or genus specific (Plohl & al., 2008). Microsatellites are also fast diverging and may be used for phylogenetic and phylogeographic analyses, but their types and genomic abundances vary substantially between lineages (Ellegren, 2004). The dispersed repeats are mostly represented by mobile genetic elements, known also as transposable elements (TEs) (Kumar & Bennetzen, 1999; Kejnovský & al., 2012). These are divided into several distinct categories, the DNA transposons (DNA-TEs) using transposition (cut-and-paste mobility), *Helitrons* which may replicate by rolling-circle replication (Kapitonov & Jurka, 2001) and retroelements that amplify using retrotransposition (RNA-mediated copy-and-paste mobility). Retroelements are further divided into LTR (long terminal repeat) and non-LTR retrotransposons (LINEs and SINEs). It is the LTR retroelements, including two major families, Ty3/*gypsy* and Ty1/*copia* retrotransposons that are the most significant dispersed repeats in higher plant genomes (Kejnovský & al., 2012). Repeats are extremely variable both in their abundance and sequence length, occurring in up to millions of copies per genome and ranging in size from a few base pairs to many thousands (Hemleben & al., 2007; Heslop-Harrison & Schwarzacher, 2011).

Analyses of the repetitive genome fraction

With the discovery of repetitive DNA in eukaryotic genomes by DNA reassociation kinetics analysis (Britten & Kohne, 1968), it was shown that plant genomes have particularly diverse and abundant repetitive families (Flavell, 1986). However, owing to the repeat complexity and sequence variability, this substantial component of the genome has typically been treated as an inconvenience in genome assemblies and of reduced interest because it is not "genic". However, it is now clear that this perspective needs reconsideration because: (1) genome size, significantly influenced by repeat abundance, impacts phenotype directly, through, for example, influences on cell size and cell cycle time (Greilhuber & Leitch, 2013); (2) repeats are prone to illegitimate recombination, which when occurring between adjacent copies can result in excision of intervening sequences (Kejnovský & al., 2012), including potentially genes; (3) TEs can, through insertions and induced DNA breaks, disrupt gene function; and (4) genes when flanked by TEs (as in the majority of genes in maize, Schnable & al., 2009) can have altered expression patterns through, for example, epigenetic silencing of elements spreading to the genes (West & al., 2014).

Well-established, traditional approaches to study repeats, involving methods such as PCR, cloning, Southern hybridization and Sanger sequencing, were limited by both financial and practical constraints. Consequently, using repeats for chromosomal and phylogenetic analyses was overwhelmingly dominated by studies using rDNA probes, exploiting the conserved nature of their genic regions. Nevertheless a number of satellite DNA sequences were well characterized and studied in a range of plant species (Hemleben & al., 2007 and references therein; Heslop-Harrison & Schwarzacher, 2011). In contrast to satellite repeats, transposable elements, which usually make up the largest fraction of plant genomes (Kumar & Bennetzen, 1999), have been less studied (Kumar & Hirochika, 2001; Petit & al., 2007). This is because isolation and characterization of full-length elements required considerable resources, although the research is essential for phylogenetic and population genetic analyses involving fingerprinting techniques such as SSAP, IRAP, REMAP (Kumar & Hirochika, 2001). To circumvent the need to clone and sequence the entire TE repeat, the conserved reverse transcriptase (rt) domains of retroelements were exploited for phylogenetic, molecular and comparative cytogenetic analyses (e.g., Park J.-M. & al., 2007).

In contrast to these established traditional approaches, next-generation sequencing (NGS)-based analyses can target repeats across the entire genome by using data from a single sequencing run. For example, the cur-

rently widely used platform, Illumina HiSeq, delivers at least hundreds of millions individual reads per sequencing lane, providing over 1–3× coverage for plant genomes of intermediate size (e.g., *Pisum sativum* with ca. 4.3 pg/1C) or 10–30× for species with small genomes (e.g., *Oryza sativa* ca. 0.43 pg/1C). Although such read volumes are insufficient for draft assemblies of all but the smallest plant genomes, they provide sufficient redundancy for analyzing repetitive elements in most genomes. In practice, for many applications a "genome skimming" approach is adequate to characterize repeats, requiring only ten-fold or even hundred-fold smaller volumes of sequence than the genome size itself. However, with the advent of large-scale NGS data came the bottleneck of bioinformatic tools to analyse the data and for efficient de novo identification of repeats. To overcome that problem, some approaches have relied on the identification of repeated motifs in DNA sequences (the repeated substring approach, Kurtz & Schleiermacher, 1999), which once identified can be merged, where possible, into longer contigs (Volfovsky & al., 2001). Other approaches have required the virtual breakup of sequencing reads into small fragments (k-mers of length 20 bp) and then recording the frequencies of all k-mers in a library of "Mathematically Defined Repeats" (MDR). Identified repeats could then be targeted back to annotate genome assemblies (Wicker & al., 2008). There also exist analytical approaches for de novo identification of repeats which exploits the large number of NGS reads as a random sample of the genome (Swaminathan & al., 2007; Flutre & al., 2011; Natali & al., 2013). Detecting repeats is achieved via direct assembly of a large number of sequence reads and conducting similarity searches to known repeat databases. The assemblies can be further analysed for putative repeats by mapping sequence reads to assembled contigs and isolating all regions with high read depth. Another approach used to facilitate a broader range of repeat analyses, the Assisted Automated Assembler of Repeat Families (AAARF) identifies sequence overlaps and walks then out to create long pseudomolecules representing the most abundant repeats in any genome (DeBarry & al., 2008; Estep & al., 2013).

Recently, an efficient pipeline called "RepeatExplorer" has been developed which allows for the de novo identification of repeats in genomes lacking a reference genome (Macas & al., 2007; Novák & al., 2010, 2013). A general outline to the approach is shown in Fig. 1. This pipeline uses a similarity-based read clustering approach that allows detection of repetitive sequences, which are identified as groups of frequently overlapping sequence reads in all-to-all read comparison. The clustering procedure employs graph-based methods that transform read similarities to a virtual graph, where reads are represented as nodes and their similarities by edges connecting the nodes. Graph

topology is subsequently used for identification of communities of densely connected nodes representing various families of repetitive DNA sequences (Fig. 2). The reads within the sequence clusters can be assembled to generate contigs that represent the repeats they contain. The pipeline is continuously being developed and improved (http://www.repeatexplorer.org/).

Studies using various types of bioinformatic analyses of NGS data have now targeted the repetitive fraction of several plant groups exhibiting, for example: (1) large or small genome sizes, or large within-group genome size variation; (2) polyploidy (both auto- and allopolyploidy); (3) variation in amounts and localization of heterochromatin; (4) holocentric chromosomes; (5) sex chromosomes and (6) supernumerary B chromosomes. The NGS-based approach has also been used to characterize the complex organisation of centromeric domains (see below), and has been used to identify novel telomere motifs in plants (Peška & al., 2015). Several other species from various genera and families have also been analysed including *Olea europea* (Oleaceae, Barghini & al., 2014), *Capsicum annuum* (Solanaceae, Park M. & al., 2012), *Helianthus* (Asteraceae, Staton & al., 2012; Natali & al., 2013), *Cucumis sativus* (Cucurbitaceae, Huang & al., 2009), and species in Poaceae including

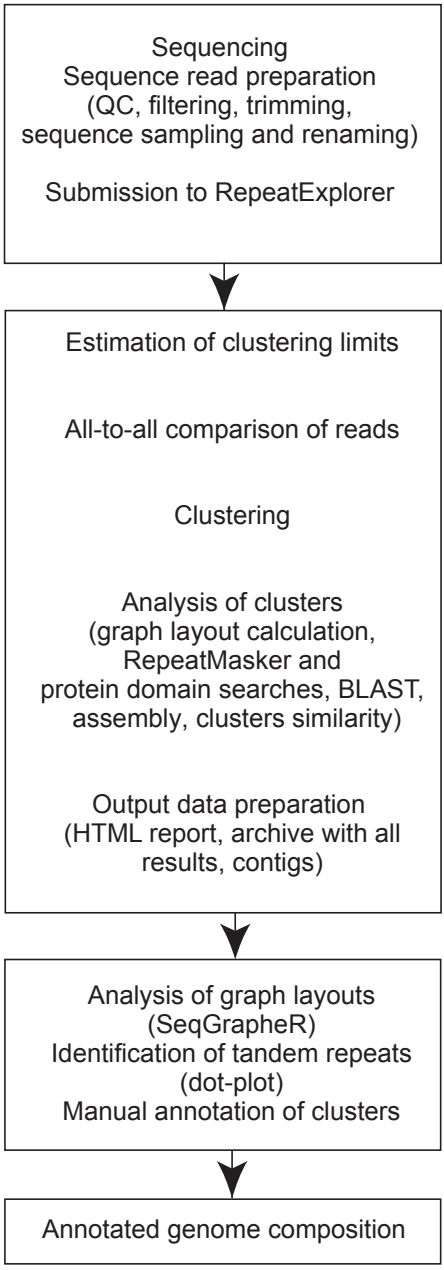

Fig. 1. Flow diagram showing the process of repeat identification with RepeatExplorer. Modified from Novák & al. (2010).

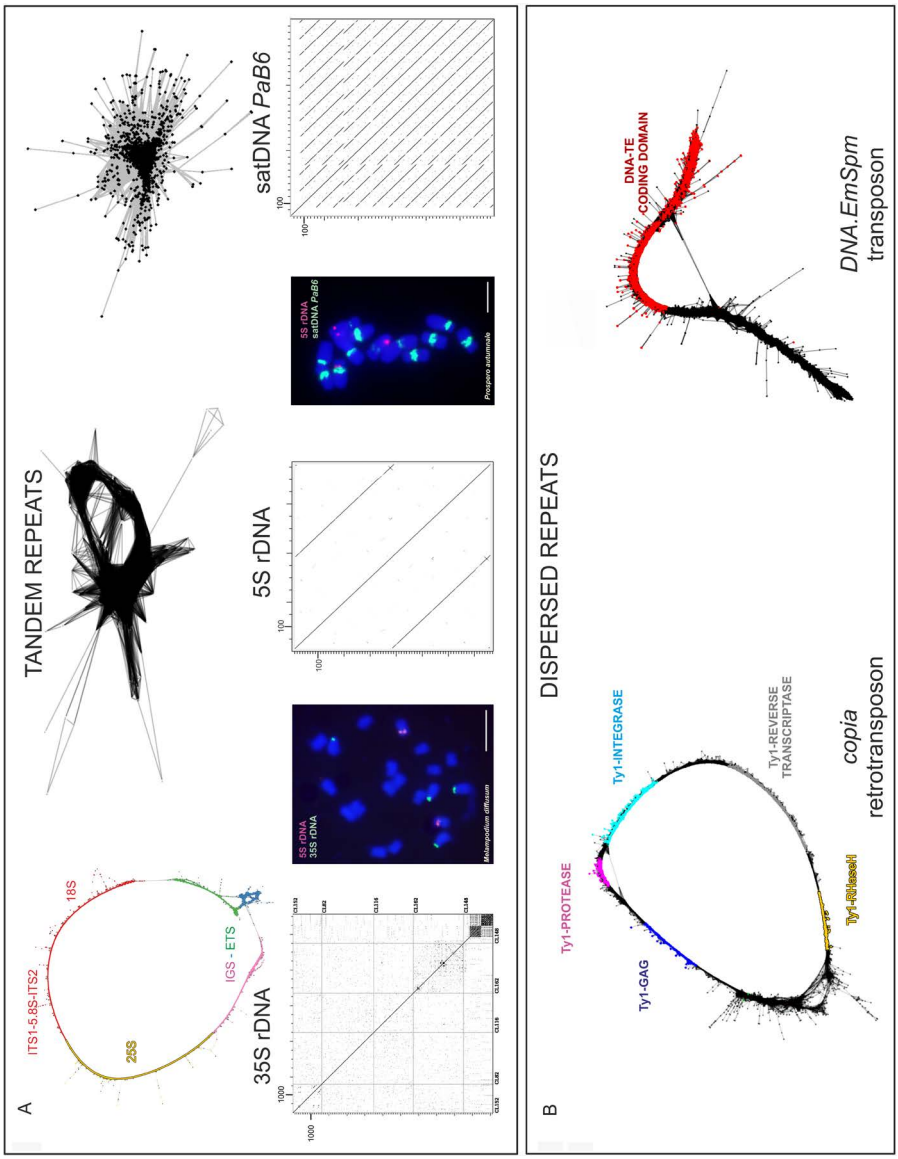

Fig. 2. Major types of repeats and their cluster characteristics. **A,** Tandem repeats include 35S rDNA, 5S rDNA and satellite DNAs; for each of these graph structures, dot-plots of the largest contigs and chromosomal localization are shown. 35S and 5S rDNA data are from genus *Melampodium*, satellite *PaB6* from genus *Prospero*. **B,** Dispersed repeats include DNA-TEs and retroelements; for each of these classes examples of graph structures and their annotations are given (both from genus *Melampodium*; unpub.).

Zea mays, *Hordeum, Triticum*, *Secale* and *Aegilops* (Wicker & al., 2006, 2008, 2009; Tenaillon & al., 2011; Ben-David & al., 2013; Senerchia & al., 2013).

This review details several of these studies in selected plant genera/families. It focuses on examples where NGS has been used to generate reliable estimations of repeat content across the genome. In particular it explores how genome skimming approaches have lead to deeper understanding of genome structures, patterns of evolution and mechanisms leading to genome size increase or reduction.

Initial studies of repeats using NGS

The first species to be analysed for their repetitive content using NGS approaches were the soybean, *Glycine max* and the pea, *Pisum sativum* in family Fabaceae (Macas & al., 2007; Swaminathan & al., 2007). *Glycine max* 454 pyrosequencing data were first analysed against the TIGR plant repeat databases, which include genomic data from *G. max* and other angiosperms. New repeat types were also identified by assembling the genome reads to contigs and evaluating the read coverage against each contig (Swaminathan & al., 2007). In a simultaneous and separate study, repeats were identified with a similarity-based read clustering approach with NGS reads from *P. sativum*, using a 454/Roche read volume equivalent to only 0.77% of the species genome size (Macas & al., 2007). Then, by combining sequence similarity searches with detection of conserved protein domains, it was possible to identify most of the abundant repeats into specific classes of elements, which collectively comprised ca. 35%–48% of the genome. Recent analysis of a novel dataset of *P. sativum* Illumina reads (0.1× genome coverage) using RepeatExplorer confirmed relative proportions of individual repeat types but provided better annotation of repeats increasing their proportion to 76% of the pea genome (Macas & al., unpub.). The majority of *P. sativum* genome was made up of LTR retrotransposons, dominated by Ty3/*gypsy* elements, in particular Ogre elements (Macas & Neumann, 2007). In addition, a large set of novel satellite repeats was identified (see below). The study also, for the first time, used NGS for whole-length element reconstruction. In addition, *rt* regions were isolated from reconstructed contigs, and used for large-scale phylogenetic analyses to reveal intragenomic relationships, providing an efficient and less biased alternative to traditional cloning approaches. This approach subsequently led to the development of the RepeatExplorer pipeline, described above (Novák & al., 2010, 2013).

Large plant genomes analysed using NGS

Gymnosperms. — Gymnosperms have large genomes in comparison with most angiosperms (gymnosperms modal 1C genome size = 10.0 pg, mean 1C = 18.8 pg, angiosperm modal 1C = 0.6 pg, mean 1C = 5.9 pg), and outside of genus *Ephedra* polyploidy is rare (Leitch A.R. & Leitch, 2012). The large genomes of gymnosperms are most likely the result of amplification of various types of repeats. Attempts to reconstruct the genome of conifers of *Pinus taeda* (Kovach & al., 2010) and *Picea abies* (Nystedt & al., 2013) using NGS is being seriously hampered by the large number of repeats. Surprisingly, however, individual repeats do not have particularly high genome proportions compared with repeats in many angiosperms, the genome being better characterized by a high abundance of heterogeneous repeats, especially LTR-retroelements (Nystedt & al., 2013). It is hypothesized that these repeats are accumulating slowly with a still lower frequency of excision. In one study, a Ty3/*gypsy* type retroelement of the Athila lineage called Gymny, appears to have amplified after the divergence of genera *Pinus* and *Picea*, but before further diversification of the subgenus *Pinus* approximately 16 Ma. In contrast two other gymnosperm retroelements, IFG7 and TPE1 (Morse & al., 2009) were shown to be shared by various genera of gymnosperms and may be more slowly diverging.

Angiosperms. — The modal genome size of angiosperms is small compared with all other land plant groups of comparable size (Leitch I.J. & Leitch, 2013). Nevertheless a few angiosperm lineages have experienced massive genome size enlargement associated with species radiation, the largest known being in genus *Fritillaria* (Liliaceae), where there is an astonishing 35 pg size range in 1C value between species (Kelly & Leitch, 2011). Analyses of repeat composition of nine species in this genus reveal, as in gymnosperms, a large heterogeneous collection of repeats (Kelly & Leitch, 2011; Kelly & al., 2015). The complexity is indicative of an evolutionary dynamics of repeats in *Fritillaria* that is similar to analyzed gymnosperms, suggesting that the character is a feature of large genomes rather than of specific evolutionary lineages.

Small plant genomes analysed using NGS

Some angiosperm genomes have undergone drastic genome size reduction in their ancestry and now contain very few repeats (Ibarra-Laclette & al., 2013). The smallest reported genome is that of *Genlisea margaretae* (0.06 pg/1C) and related taxa in family Lentibulariaceae (Greilhuber & al., 2006). Recent

sequencing of the complete genome of *Utricularia gibba* (0.09 pg/1C) using a hybrid (454/Illumina/Sanger) sequencing strategy revealed a paucity of repetitive DNA, which amounts to only 3.12% of the genome, largely composed of LTR retrotransposons (2.6%). Among 369 identified copies of retroelements only 58 were complete, representing 0.5% of the genome. The high frequency of incomplete TEs or fragmented LTR retrotransposons associated with numerous microdeletions encompassing segments or whole genes of retroelements in *U. gibba* indicated that any genome expansion through retrotransposon amplification has been counterbalanced, probably at a faster rate over recent evolutionary history, by removal of the elements via illegitimate recombination (mechanistically as suggested in *Arabidopsis*, Ibarra-Laclette & al., 2013).

Repeats analyses across families and genera using NGS

Musaceae. — A study of *Musa acuminata* (cv. 'Calcutta 4') using clustering analysis of 454/Roche reads revealed that repeats make up about 30% of its genome (ca. 0.64 pg/1C). The repeats are dominated by LTR-retrotransposons, with Ty1/*copia* being about two-fold more abundant than Ty3/*gypsy*. DNA-TEs were found to be rare and only two new satellite DNA repeats were identified (Hřibová & al., 2010; Čížková & al., 2013). In a follow-up study the analyses of the repeats were extended to six related species, selected to represent various phylogenetic groups within Musaceae, these being five species of *Musa* (in sect. *Musa* and sect. *Callimusa*) and *Ensete gilletii* (Novák & al., 2014). LTR retrotransposons of the Maximus/SIRE and Angela lineages of Ty1/*copia* and the chromovirus lineage of Ty3/*gypsy* elements were dominant elements in all species (representing 14%–34.5% of the genomes). However, there were quantitative differences and sequence variation detected for all repeat families. These differences occurred at the intersectional level, whereas pairs of closely related species shared more similar populations of repetitive elements. This data suggested that the total repetitive fraction of the genome contains phylogenetic and evolutionary signal (Novák & al., 2014; see the development of a novel phylogenetic approach below).

Orobanchaceae. — NGS analyses (454/Roche platform) to identify the repetitive DNA fraction of representatives of several genera have provided insights into the repeat composition of species in two holoparasitic (non-photosynthetic) genera of plants, *Orobanche* and *Phelipanche* (Piednoël & al., 2012, 2013). These genera are closely related but differ in chromosome numbers (x = 19 and 12, respectively; Schneeweiss & al., 2004) and genome

size (1.5–3.3 pg/1C in analysed *Orobanche* species and about 4.4 pg/1C in *Phelipanche*). Their repeat composition also differs in the prevalence of Ty3/*gypsy* elements in *Orobanche* genomes and Ty1/*copia* elements in most analysed *Phelipanche* genomes (except for *P. purpurea*). DNA-TEs constitute about 2% of the genomes, while satellite DNAs constitute between 2%–5%, with satellite DNAs content being higher in *Orobanche*. Two outgroup taxa, *Schwalbea* and *Lindenbergia*, exhibited relatively low overall retroelement content (about 13% and 20%, respectively), thus also total repeat content, while the levels of DNA-TEs and satellite DNAs corresponded to those in their holoparasitic relatives (Piednoël & al., 2012). The higher proportions of repetitive DNA sequences in *Orobanche* and *Phelipanche* might reflect relaxed selection on genome size in parasitic organisms (Piednoël & al., 2012). *Orobanche* species have smaller genomes but higher proportions of repetitive DNA than those of *Phelipanche*, mostly due to a diversification of repeats. In the relatively recently formed tetraploid *O. gracilis*, tetraploidization may have contributed to Ty3/*gypsy* element enrichment and led to the emergence of seven large species-specific families of chromoviruses (Piednoël & al., 2012), perhaps caused by relaxed epigenetic regulation of repeats associated with polyploidy.

Solanaceae. — The genus *Nicotiana* has been used to study the genomic consequences of polyploidy. It consists of about 70 species of which 40% are documented allotetraploids derived from six independent polyploidy events (Leitch I.J. & al., 2008). It is one of most exhaustive examples of the applications of NGS for comparative analyses of repeats in diploid and related polyploid species. Tobacco (*Nicotiana tabacum*, $2n = 4x = 48$; 5.2 pg/1C) is an allopolyploid that is thought to have originated less than 200,000 years ago from diploid progenitors most closely related to *N. sylvestris* and *N. tomentosiformis* (each with $2n = 2x = 24$). Genome skimming of *N. tabacum* and its diploid relatives using 454/Roche platform (representing about 0.5% of each of the genomes) allowed comparisons of the genomes (Renny-Byfield & al., 2011). The major repeats found were Ty3/*gypsy* (17%–23%) and Ty1/*copia* (2%–3.5%) retrotransposons. However, the three genomes had experienced distinct evolutionary trajectories, with recent bursts of sequence amplification and/or homogenization in diploid maternal parent *N. sylvestris*, stability of repeat abundance in the paternal parent *N. tomentosiformis*, and genome downsizing and sequence loss across most major repeat types in *N. tabacum*, with particular losses of Ty3/*gypsy* elements and rDNA. An analysis of resynthesised *N. tabacum* from the diploid parents also revealed the loss of some repeats by the fourth generation, indicating independent, targeted, and directional loss of certain repeats from the onset of allopolyploid genome evolution (Renny-Byfield & al., 2012).

In four older allotetraploid species in *Nicotiana* sect. *Repandae*, radiating from an allopolyploidy event about 5 Ma, genome analysis showed 14%–24% genome upsizing in three species (one of which, *N. repanda*, was analysed in more detail) and 20% genome downsizing in another species (*N. nudicaulis*). Thus, despite a common origin, these allotetraploids have subsequently experienced different patterns of genome evolution. Over 85% of the repeats identified by RepeatExplorer in *N. repanda* and *N. nudicaulis* had lower abundances than expected by additivity of parental values. Genome downsizing in *N. nudicaulis* was predominantly associated with the loss of high-copy sequences, while genome expansion in *N. repanda* resulted from increase of copy numbers of repeats that have already been inherited in high copy numbers from the diploid parental taxa, particularly of CRM-like Ty3/*gypsy* elements. The loss of low-copy sequences was associated with ongoing genome diploidization in both allotetraploids, whilst differences in genome size were manifested through differential accumulation and/or deletion of high-copy-number sequences (Renny-Byfield & al., 2013).

Satellite DNAs identified using NGS for karyotype analysis

Satellite DNA makes up significant proportion of many eukaryotic genomes and consists of tandemly arranged repetitive units up to thousands of nucleotides long that can exist in millions of copies in the genome (Macas & al., 2002; Plohl & al., 2008). Higher-plant genomes might contain from a few to many families of satellite DNAs (Hemleben & al., 2007; Macas & al., 2007, 2011). Despite their genomic abundance, they are relatively poorly characterized even in extensively sequenced species, mainly due to limitations of traditional methods of analysis. However NGS and genome skimming approaches allows for efficient and comprehensive identification of satellite DNAs, regardless of monomer length, enabling comprehensive characterization of satellite DNA diversity within individual genomes and inferences to be made on evolutionary histories.

Studies using these approaches have shown that some genomes harbor many families of satellite DNAs (*Luzula elegans*: Heckmann & al., 2013; *Pisum sativum*: Macas & al., 2007; Neumann & al., 2012), whilst others contain very few (*Musa acuminata*: Čížková & al., 2013; *Prospero autumnale*: Emadzade & al., 2014; Weiss-Schneeweiss & al., unpub.). There seems to be no correlation between the number of families of satellite DNAs and their contribution to genome size. In addition, a few satellite DNA families might be highly amplified in the genome and contribute a large portion of the genome

(e.g., *Olea europea*: Barghini & al., 2014) or a small proportion of the genome might be composed of many satellite DNAs (*Pisum sativum*: Macas & al., 2007; *Luzula*: Heckmann & al., 2013). In addition, the satellite DNA repertoire of two related taxa can either vary significantly (e.g., between species of *Fritillaria*, Ambrožová & al., 2011 and cytotypes of *Prospero autumnale*, Emadzade & al., 2014) or be very similar (e.g., in genus *Musa*: Čížková & al., 2013). In *P. autumnale* complex the satellite *PaB6 is* predominantly located in pericentromeric regions of all cytotypes but its copy number and hence genome proportion varies widely (from 0.1% to 10%, Fig. 2A; Emadzade & al., 2014).

Satellite reconstruction from NGS data has also contributed to our understanding of the complex ancestry of the allopolyploid *Cardamine schulzii*. This species was thought to be an autohexaploid of the semi-fertile triploid *C.* ×*insueta* ($2n = 24$, genome designation RRA) which formed about 110–150 years ago in the Swiss Alps (Mandáková & al., 2013). *Cardamine insueta* itself is a hybrid of two diploid species (both $2n = 2x = 16$), *C. amara* (genome designation AA) and *Cardamine rivularis* (genome designation RR). However, recent studies of *C. schulzii* revealed greater complexity with the involvement of *Cardamine pratensis* (PPPP, $2n = 4x–2 = 30$), and the occurrence of *C. schulzii* at two ploidy levels, each with fewer chromosomes than expected ($2n = 5x–2 = 38$, PPRRA and $2n = 6x–2 = 46$, PPPPRA) (Mandáková & al., 2013). Low-pass 454/Roche pyrosequencing provided support through the abundances and monomer types of two satellite DNAs *Crambo* and *Prasat* (Mandáková & al., 2013). This hypothesis was further corroborated by mining NGS 454 data for parental monomer types of 35S rDNA repeats (Zozomová-Lihová & al., 2014).

Until recently the typical tandem repeat monomers in angiosperms were thought to be about 180 or 360 bps in length (Macas & al., 2002; Heslop-Harrison & Schwarzacher, 2011). However, whole-genome repeat analyses of several species clearly demonstrate that many satellite DNAs deviate from these lengths, with a nearly continuous length range from few base pairs (sex chromosomes of *Rumex acetosa*: Kejnovský & al., 2013; holocentric chromosomes of *Luzula elegans*: Heckmann & al., 2013) through tens of base pairs (e.g., *Luzula elegans*: Heckmann & al., 2013; *Silene latifolia*: Macas & al., 2011) to satellites with monomer sizes longer than 100 bp, with upper limit exceeding 1000 bp (*Nicotiana*: Renny-Byfield & al., 2012; *Solanum*: Gong & al., 2012; Zhang H. & al., 2014; *Musa*: Hřibová & al., 2010; Čížková & al., 2013; *Pisum sativum*: Neumann & al., 2012).

ChIP-seq for characterization of plant centromeric repeats

Satellite repeats are common at the centromeres of most higher eukaryotes, but the composition of centromeric regions has been analysed in only a very few plant taxa, mostly model organisms for which genomic resources are available (reviewed in Hirsch & Jiang, 2013). In addition to satellites, centromeres are sometimes accompanied by TEs, particularly chromoviruses of Ty3/*gypsy* superfamily (Neumann & al., 2011; Sharma & al., 2013). Now NGS in combination with chromatin immunoprecipitation (ChIP-seq) offers unparalleled opportunities to analyse centromere composition in a wide range of plant groups. The procedure starts with the isolation of gene(s) coding for centromere-specific histone CENH3, using, e.g., whole-transcriptome sequence data. Following successful gene identification antibodies are developed against species-specific, centromere-specific histone CENH3. These are subsequently used for ChIP, which allows enrichment of centromere-associated DNA sequences. This enriched genomic DNA fraction is sequenced and the sequences compared to NGS data obtained from control sample of whole genomic DNA of the same species (Macas & al., 2010).

ChIP-seq (using Illumina sequencing) applied to potato (*Solanum tuberosum*, $2n = 24$) has revealed surprisingly variable DNA compositions at the centromeres (Gong & al., 2012). Whilst six of twelve chromosomes contained megabase-sized satellite repeat arrays unique to individual centromeres, five chromosomes had centromeres composed primarily of single- or low-copy DNA sequences, resembling neocentromeres, and one chromosome possessed a centromere composed of both repeats and low-copy DNA sequences. At least four of the centromeric repeats were amplified from retrotransposon-related sequences and possessed very long monomers that were not detected in other closely related *Solanum* species (Zhang H. & al., 2014). Comparisons of *S. tuberosum* centromeres with those from *S. verrucosum* (a likely progenitor of cultivated potato) revealed homeologous centromeric sequences on only one chromosome. Four centromeres of *S. verrucosum* contained unique satellite repeats amplified from retrotransposon-related sequences. Further comparisons with other *Solanum* species revealed high evolutionary dynamics and rapid sequence divergence (Zhang H. & al., 2014). Some centromeres of a few *Solanum* species also contained telomere-like repeats (He & al., 2013). Interestingly, none of the potato centromeric repeats resembled the "classical" centromeric satellites with monomer sizes of 150 to 180 bp (Gong & al., 2012).

In another study to *Pisum sativum*, satellite repeats were identified that mapped to the pericentric regions of some or all chromosomes (Macas & al., 2007). The species also has remarkable extended primary constrictions

containing 3 to 5 discreet CENH3-containing regions, spanning between 69 and 107 Mbp, generating a novel "meta-polycentric" type of centromere with multiple centromere domains. These domains are composed of repetitive DNA sequences belonging to 13 distinct families of satellite DNAs and one centromeric retrotransposon. The satellite DNAs are unevenly distributed among the chromosomes, often occurring in various combinations at individual centromeres. These data point to a centromere that is determined by its epigenetic properties (e.g., folding, epigenetic marks) rather than sequence identity per se. They also indicate that functional centromere domains might not require any repeat to segregate properly and that satellites are generated in these regions following centromere establishment.

NGS studies to chromosomes of *Luzula elegans* (family Juncacaeae) have been informative in understanding chromosomes that lack a clearly defined centromere and have holokinetic kinetochores (called also diffuse or non-localized kinetochores) that are distributed along most of the poleward surface of the chromatids. To these kinetochores the spindle fibers attach (Heckmann & al., 2013). The repetitive genome fraction of *L. elegans* includes about 61% repetitive DNA, mostly comprising retrotransposons (34.3% Ty1/*copia* and 1.1% Ty3/*gypsy*) and an exceptional diversity of satellite DNAs, represented by over 30 families (9.9%), including several microsatellite types. These satellites occur as bands distributed towards the chromosome termini. None of the satellite DNAs showed a chromosome-wide distribution pattern or had associated LTR retrotransposons that may suggest centromeric sequence activity, as found in typical monocentric plant species. In addition there were no distinguishable large-scale patterns of eu- and heterochromatin specific epigenetic marks along mitotic chromosomes.

Sex chromosomes analysed using NGS

The majority of angiosperms have bisexual flowers, but about 6% of plants are dioecious with different individuals carrying either male or female flowers (Renner & Ricklefs, 1995; Ming & al., 2011). Some dioecious plants have differentiated sex chromosomes and these are best studied in *Rumex acetosa* (Polygonaceae; $2n = 14$ or 15) and *Silene latifolia* (Caryophyllaceae; $2n = 24$). Certain types of repetitive DNA have been proposed to contribute towards sex chromosome differentiation (Kejnovský & al., 2009), leading to NGS and analysis of the repetitive genome fraction in these two model systems.

Silene latifolia possesses heteromorphic sex chromosomes, with heterogametic males (XY) and homogametic females (XX). RepeatExplorer analy-

sis of 454 sequence reads allowed characterization of the major repeat types (Macas & al., 2011). These assembled repeats then became the backbone for mapping shorter but more numerous Illumina reads obtained separately for male and female individuals. About 62% of the genome is repetitive, 50% of which comprises retroelements. Of these, Ty3/*gypsy* retroelements (35%–37%) are the most abundant, with near equal proportions of Ogre and Athila family members (each 12%–13%). Ty1/*copia* retrotransposons (13.7%) were mostly represented by Angela lineage (11.8%). Satellite DNA repeats were represented by four major families amounting to 3% of the genome. However, there was little divergence in repeat composition between the sex chromosomes, consistent with a recent evolutionary origin of the sex chromosome system.

Rumex acetosa also possesses heteromorphic sex chromosomes with an XY_1Y_2 sex determination system which probably arose about 12 Ma (Navajas-Perez & al., 2005). The X chromosome is larger than individual Y chromosomes, although together the Ys are larger than the X chromosome. Earlier analyses indicated that the two heterochromatic Y chromosomes have different compositions of tandem repeats. Low-pass 454/Roche sequencing of male and female genomic DNAs allowed identification of the major repeats (in total about 60% of the genome), with Ty1/*copia* retroelements (ca. 35% Maximus/Sirevirus families) dominating over Ty3/*gypsy* superfamily (especially chromovirus and Tat/Ogre lineages, each ca. 6% of the genome). Many retrotransposons were abundant on autosomes and X chromosome, but were depleted on Y_1 and Y_2. Only one Ty1/*gypsy* subfamily had accumulated on Y_1 and Y_2 chromosomes. In contrast, most of the seven identified satellite DNA families, were most abundant on either or both Y_1 and Y_2, as were microsatellites (Steflova & al., 2013).

Supernumerary B chromosomes analysed using NGS

Plant and animal genomes might contain, in addition to regular chromosome complements, accessory B chromosomes (Houben & Carchilan, 2012). B chromosomes are dispensable, do not recombine with the chromosomes of the regular chromosome complement and can follow their own evolutionary trajectory. Little is known about the origin and molecular makeup of B chromosomes. Recent analyses of sorted B chromosomes of rye (*Secale cereale*) and sequence analysis using NGS (454/Roche platform) revealed that they are composed of A chromosome fragments, accumulated repeats shared with A chromosomes, and B-specific repeats (Martis & al., 2012). B chromosomes contain B-specific repeats as well as larger amounts of all parts of organellar

genomes, both plastid (NUPTs) and mitochondrial (NUMTs) than A chromosomes. An accumulation of B-enriched satellites was found mostly in the transcriptionally active and late-replicating nondisjunction control region of the B chromosomes and in the pericentromere region of the B chromosomes (Klemme & al., 2013).

Patterns of rDNA divergence using NGS

Both 35S and 5S rDNA sequences are frequently used as phylogenetic and chromosomal markers. However, the different domains of 35S and 5S rDNA units differ in their overall conservation between species, ranging from the rRNA genes, which are highly conserved to the intergenic spacers which can diverge sufficiently rapidly that they can be used to distinguish species (e.g., ITS—internal transcribed spacer of 35S rDNA, NTS—non transcribed spacer of 5S rDNA) or populations of a species (IGS—intergenic spacer of 35S rDNA). In addition, because the units are in multiple copies, some of which may be redundant, there can be further divergence between units of the array. The extent of intragenomic variation reflects a balance between unit mutation, selection and drift, as well as the frequency of array homogenisation in the species ancestry (e.g., Kovařík & al., 2008; Matyášek & al., 2012). Addressing the underlying biology of array homogenisation using traditional approaches is extremely challenging because of the high copy numbers of rDNA units, and consequently our understanding of the mechanisms involved is limited. The availability of NGS technology is now changing that perspective. Comparative studies of 35S rDNA repeat dynamics in four *Nicotiana* diploids using NGS (Illumina and/or 454/Roche data) suggested a relationship between rDNA unit diversity and rDNA loci number (Matyášek & al., 2012). The authors proposed that the diversity of rDNA units is best explained by higher rates of intralocus than interlocus homogenization, and that these rates exceed the rates of speciation in the genus.

Another study using 5S and 35S rDNA sequences obtained via NGS focused on elucidating the origin of the hexaploid *Helianthus tuberosus* (cultivated Jerusalem Artichoke) from its closest diploid and tetraploid relatives (Bock & al., 2014). Phylogenies based on rDNAs, as well as whole plastid and partial mitochondrial genome phylogenies, supported a hypothesis of polyploid origin of *H. tuberosus* involving autotetraploid *H. hirsutus* and diploid *H. grosseserratus*.

Phylogenetic analyses from repeats using NGS

One goal of molecular systematics is to use NGS data to infer species evolutionary relationships. Although NGS of gene-enriched DNAs allows identification of numerous low- and single-copy genes, there are significant sequencing costs and labour demands for sequence assembly, alignment and tree building. The repeat composition of the genome, analysed by genome skimming approaches requires, by contrast, relatively cheaply derived data that is more easily analysed. The data, especially from closely related taxa (within a genus) can be used for fast and effective phylogenetic analyses (Dodsworth & al., 2015). The approach uses the abundance and occurrence of repeats in the genome as continuously varying characters for that analysis. The feasibility of the approach was demonstrated in studies to six diverse groups of eukaryotes, including five angiosperms and one insect group. The inferred phylogenies provided well-resolved and supported species relationships, the patterns of branching being broadly similar to more traditionally applied sequence-based approaches (Dodsworth & al., 2015). The method is easy to implement, requiring only DNA extraction and Illumina sequencing, comparative clustering of reads (RepeatExplorer) and tree building. This methodology may prove particularly useful in groups with little genetic differentiation across commonly applied phylogenetic markers, yielding at the same time a wealth of data for detailed studies and better understanding of genome evolution in the context of repetitive DNA. It provides an important extension to molecular systematics methods and will be a useful additional tool in comparative phylogenomics.

Conclusions

Plant genomes are rich in repetitive DNAs and NGS provides a route to their study, enabling the repeat types and proportions to be determined, generating understanding of their origin, function and dynamics. The studies published so far come from only a tiny fraction of plant diversity, but even so they are illuminating. We know that there can be astonishing diversity, sometimes even between closely related taxa in repeat occurrence and abundance. Whilst very small genomes possess relatively little repetitive DNA, larger genomes have plentiful repeats organized in a myriad of combinations. Retrotransposons are mostly the dominant fraction of repeats in plants, but some genomes can be highly enriched in satellite repeats too. In the majority of plant genomes in which retroelements dominate, the ratios between Ty3/*gypsy* and Ty1/*copia*

retrotransposon can vary significantly between lineages. Some genomes are dominated by one particular type of element whilst others are composed of a very broad spectrum of repeat types.

The ability to characterize and understand the genomic diversity amongst plant species is swiftly gathering momentum, accompanied by the ever increasing frequency of published NGS data and the improving bioinformatic tools available to analyse those data. NGS provides sequence data for isolation of any number and type of DNA markers to be used in genetic diversity studies, in marker-assisted selection (e.g., development of microsatellite primers), and for the identification and reconstruction of repeats, with applications in characterizing genomes and their activities. NGS allows for the rapid generation of datasets for multiple plant lineages, which in turn allows understanding of mechanisms contributing the dynamic nature of the repeat repertoire.

Acknowledgements

The authors acknowledge financial support of Austrian Science Fund (FWF; projects P21440 and P25131 to HW-S) and Czech Science Foundation and Czech Academy of Sciences (projects GBP501/12/G090 and RVO:60077344 to JM).

Literature cited

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I.J., Koblížková, A., Macas, J. & Lysak, M.A. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Ann. Bot. (Oxford)* 107: 255–268. http://dx.doi.org/10.1093/aob/mcq235

Barghini, E., Natali, L., Cossu, R.M., Giordani, T., Pindo, M., Cattonaro, F., Scalabrin, S., Velasco, R., Morgante, M. & Cavallini, A. 2014. The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. *Genome Biol. Evol.* 6: 776–791. http://dx.doi.org/10.1093/gbe/evu058

Ben-David, S., Yaakov, B. & Kashkush, K. 2013. Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat. *Plant J.* 76: 201–210. http://dx.doi.org/10.1111/tpj.12285

Bock, D.G., Kane, N.C., Ebert, D.P. & Rieseberg, L.H. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: Neither from Jerusalem nor an artichoke. *New Phytol.* 201: 1021–1030. http://dx.doi.org/10.1111/nph.12560

Britten, R.J. & Kohne, D.E. 1968. Repeated sequences in DNA. *Science* 161: 529–540. http://dx.doi.org/10.1126/science.161.3841.529

Čížková, J., Hřibová, E., Humplíková, L., Christelová, P., Suchánková, P. & Doležel, J. 2013. Molecular analysis and genomic organization of major DNA satellites in banana (*Musa* spp.). *PLOS ONE* 8: e54808. http://dx.doi.org/10.1371/journal.pone.0054808

DeBarry, J.D., Liu, R. & Bennetzen, J.L. 2008. Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. *B. M. C. Bioinf.* 9: 235. http://dx.doi.org/10.1186/1471-2105-9-235.

Dodsworth, S., Chase, M.W., Kelly, L.J., Leitch, I.J., Macas, J., Novák, P., Piednoël, M., Weiss-Schneeweiss, H. & Leitch, A.R. 2015. Genomic repeat abundances contain phylogenetic signal in diverse eukaryotic groups. *Syst. Biol.* 64: 112–126. http://dx.doi.org/10.1093/sysbio/syu080

Ellegren, H. 2004. Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* 5: 435–445. http://dx.doi.org/10.1038/nrg1348

Emadzade, K., Jang, T.S., Macas, J., Kovařík, A., Novák, P., Parker, J. & Weiss-Schneeweiss, H. 2014. Differential amplification of satellite *PaB6* in chromosomally hypervariable *Prospero autumnale* complex (Hyacinthaceae). *Ann. Bot. (Oxford)* 114: 1597–1608. http://dx.doi.org/10.1093/aob/mcu178

Estep, M.C., Estep, M.C., DeBarry, J.D. & Bennetzen, J.L. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity* 110: 194–204. http://dx.doi.org/10.1038/hdy.2012.99

Flavell, R.B. 1986. Repetitive DNA and chromosome evolution in plants. *Philos. Trans., Ser. B* 312: 227–242. http://dx.doi.org/10.1098/rstb.1986.0004

Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLOS ONE* 6: e16526. http://dx.doi.org/10.1371/journal.pone.0016526

Gong, Z., Wu, Y., Koblížková, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C.R., Macas, J. & Jiang, J. 2012. Repeatless and repeat-based centromeres in potato: Implications for centromere evolution. *Pl. Cell* 24: 3559–3574. http://dx.doi.org/10.1105/tpc.112.100511

Greilhuber, J. & Leitch, I.J. 2013. Genome size and the phenotype. Pp. 323–344 in: Leitch, I.J., Greilhuber, J., Doležel, J. & Wendel, J.F. (eds.), *Plant genome diversity*, vol. 2, *Physical structure, behaviour and evolution of plant genomes*. Vienna: Springer. http://dx.doi.org/10.1007/978-3-7091-1160-4_20

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S. & Barthlott, W. 2006. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Pl. Biol.* 8: 770–777. http://dx.doi.org/10.1055/s-2006-924101

He, L., Liu, J., Torres, G.A., Zhang, H.Q., Jiang, J.M. & Xie, C.H. 2013. Interstitial telomeric repeats are enriched in the centromeres of chromosomes in *Solanum* species. *Chromosome Res.* 21: 5–13. http://dx.doi.org/10.1007/s10577-012-9332-x

Heckmann, S., Macas, J., Kumke, K., Fuchs, J., Schubert, V., Ma, L., Novák, P., Neumann, P., Taudien, S., Platzer, M. & Houben, A. 2013. The holocentric species of *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant J.* 73: 555–565. http://dx.doi.org/10.1111/tpj.12054

Hemleben, V., Kovařík, A., Torres-Ruiz, R.A., Volkov, R.A. & Beridze, T. 2007. Plant highly repeated satellite DNA: Molecular evolution, distribution and use for identification of hybrids. *Syst. Biodivers.* 5: 277–289. http://dx.doi.org/10.1017/S147720000700240X

Heslop-Harrison, J.S. & Schmidt, T. 1998. Genomes, genes and junk: The large-scale organization of plant genomes. *Trends Pl. Sci.* 3: 195–199. http://dx.doi.org/10.1016/S1360-1385(98)01223-0

Heslop-Harrison, J.S. & Schwarzacher, T. 2011. Organisation of the plant genome in chromosomes. *Plant J.* 66: 18–33. http://dx.doi.org/10.1111/j.1365-313X.2011.04544.x

Hirsch, C.D. & Jiang, J. 2013. Centromeres: Sequences, structure, and biology. Pp. 59–70 in: Wendel, J., Greilhuber, J., Dolezel, J. & Leitch, I.J. (eds.), *Plant genome diversity*, vol. 1, *Plant genomes, their residents, and their evolutionary dynamics*. Vienna: Springer. http://dx.doi.org/10.1007/978-3-7091-1130-7_4

Houben, A. & Carchilan, M. 2012. Plant B chromosomes: What makes them different? Pp. 59–77 in: Bass, H.W. & Birchler, J.A. (eds.), *Plant cytogenetics*. Vienna: Springer. http://dx.doi.org/10.1007/978-0-387-70869-0_1

Hřibová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J. & Doležel, J. 2010. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *B. M. C. Pl. Biol.* 10: 204. http://dx.doi.org/10.1186/1471-2229-10-204

Huang, S., Li, R., Zhang Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, Y., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., Van der Vossen, E.A.G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W.K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y. & Li, S. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41: 1275–1281. http://dx.doi.org/10.1038/ng.475

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S.C., Himmelbauer, H., Minoche, A.E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Pérez, S.A., Ortega-Estrada, M.J., Cervantes-Luevano, J.I., Michael, T.P., Mockler, T., Bryant, D., Herrera-Esterlla, A., Albert, V.A. & Herrera-Esterella, L. 2013. Architecture and evolution of a minute plant genome. *Nature* 498: 94–98. http://dx.doi.org/10.1038/nature12132

Kapitonov, V.V. & Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 98: 8714–8719. http://dx.doi.org/10.1073/pnas.151269298

Kejnovský, E., Hobza, R., Cermak, T., Kubat, Z. & Vyskot, B. 2009. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity* 102: 533–541. http://dx.doi.org/10.1038/hdy.2009.17

Kejnovský, E., Hawkins, J. & Feschotte, C. 2012. Plant transposable elements: Biology and evolution. Pp. 17–34 in: Wendel, J., Greilhuber, J., Dolezel, J. & Leitch, I.J. (eds.), *Plant genome diversity*, vol. 1, *Plant genomes, their residents, and their evolutionary dynamics.* Vienna: Springer. http://dx.doi.org/10.1007/978-3-7091-1130-7_2

Kejnovský, E., Michalovova, M., Steflova, P., Kejnovska, I., Manzano, S., Hobza, R., Kubat, Z., Kovarik, J., Jamilena, M. & Vyskot, B. 2013. Expansion of microsatellites on evolutionary young Y chromosome. *PLoS ONE* 8: e45519. http://dx.doi.org/10.1371/journal.pone.0045519

Kelly, L.J. & Leitch, I.J. 2011. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Res.* 19: 939–953. http://dx.doi.org/10.1007/s10577-011-9246-z

Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysák, M., Day, P.D., Berger, M., Fay, M.F., Nichols, R.A., Leitch, A.R. & Leitch, I.J. 2015. Analysis of

the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol.* (early view). http://dx.doi.org/10.1111/nph.13471

Klemme, S., Banaei-Moghaddam, A.M., Macas, J., Wicker, T., Novák, P. & Houben, A. 2013. High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytol.* 199: 550–558. http://dx.doi.org/10.1111/nph.12289

Knoll, A., Fauser, F. & Puchta, H. 2014. DNA recombination in somatic plant cells: Mechanisms and evolutionary consequences. *Chromosome Res.* 22: 191–201. http://dx.doi.org/10.1007/s10577-014-9415-y

Kovach, A., Wegrzyn, J., Parra, G., Holt, C., Bruening, G., Loopstra, C., Hartigan, J., Yandell, M., Langley, C., Korf, I. & Neale, D.B. 2010. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *B. M. C. Genomics* 11: 420. http://dx.doi.org/10.1186/1471-2164-11-420

Kovařík, A., Dadejová, M., Lim, Y.K., Chase, M.W., Clarkson, J.J., Knapp, S. & Leitch, A.R. 2008. Evolution of rDNA in *Nicotiana* allopolyploids: A potential link between rDNA homogenization and epigenetics. *Ann. Bot. (Oxford)* 101: 815–823. http://dx.doi.org/10.1093/aob/mcn019

Kumar, A. & Bennetzen, J.L. 1999. Plant retrotransposons. *Annual Rev. Genet*. 33: 479–532. http://dx.doi.org/10.1146/annurev.genet.33.1.479

Kumar, A. & Hirochika, H. 2001. Applications of retrotransposons as genetic tools in plant biology. *Trends Pl. Sci.* 6: 127–134. http://dx.doi.org/10.1016/S1360-1385(00)01860-4

Kurtz, S. & Schleiermacher, C. 1999. REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* 15: 426–427. http://dx.doi.org/10.1093/bioinformatics/15.5.426

Leitch, A.R. & Leitch, I.J. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* 194: 629–646. http://dx.doi.org/10.1111/j.1469-8137.2012.04105.x

Leitch, I.J. & Leitch, A.R. 2013. Genome size diversity and evolution in land plants. Pp. 307–322 in: Leitch, I.J., Greilhuber, J., Doležel, J. & Wendel, J.F. (eds.), *Plant genome diversity*, vol. 2, *Physical structure, behaviour and evolution of plant genomes.* Vienna: Springer. http://dx.doi.org/10.1007/978-3-7091-1160-4_19

Leitch, I.J., Hanson, L., Lim, K.Y., Kovařík, A., Chase, M.W., Clarkson, J.J. & Leitch, A.R. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot. (Oxford)* 101: 805–814. http://dx.doi.org/10.1093/aob/mcm326

Macas, J. & Neumann, P. 2007. *Ogre* elements: A distinct group of plant Ty3/*gypsy*-like retrotransposons. *Gene* 390: 108–116. http://dx.doi.org/10.1016/j.gene.2006.08.007

Macas, J., Meszaros, T. & Nouzová, M. 2002. PlantSat: A specialized database for plant satellite repeats. *Bioinformatics* 18: 28–35. http://dx.doi.org/10.1093/bioinformatics/18.1.28

Macas, J., Neumann, P. & Navrátilová, A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *B. M. C. Genomics* 8: 427. http://dx.doi.org/10.1186/1471-2164-8-427

Macas, J., Neumann, P., Novák, P. & Jiang, J. 2010. Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. *Bioinformatics* 26: 2101–2108. http://dx.doi.org/10.1093/bioinformatics/btq343

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A. & Vyskot, B. 2011. Next generation sequencing-based analysis of repetitive DNA in the model dioceous plant *Silene latifolia*. *PLOS ONE* 6: e27335. http://dx.doi.org/10.1371/journal.pone.0027335

Mandáková, T., Kovařík, A., Zozomová-Lihová, J., Shimizu-Inatsugi, R., Shimizu, K.K., Mummenhoff, K., Marhold, K. & Lysak, M.A. 2013. The more the merrier: Recent hybridization and polyploidy in *Cardamine*. *Pl. Cell* 25: 3280–3295. http://dx.doi.org/10.1105/tpc.113.114405

Martis, M.M., Klemme, S., Banaei-Moghaddam, A.M., Blattner, F.R., Macas, J., Schmutzer, T., Scholz, U., Gundlach, H., Wicker, T., Šimková, H., Novák, P., Neumann, P., Kubaláková, M., Bauer, E., Haseneyer, G., Fuchs, J., Doležel, J., Stein, N., Mayer, K.F.X. & Houben, A. 2012. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc. Natl. Acad. Sci. U.S.A.* 109: 13343–13346. http://dx.doi.org/10.1073/pnas.1204237109

Matyášek, R., Renny-Byfield, S., Fulneček, J., Macas, J., Grandbastien, M.-A., Nichols, R., Leitch, A.R. & Kovařík, A. 2012. Next generation sequencing analysis reveals a relationship between rDNA unit diversity and locus number in *Nicotiana* diploids. *B. M. C. Genomics* 13: 722. http://dx.doi.org/10.1186/1471-2164-13-722

Ming, R., Bendahmane, A. & Renner, S. 2011. Sex chromosomes in land plants. *Annual Rev. Pl. Biol.* 62: 485–514. http://dx.doi.org/10.1146/annurev-arplant-042110-103914

Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., Kubisiak, T.L., Amerson, H.V., Carlson, J.E., Nelson, C.D. & Davis, J.M. 2009. Evolution of genome size and complexity in *Pinus*. *PLOS ONE* 4: e4332. http://dx.doi.org/10.1371/journal.pone.0004332

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg, L. & Cavallini, A. 2013. The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *B. M. C. Genomics* 14: 686. http://dx.doi.org/10.1186/1471-2164-14-686

Navajas-Pérez, R., De la Herran, R., Jamilena, M., Lozano, R., Rejón, C.R., Rejón, M.R. & Garrido-Ramos, M.A. 2005. Reduced rates of sequence evolution of Y-linked satellite DNA in *Rumex* (Polygonaceae). *J. Molec. Evol.* 60: 391–399. http://dx.doi.org/10.1007/s00239-004-0199-0

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. & Macas, J. 2011. Plant centromeric retrotransposons: A structural and cytogenetic perspective. *Mobile DNA* 2: 4. http://dx.doi.org/10.1186/1759-8753-2-4

Neumann, P., Navrátilová, A., Schroeder-Reiter, E., Koblížková, A., Steinbauerová, V., Chocholová, E., Novák, P., Wanner, G. & Macas, J. 2012. Stretching the rules: Monocentric chromosomes with multiple centromere domains. *PLoS Genet.* 8: e1002777. http://dx.doi.org/10.1371/journal.pgen.1002777

Novák, P., Neumann, P. & Macas, J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *B. M. C. Bioinf.* 11: 378. http://dx.doi.org/10.1186/1471-2105-11-378

Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. 2013. RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* 29: 792–793. http://dx.doi.org/10.1093/bioinformatics/btt054

Novák, P., Hřibová, E., Neumann, P., Koblížková, A., Doležel, J. & Macas, J. 2014. Genome-wide analysis of repeat diversity across the family Musaceae. *PLOS ONE* 9: e98918. http://dx.doi.org/10.1371/journal.pone.0098918

Nystedt, B., Street, N.R, Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, A., Rilakovic, N., Ritland, C., Rosselló, J.A., Sena, J., Svensson, T., Talavera-

López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T.R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Thompson, S.L., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P.K., Lundeberg, J. & Jansson, S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497: 579–584. http://dx.doi.org/10.1038/nature12211

Park, J.-M., Schneeweiss, G.M., Weiss-Schneeweiss, H. 2007. Diversity and evolution of Ty1-*copia* and Ty3-*gypsy* retroelements in the non-photosynthetic flowering plants *Orobanche* and *Phelipanche* (Orobanchaceae). *Gene* 387: 75–86. http://dx.doi.org/10.1016/j.gene.2006.08.012

Park, M., Park, J., Kim, S., Kwon, J.-K., Park, H.M., Bae, I.H., Yang, T.-J., Lee, Y.-H., Kang, B.-C. & Choi, D. 2012. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *Plant J.* 69: 1018–1029. http://dx.doi.org/10.1111/j.1365-313X.2011.04851.x

Pellicer, J., Fay, M.F. & Leitch, I.J. 2010. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* 164: 10–15. http://dx.doi.org/10.1111/j.1095-8339.2010.01072.x

Peška, V., Fajkus, P., Fojtová, M., Dvořáčková, M., Hapala, J., Dvořáček, V., Polanská, P., Leitch, A.R., Sýkorová, E. & Fajkus, J. 2015. Characterisation of an unusual telomere motif $(TTTTTTAGGG)_n$ in the plant *Cestrum elegans* (Solanaceae), a species with a large genome. *Plant J.* 82: 644–654. http://dx.doi.org/10.1111/tpj.12839

Petit, M., Lim, K.Y., Julio, E., Poncet, C., Dorlhac de Borne, F., Kovařík, A., Leitch, A.R., Grandbastien, M.A. & Mhiri, C. 2007. Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). *Molec. Genet. Genomics* 278: 1–15. http://dx.doi.org/10.1007/s00438-007-0226-0

Piednoël, M., Aberer, A.J., Schneeweiss, G.M., Macas, J., Novák, P., Gundlach, H., Temsch, E.M. & Renner, S.S. 2012. Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Molec. Biol. Evol.* 29: 3601–3611. http://dx.doi.org/10.1093/molbev/mss168

Piednoël, M., Carrete-Vega, G. & Renner, S.S. 2013. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. *Plant J.* 75: 699–709. http://dx.doi.org/10.1111/tpj.12233

Plohl, M., Luchetti, A., Mestrović, N. & Mantovani, B. 2008. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* 409: 72–82. http://dx.doi.org/10.1016/j.gene.2007.11.013

Renner, S.S. & Ricklefs, R.E. 1995. Dioecy and its correlates in the flowering plants. *Amer. J. Bot.* 82: 596–606.

Renny-Byfield, S., Chester, M., Kovařík, A., Le Comber, S.C., Grandbastien, M.-A., Deloger, M., Nichols, R.A., Macas, J., Novák, P., Chase, M.W. & Leitch, A.R. 2011. Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Molec. Biol. Evol.* 28: 2843–2854. http://dx.doi.org/10.1093/molbev/msr112

Renny-Byfield, S., Kovařík, A., Chester, M., Nichols, R.A., Macas, J., Novák, P. & Leitch, A.R. 2012. Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of *Nicotiana tabacum*. *PLoS ONE* 7: e36963. http://dx.doi.org/10.1371/journal.pone.0036963

Renny-Byfield, S., Kovařík, A., Kelly, L.J., Macas, J., Novák, P., Chase, M.W., Nichols, R.A., Pancholi, M.R., Grandbastien, M.-A. & Leitch, A.R. 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* 74: 829–839. http://dx.doi.org/10.1111/tpj.12168

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. & Wilson, R.K. 2009. The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326: 1112–1115. http://dx.doi.org/10.1126/science.1178534

Schneeweiss, G.M., Palomeque, T., Colwell, A.E. & Weiss-Schneeweiss, H. 2004. Chromosome numbers and karyotype evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. *Amer. J. Bot.* 91: 439–448. http://dx.doi.org/10.3732/ajb.91.3.439

Senerchia, N., Wicker, T., Felber, F. & Parisod, C. 2013. Evolutionary dynamics of retrotransposons assessed by high-throughput sequencing in wild relatives of wheat. *Genome Biol. Evol.* 5: 1010–1020. http://dx.doi.org/10.1093/gbe/evt064

Sharma, A., Wolfgruber, T.K. & Presting, G.G. 2013. Tandem repeats derived from centromeric retrotransposons. *B. M. C. Genomics* 14: 142. http://dx.doi.org/10.1186/1471-2164-14-142

Staton, S.E., Bakken, B.H., Blackman, B.K., Chapman, M.A., Kane, N.C., Tang, S., Ungerer, M.C., Knapp, S.J., Rieseberg, L.H. & Burke, J.M. 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. *Plant J.* 72: 142–153. http://dx.doi.org/10.1111/j.1365-313X.2012.05072.x

Steflova, P., Tokan, V., Vogel, I., Lexa, M., Macas, J., Novák, P., Hobza, R., Vyskot, B. & Kejnovský, E. 2013. Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY_1Y_2) in the dioecious plant *Rumex acetosa*. *Genome Biol. Evol.* 5: 769–782. http://dx.doi.org/10.1093/gbe/evt049

Swaminathan, K., Varala, K. & Hudson, M.E. 2007. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *B. M. C. Genomics* 8: 132. http://dx.doi.org/10.1186/1471-2164-8-132

Tenaillon, M.I., Hufford, M.B., Gaut, B.S. & Ross-Ibarra, J. 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* 3: 219–229. http://dx.doi.org/10.1093/gbe/evr008

Volfovsky, N., Haas, B.J. & Salzberg, S.L. 2001. A clustering method for repeat analysis in DNA sequences. *Genome Biol.* 2: research0027-research0027.11. http://dx.doi.org/10.1186/gb-2001-2-8-research0027

Weiss-Schneeweiss, H. & Schneeweiss, G.M. 2013. Karyotype diversity and evolutionary trends in angiosperms. Pp. 209–230 in: Leitch, I.J., Greilhuber, J., Dolezel, J. & Wendel, J.F. (eds.), *Plant genome diversity*, vol. 2, *Physical structure and evolution of plant genomes*. Vienna: Springer. http://dx.doi.org/10.1007/978-3-7091-1160-4_13

West, P.T., Li, Q., Ji, L., Eichten, S.R., Song, J., Vaughn, M.W., Schmitz, R.J. & Springer, N.M. 2014. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLOS ONE* 9: e105267. http://dx.doi.org/10.1371/journal.pone.0105267

Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B. & Stein, N. 2006. 454 sequencing put to the test using the complex genome of barley. *B. M. C. Genomics* 7: 275. http://dx.doi.org/10.1186/1471-2164-7-275

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T.H., Graner, A., Ware, D. & Stein, N. 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *B. M. C. Genomics* 9: 518. http://dx.doi.org/10.1186/1471-2164-9-518

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. & Stein, N. 2009. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* 59: 712–722. http://dx.doi.org/10.1111/j.1365-313X.2009.03911.x

Zhang, H., Koblížková, A., Wang, K., Gong, Z., Oliveira, L., Torres, G.A., Wu, Y., Zhang, W., Novák, P., Buell, C.R., Macas, J. & Jiang, J. 2014. Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: Rapid evolution of DNA sequences associated with centromeres. *Pl. Cell* 26: 1436–1447. http://dx.doi.org/10.1105/tpc.114.123877

Zhang, J., Zuo, T. & Peterson, T. 2013. Generation of tandem direct duplications by reversed ends transposition of maize Ac elements. *PLoS Genet*. 9: e1003691. http://dx.doi.org/10.1371/journal.pgen.1003691

Zozomová-Lihová, J., Mandáková, T., Kovaříková, A., Mühlhausen, A., Mummenhoff, K., Lysak, M.A. & Kovařík, A. 2014. When fathers are instant losers: Homogenization of rDNA loci in recently formed *Cardamine schulzii* trigenomic allopolyploid. *New Phytol.* 203: 1096–1108. http://dx.doi.org/10.1111/nph.12873