HAL
open science

The Repetitive Content in Lupin Genomes

Abdelkader Ainouche, Aurore Paris, Delphine Giraud, Jean Keller, Pauline Raimondeau, Frédéric Mahe, Pavel Neuman, Petr Novak, Jiri Macas, Lily Ainouche Malika, et al.

To cite this version:

Abdelkader Ainouche, Aurore Paris, Delphine Giraud, Jean Keller, Pauline Raimondeau, et al.. The Repetitive Content in Lupin Genomes. Singh Karam B. (ed.); Kamphuis Lars G. (ed.); Nelson Matthew N. (ed.). The Lupin Genome, Springer, pp.161-186, 2020, 978-3-030-21270-4; 978-3-030-21269-8. 10.1007/978-3-030-21270-4_12 . hal-02879117

HAL Id: hal-02879117

https://univ-rennes.hal.science/hal-02879117v1

Submitted on 23 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

12 The repetitive content in lupin genomes

Abdelkader Aïnouche[1], Aurore Paris[1], Delphine Giraud[1], Jean Keller[1,2], Pauline Raimondeau[1], Frédéric Mahé[3], Pavel Neuman[4], Petr Novak[4], Jiri Macas[4], Malika Aïnouche[1], Armel Salmon[1], Guillaume E. Martin[5].

Corresponding author: kader.ainouche@univ-rennes1.fr

[1] UMR CNRS 6553 ECOBIO, Université de Rennes 1, 35042 Rennes cedex, France.

[2] Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 24 chemin de Borde Rouge, Auzeville, BP42617, 31326 Castanet-Tolosan, France

[3] CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), Campus international de Baillarguet, 34398 Montpellier Cedex 5, France

[4] Biology Centre of the Czech Academy of Sciences, Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Branisovska 31, Ceske Budejovice, CZ-37005, Czech Republic

[5] CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France

Accepted Manuscript

Accepted Manuscript

Abstract In this chapter, we present the first detailed evaluation of the repetitive compartment in *Lupinus* genomes. Low-depth next generation sequencing (NGS) genomic resources from four closely related smooth-seeded Mediterranean lupin species (*L. albus*, *L. angustifolius*, *L. luteus* and *L. micranthus*), exhibiting remarkable differences in genome size and chromosome number have been investigated. The repetitive compartment is composed of a wide diversity of repeats and represents 23 to 51 % of the genomes. This compartment is essentially comprised of transposable elements (43 to 85%), mainly represented by *copia* and *gypsy* LTR retrotransposon families. Among the latter, some prominent families (*Tekay*, *Athila*, *Maximus-SIRE*) significantly contribute to genome size differences among species and in shaping different species-specific repeat profiles, regardless of their chromosome numbers. Also particular lineages of these elements have been differentially and recently amplified within species, such as in *L. luteus*, *L. albus* and *L. angustifolius*. Moreover, this study highlighted the diversity of tandem repeats in lupin genomes, with minisatellites and satellites mostly being species-specific, whereas microsatellites (SSRs) are ubiquitously distributed. Strikingly, *L. angustifolius* exhibited a tremendous amount of tandem repeats in its genome (26%), including a noteworthy accumulation of one particular hexamer SSR (15.24% of the genome), which demonstrate that also tandem repeats may greatly contribute to genome obesity and dynamics in lupins. Therefore, differential lineage-specific amplifications of retrotransposons and tandem repeats occurred among lupins. Accordingly, this strongly

Accepted Manuscript

suggests that different processes and mechanisms regulating amplification, proliferation and clearance of repeats have differentially operated within the same genus among closely related Mediterranean species over the last ~10-12 Myr. Further extension of such evaluation to various representatives of the lupins diversity and outgroups will provide a better overview of the repetitive compartment and its evolutionary dynamics in the genus. Additionally, the genomic resources generated by this work represent a valuable basis to start building a repeats database specifically dedicated to best understand the genomic landscape, repeats distribution and localization in lupins. This will facilitate further investigations on the functional and evolutionary impact of repeats on genes of interest, such as, those responsive for important agronomical, adaptive and defense features.

Accepted Manuscript

13.1 INTRODUCTION

Genome size variation, with a magnitude order of 200,000, is one of the most remarkable biological features in Eucaryotes (Bennetzen and Wang, 2014; Biscotti *et al*., 2015). As this variation is not correlated with the morphological or physiological complexity of organisms it has been termed the "C-value paradox" (Thomas, 1971) and later the "C-value enigma" (Gregory, 2005). In angiosperms, genome size (GS) ranges from 60 Mb (in *Genlisea aurea*) to 150 Gb (in *Paris japonica*), which corresponds to approximately a 2,400-fold variation (Greilhuber *et al.*, 2006; Leitch & Leitch, 2008; Vu *et al.*, 2015; Jaume Pellicer *et al.*, 2018). Moreover, GS variation occurs at various taxonomic levels, including among closely related species within genera (Greilhuber *et al.*, 2006; Hu *et al.*, 2011) or even among various accessions within species, such as in *Hordeum spontaneum* (Kalendar *et al.*, 2000) or in *Arabidopsis thaliana* (Schmuths, 2004). Apart from whole-genome duplication, triplication, or polyploidy (Soltis *et al.*, 2009; Renny-Byfield & Wendel, 2014), it is now obvious that repetitive sequences may account for a large proportion in the plant genomes, regardless of the number of protein coding genes, the ploidy level or the past paleopolyploid history (Bennetzen, 2002, 2005; Wendel *et al.*, 2016). While, the repetitive sequences were previously regarded as "junk", "parasitic" or "selfish" DNA (Doolittle & Sapienza, 1980; Orgel *et al.*, 1980; Lönnig & Saedler, 1997), nowadays they are not only considered as a determinant fraction involved in GS variation (expansion / contrac-

Accepted Manuscript

tion), but also that they play a major role in their evolutionary dynamics and are crucial for living organisms (Biémont & Vieira, 2006; Oliver *et al.*, 2013; Wendel *et al.*, 2016; Hosaka & Kakutani, 2018; Pellicer *et al.*, 2018). Two types of repetitive DNA sequences proliferate in the genomes: tandem repeats (or satellites *sensu lato*) and interspersed repeats (or transposable elements).

13.1.1 Tandem repeats

Tandem repeats (TR) consist of basic nucleotide units (or monomers) that are repeated head-to-tail to form TR arrays. According to the size of the repeated unit the tandem repeats are classified as: microsatellites or SSR (Simple Sequence Repeats) with motives shorter than 10-12 bp, minisatellites with motives between 12 to ~60 bp, and satellites with longer monomers (>60 to 100 bp or even several kilobases). Increase (or decrease) of the number of repeated units in microsatellites, for instance, generally results from a "slipped-strand mispairing" mechanism due to a polymerase shift during DNA replication (Levinson & Gutman, 1987) or unequal crossovers (Petes, 1980). Satellite DNA can represent up to half of the genome in some eukaryotes (Satović *et al.*, 2018). Microsatellites (SSRs) are ubiquitous in genomes and are widely used as genetic markers for genotyping (Parra-González *et al*., 2012;.Raman *et al*., 2014; Kamphuis *et al*., 2015; Atnaf *et al*., 2017). The other larger arrays of TR, minisatellites, satelittes, including highly repetitive gene families such as nuclear ribosomal DNA, are helpful for chromosome fingerprinting. They are usually associated to centromeric, peri-

Accepted Manuscript

centromeric and telomeric regions and seem to have a significant functional regulatory role (Streelman & Kocher, 2002; Li, 2004; Lower *et al.*, 2018), but they yet remain poorly investigated and were the subject of only few comparative genomics studies (Shi *et al.*, 2013; Ruiz-Ruano *et al.*, 2016; Usai *et al.*, 2017).

13.1.2 Transposable elements

Transposable elements (TEs) are very diverse interspersed repetitive DNA sequences (or jumping genetic elements of B. McClintock, 1948) able to duplicate themselves and to insert their copies at different positions in the genome *via* a transposition mechanism (Kumar & Bennetzen, 1999; Bennetzen, 2002). Following the classification of (Wicker *et al.*, 2007), TEs are divided into two main classes, according to the type of intermediate (DNA or RNA) used in their transposition mode. Class I elements, or retrotransposons, follow a transposition mode using an RNA intermediate called "copy / paste", which may dramatically increase their copy number in genomes (Vicient *et al.*, 1999; Bennetzen, 2002, 2005; Piegu *et al.*, 2006). Five orders are distinguished within this class: LTR elements (Long Terminal Repeats) DIRS elements (Dictyostelium intermediate repeat sequence), PLEs (Penelope-like elements), LINEs (Long Interspersed Nuclear elements) and SINEs (Small Interspersed Nuclear Element). Within each order, elements are clustered into superfamilies based on the structure of their protein and non-coding domains. The Class II elements (or DNA transposons) transpose *via* a DNA intermediate in a mode called "cut / paste", which results in

Accepted Manuscript

their excision from their genomic location and their insertion elsewhere in the genome. Two subclasses are recognized: subclass 1 mainly correspond to TIR elements, which are characterized by their Terminal Inverted Repeats (TIR) at their extremities; and subclass 2 which correspond to Helitron and Maverick elements (Wicker *et al.*, 2007).

In plants, amplification and accumulation of Class I elements represent the major source of GS increase. For example, LTR-retrotransposons may reach between ~70 to 76% of the genomes in maize, bread wheat and barley (Mayer *et al.*, 2011; Oliver *et al.*, 2013; Wicker *et al.*, 2018). TEs amplification can be activated by various environmental (biotic and abiotic) and genomic (e.g., hybridizations) stresses during the evolutionary history of organisms (Kalendar *et al.*, 2000; Liu & Wendel, 2000; Jiang *et al.*, 2003; Grandbastien *et al.*, 2005; Wessler, 2006). In turn, different regulatory mechanisms are triggered at the cellular and molecular levels to control their proliferation and counteract genome expansion *via* epigenetic mechanisms (small RNA, DNA methylation, histone modification) and removal (Bennetzen, 2005; Hawkins *et al.*, 2006, 2009; Slotkin & Martienssen, 2007; Lisch, 2009; Yaakov & Kashkush, 2012; Axtell, 2013; Castel & Martienssen, 2013). The repeated waves of TE amplification and regulatory mechanisms thus have a deep impact on the host genomes. They may drive structural genomic rearrangements and generate genetic diversity which accompanies the adaptation and diversification of species in their environments (Bennett, 2005; Morgante *et al.*, 2005; Chénais *et al.*, 2012).

Accepted Manuscript

Following their insertion into or near genes, they may modify expression and function of various genes which may induce variable phenotypic changes (Jiang *et al.*, 2003; Kashkush *et al.*, 2003; Lisch, 2013). Also there is evidence that they contribute to the formation of new genes and represent an important source of evolutionary novelties (Biémont & Vieira, 2006; Oliver *et al.*, 2013; Lynch *et al.*, 2015).

13.1.3 Advancing the discovery of repetitive sequences using Next Generation Sequencing technology

Regarding their importance, investigations on the repetitive sequences greatly benefited from the advances of high throughput sequencing technologies. Several strategies and bioinformatics programs have been developed for the detection and identification of repeated elements in fully sequenced genomes of model organisms (Quesneville *et al.*, 2005; Lerat, 2010; Flutre *et al.*, 2011; Treangen & Salzberg, 2011; Wajid & Serpedin, 2012). However, assembly, annotation and precise location of massive similar repeated short-reads, representing regions which underwent various processes of recombination/deletion, is challenging and generally results in incompletely assembled genomes with large gap-spaces and potentially chimerical structures (Jiang *et al.*, 2004; Sequencing Project, 2005). Combination of short-reads technologies (Illumina HiSeq) with long-reads sequencing ones (Pacific BioSiences and Oxford Nanopore) will yield higher quality genomes to accurately assemble and circumscribe repeated structures. Other programs have been de-

Accepted Manuscript

signed to directly evaluate the repetitive content from raw unassembled short read sequences generated from various high-throughput DNA sequencing technology platforms, such as for example: RepeatExplorer (Novák *et al*., 2010; Novak *et al*., 2013), Transposome (Staton, Burke, 2015), REPdenovo (Chu *et al*., 2016). Such programs use various tools which allow detection, quantitative estimation, reconstruction and annotation of repetitive elements in NGS data. They are based on all-to-all read sequence similarities, graph-based clustering methods, and repeats identification using complementary Blast methods and search of conserved specific TEs protein coding domains against reference databases. These toolkits demonstrated their efficiency for evaluating the repetitive compartment from a reduced sample of low-pass genome sequence data (even less than 1% genome coverage) in various plant taxa (Macas *et al*., 2007; Hřibová *et al*., 2010; Novák *et al*., 2010, 2013; Renny-Byfield *et al*., 2011; Piednoël *et al*., 2013; Staton, Burke, 2015; Vu *et al*., 2015; Wu *et al*., 2019). They not only allow rapid investigation of the repetitive compartment in many non-model genomes, but also may provide crucial information to assist assembly and annotation of complex genomes.

13.1.4 *Lupinus*: A system of interest to evaluate the dynamics of the genomic repetitive compartment

In this context, the genistoid legume *Lupinus* (Fabaceae) is a system of particular interest to explore the evolutionary dynamics of repetitive sequences and their impact on the evolution of host genomes.

Accepted Manuscript

Indeed, *Lupinus* is a large genus which is composed of hundreds of species adapted to very divers ecological conditions which diversified during the last ~16 Myr (mean age of the stem node of the genus according to Hughes and Eastwood, 2006) in two major regions of the World: about 250-300 species in the New World and around 20 in the Old World (Gladstones *et al.*, 1998; Ainouche *et al.*, 2004; Eastwood *et al.*, 2008). Among the latter, the smooth-seeded lupins (or *Malacosperma*), which are mainly circum-Mediterranean are distinguished from the rough-seeded lupins (or *Scabrispermae*), which are predominantly North African. Previous studies have shown that lupins exhibit a remarkable variation of their chromosome number (2n = 32 to 52) and their genome size (2C = 1 to ~ 2.6 Gb), including between closely related taxa, regardless of their chromosome number (Gladstones *et al.*, 1998; Naganowska, 2003; Naganowska *et al.*, 2005; Conterato & Schifino-Wittmann, 2006; Mahé, 2009). A first PCR-based screening of the repeated compartment revealed a significant diversity of LTR-retrotransposons in lupin genomes (Mahé, 2009). Because of their beneficial properties for agriculture, human health and nutrition (Gladstones *et al.*, 1998; Cabello-Hurtado *et al.*, 2016), and for their novel status as model plants for studying symbiosis, proteoid roots and Pi uptake (O'Rourke *et al.*, 2013; Keller *et al.*, 2018), the smooth-seeded Mediterranean lupins, which include three crops (*L. albus*, *L. luteus* and *L. angustifolius*), are under increasing attention. Several transcriptomes and genomic resources have been generated (see Section 4.X and 6.Y) (Parra-González *et al.*, 2012; O'Rourke *et al.*, 2013; Kamphuis *et al.*, 2015;

Accepted Manuscript

Keller *et al.*, 2018) and a first draft genome has been recently released (Hane *et al.*, 2017; this book), providing the raw material to best understand structure, evolution and functional potential of the lupin genome, including its repetitive compartment.

Therefore, as a first step to develop our knowledge on this enigmatic genomic compartment, we report results from: (i) a preliminary survey of the diversity of LTR-retotransposons (*Ty1-copia* and *Ty3-gypsy-like* elements) in *Lupinus* and allied Genistoid taxa, based upon analysis of their reverse transcriptase sequences (RTs) (Flavell *et al.*, 1992; Alix & Heslop-Harrison, 2004; Mahé, 2009); and (ii) a detailed evaluation of the repetitive compartment of four smooth-seeded Mediterranean lupin taxa, from the analysis of low-depth NGS genomic resources, using different programs to identify and estimate the repetitive sequences (Benson, 1999; Novák *et al.*, 2010, 2017; Novak *et al.*, 2013).

13.2 Exploring retrotransposons diversity in genomes of lupins and allied Genistoids

Ty1-copia-like and *Ty3-gypsy-like* superfamilies of class I retrotransposons are ubiquitous in eukaryote genomes and most often involved in genome size (GS) variation. A preliminary investigation of their diversity, was conducted in 44 accessions belonging to 27 lupin taxa (16 from the Old World and 11 from the New World) and 8 other Genistoid representatives; Table 13.1). This was carried out

Accepted Manuscript

through analysis of their constitutive reverse transcriptase sequences (RTs) (Mahé, 2009). Accordingly, conserved coding RT domains were amplified, cloned and sequenced from genomic DNA samples using universal primers (Flavell *et al.*, 1992) following the procedure described by Alix & Heslop-Harrison, (2004). After (i) removing the low-quality sequences from the hundreds of amplicons generated, (ii) verifying their homology with known RTs from public databases (via Blastn, Blastx and RepeatMasker; http://www.ncbi.nlm.nih.gov and http://www.repeatmasker.org/), and (iii) size-filtering, a total of 367 retrotransposon-*like* RT sequences were selected for further analysis. Among them 260 amplicons ranged in size from 248 to 295 bp, with pairwise identity varying from 38.6 to 100% for *copia* elements (GenBank accession numbers GU189754 to GU190013); and 107 ranged from 366 to 564 bp with a pairwise identity of 32.4 to 100% for *gypsy* elements (accession numbers GU190014 to GU190133).

Within this set of amplicons, 305 were from the lupin species (211 RT-copia, 89 RT-gypsy and five unidentified) and 62 were from the eight Genistoid representatives (including 40 RT-copia, 17 RT-gypsy and five unannotated). Altogether, these 367 DNA sequences were aligned with MAFFT (Katoh & Standley, 2013). The sequence data matrix was then subjected to a maximum likelihood (ML) phylogenetic analysis, using the best-fitted evolutionary model (GTR+R7: General Time Reversible model, rates Gamma distributed) identified with ModelFinder (Kalyaanamoorthy *et al.*, 2017) as

Accepted Manuscript

implemented in IQ-TREE v1.5.5 (Nguyen *et al.*, 2015). The robustness of the nodes was estimated with 10,000 ultrafast bootstrap replicates (Hoang *et al.*, 2017). The phylogenetic tree resulting from this analysis is shown in Figure 13.1, where each terminal branch representing RT-*copia* or RT-*gypsy* amplicons is colored according to its taxonomic and geographic origin (see Fig. legends) and its annotation assignment indicated by different colors in the outer circle.

Table 13.1: List of accessions from *Lupinus* and other Genistoid taxa surveyed for retrotransposons diversity. The origin, geographic distribution, and accession reference are indicated for each sample. OW = Old World species; NW = New World species.

Taxon	**2*n***	**Origin/Distribution**	**Sample source & Reference number**
L. affinis	48	Oregon/NW, West NA	USDA/504315/N20
L. albus	50	Algeria/OW, Med	INAE-DZ/M20
L. anatolicus	42	Turkey/OW, Afr	AKA/K32
L. angustifolius ssp. reticulatus	40	France/OW, Med	AKA/T25
L. angustifolius ssp. angustifolius	40	Algeria/OW, Med	AKA-M1/T24
L. atlanticus	38	Morocco/OW, Afr	USDA/384612-FM83/T1
——	38	Morocco/OW, Afr	INRA-SAPF/T11
——	38	Morocco/OW, Afr	USDA/384613-FM87/T2
L. bracteolaris	32-34	Brazil/NW, South-East SA	USDA/404349/S80
L. concinnus	?	USA/NW	N19
L. cosentinii	32	?/OW, Med	INRAL-FR/T15
L. diffusus	?	Florida/NW	K35
L. digitatus	36	Egypt/OW, Afr-Med	WADA-PI26877/T4
L. elegans	48	Mexico/NW, West SA	USDA/185099/S33
L. hirsutissimus	?	USA/NW	AKA/N85
L. hispanicus ssp. bicolor	52	Spain/OW, Med	USDA/PI 384554/T23
L. hispanicus ssp. hispanicus	52	Portugal /OW, Med	USDA/384555/T22
L. luteus	52	Algeria/OW, Med	AKA/M5
——	52	Algeria/OW, Med	AKA/T20

Accepted Manuscript

——	52	Algeria/OW, Med	AKA/T21
L. mariae-josephi	52?	Spain/OW, Med	H. Pascual/MJ1
L. micranthus	52	Algeria/OW, Med	AKA/T19
——	52	Algeria/OW, Med	T 28
L. mutabilis	48	Perou/NW, West SA	INAE-DZ/S35/MU23
L. nanus	48	USA/NW	N42
L. palaestinus	42	Near-East/OW, Afr-Med	INRA-FR/T14
L. paraguariensis	36	Brazil/NW, East SA	BRA-02828/BZ1
L. pilosus	42	Algeria/OW, Afr-Med	INAE-DZ/T6
——	42	Algeria/OW, Afr-Med	INAE-DZ/T9
——	42	North-Africa/OW, Afr-Med	USDA/W6 PI 11995/T13
L. pilosus tassilicus	?	Lybia/OW, Afr	AKA/A641
L. polyphyllus	48	USA/NW, NA	USDA/504404/T26
L. princei	38	Kenya/OW, Afr	WADA P 23021/T0
——	38	Kenya/OW, Afr	RP Chyulu 1800/T16
——	38	Kenya/OW, Afr	RP Chyulu 1915/T17
L. texensis	36	USA/NW, South NA	USDA/577291/N45
Anarthrophyllum cumingii	?	?/NW, South SA	AKA/201
Argyrolobium uniflorum	?	OW	AKA/G25
Chamaecytisus mollis	?	OW	AKA/C84
Crotalaria podocarpa	?	OW	AKA/K50
Cytisus heterochrous	?	OW	AKA/G8
Genista tinctoria	?	OW	AKA/G56
Thermopsis rhombifolia	?	NW	AKA/G46
Ulex parviflorus	?	Spain/OW, Med	AKA/G24

Accepted Manuscript

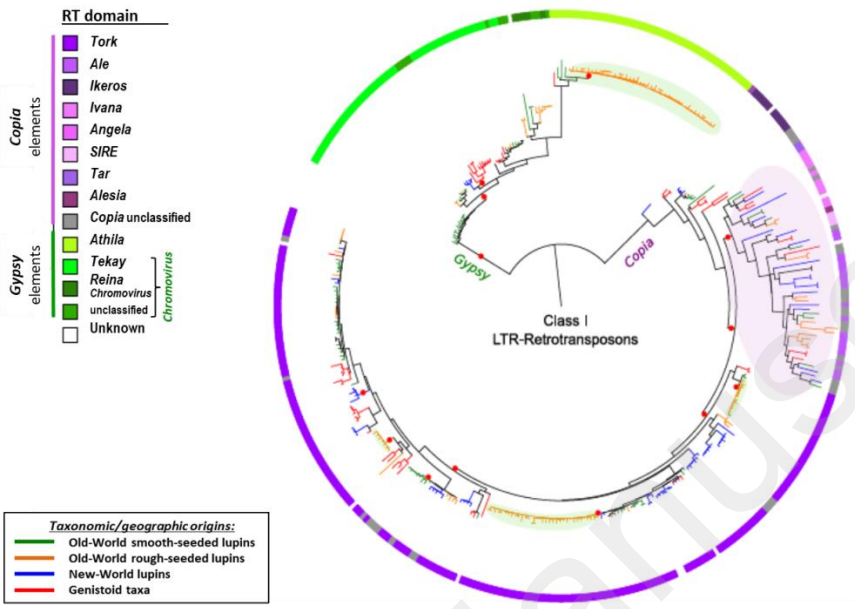

Figure 13.1: Maximum likelihood phylogenetic tree of 367 *copia* and gypsy RT fragments amplified from 44 accessions belonging to 27 Old World and New World lupin taxa (305 sequences) and from 8 other Genistoid representatives (62 sequences). Each terminal branch is colored according to its taxonomical/geographical group of origin and its *copia* or *gypsy* annotation assignment indicated by different colors in the outer circle (see legends in the figure). Red dots on the tree indicate remarkable well supported nodes (by bootstrap estimate). Some remarkable groups of *copia* or *gypsy* amplicons likely representing ubiquitous elements of ancient origin (shaded in mauve) or recent lineage-specific amplifications (shaded in light green) are also indicated.

A remarkable diversity of retrotransposon elements was detected within genomes of both lupin and Genistoid species. Random amplified RT sequences using universal primers allowed a clear segrega-

Accepted Manuscript

tion of the two retrotransposon superfamilies, with a higher amplification success for the *copia* ones. Within each of these superfamilies, several families have been identified: mainly *Tork copia*-like elements and at a lower scale other *copia* families (*Ale, Ikeros, Ivana, Angela, SIRE, Tar* and *Alesia*). Whereas *gypsy*-like elements were mainly represented by *Tekay* and *Athila* elements, and few Reina and *Tat/Ogre* ones. Most of them showed significant levels of RT-identity with those from other Fabaceae available in databases for *Cicer* (52.5-88.4 %), *Vigna* (51-86 %), *Vicia*, or Soybean, suggesting that these elements were most likely inherited from a common Papilionoid ancestor or even from earlier origin.

Although the number of amplicons is low for some taxa, the main retrotransposons families detected appear to be ubiquitous throughout the lupine and Genistoid genomes, as illustrated by the presence of multicolored branches in each of the main *copia* and *gypsy* clades in Figure 13.1). One remarkable and well-supported multicolored clade (shaded in mauve) of related *copia* elements (*Alesia, SIRE, Angela, Ivana*) includes highly divergent RT-sequences (with long evolutionary branches) from most lupin and Genistoid taxonomic and geographical groups. This suggests that they likely derive from a common and ancient ancestor. The other noteworthy groups revealed by the phylogeny are those including weakly divergent sequences isolated from the same taxonomic or geographical lineage (shaded in light green), indicating recent lineage-specific transpositional activities. This is particularly well exemplified by several spe-

Accepted Manuscript

cific retrotransposon families (groups with monochrome branches) observed in the tree for: the rough seeded Old World lupins with orange branches (*Athila* and *Tork*); the smooth seeded Old World lupins with green branches (*Tekay* and *Tork*); the New World lupins with blue branches (*Tork* groups); and in the Genistoids with red branches (a *Tekay* group). Other homogeneous lineage-specific groups are composed of more divergent RT sequences likely deriving from earlier transposition events, such as for instance in the Genistoids (red branches) which show specific lines of *Athila* and *Tork* elements.

Within the collection of conserved RT domains generated from the lupins, five were amplified from a sample of *L. angustifolius* subsp. *angustifolius* (originating from North Africa), three RT-*gypsy* and two RT-*copia* clones. These clones have been used as queries in a rapid screening of the current reference NLL genome (of *L. angustifolius* cultivar. Tanjil; (Hane *et al.*, 2017)) to estimate a potential number of PCR-based amplified products that could be expected from this genome. Interestingly, no less than 997 and 1209 non redundant hits were found for the *gypsy* and the *copia* elements, respectively, using the easy Blast search tool (with evalue threshold: 1.0e-5) implemented in the Lupin Genome Portal https://www.lupinexpress.org/ (Priyam et al., 2015).

Therefore, despite an inevitably biased sampling due to the intrinsic limits of the method (related to variable rates of RT degeneracy within and among genomes, to the performance of the "universal

Accepted Manuscript

primers", and depending on the cloning and sequencing depth), the PCR-based exploration of the lupin and Genistoid genomes allowed detection of a wide diversity of *copia*-like and *gypsy*-like LTR-retrotransposons families. Most of them are ubiquitous throughout the lupins and Genistoids. Moreover, phylogenetic analysis of the RT-sequences provided clues which suggest that some retrotransposons subfamilies seem to have differentially and specifically proliferated (bursts) during the recent evolutionary history of the genus in the New and the Old World lupins. Besides, a fluorescence in situ hybridization (FISH) test performed on metaphase root cells of Old World lupines using *copia* ant *gypsy* RT-probes (Mahé, 2009) indicated a much higher accumulation of retrotransposons in the large genome of the Mediterranean species *L. luteus* (2367 Mb/2C) than in the small genome of *L. micranthus* (1147 Mb/2C). Thus, altogether these results emphasized the need to more accurately identify and evaluate the diversity and relative abundance of transposable elements.

13.3 NGS-based evaluation of the repetitive compartment in lupin genomes

As highlighted, lupins are characterized by a noteworthy genome sizes variation (GS = 2C = nuclear DNA amount per cell) ranging from 1.05 to 2.6 Gb, regardless of their various chromosome numbers (varying from 2n = 32, 34, 36, 38, 40, 42, 48, 50 to 52). This is observable even between taxa having the same chromosome number (such as *L. luteus* which has more than twice the size of that of *L.*

Accepted Manuscript

micranthus), as well as regardless of their Old World or New World origins and of their phylogenetic relationships.

Therefore, in order to deepen our understanding of the lupin genome dynamics, four lupin accessions with small and large genomes, belonging to different Mediterranean Old World smooth-seeded species (Table 13.2; Figure 13.2), were subjected to a comparative NGS-based analysis of their genomic repetitive compartment*: L. albus* (2n=50; 2C=1.13 Gb), *L. angustifolius* (2n=40; 2C=1.85 Gb), *L. luteus* (2n=52; 2C=2.37 Gb), and *L. micranthus* (2n=52; 2C=1.15 Gb). For this purpose, a sequence dataset of 1,200,000 Paired-End 100 bp reads (120 Mb) per accession, extracted from resources generated by low-depth genomic Illumina HiSeq sequencing, and representing 5 to 10% of each genome (Table 13.2), was analyzed with RepeatExplorer (Novák *et al.*, 2010; Novak *et al.*, 2013). Following analysis of a combined data set (including 4,800,000 reads, each labeled according to its species origin), 293,635 clusters were obtained. Among the 744 clusters containing more than 48 reads (the largest having 265,540 reads), 176 were annotated as transposable elements and 207 as simple sequence repeats; the remaining clusters corresponded to organelle or to unclassified sequences.

Accepted Manuscript

Table 13.2: Origins and characteristics of the genomic resources of four *Lupinus* species used in this study.

Species	Acces-sion code	Origins	2*n*	2C DNA amount* (in pg)	Genome Size** (2C in Mb)	Total length of reads sequenced (Gb)	Genome coverage (x folds)	% of genome analyzed with RE***
L. albus	M20	Egypt	50	1,16	1134,48	1.2	1.06	10.57 %
L. angusti-folius	IPG2	Morocco	40	1.89	1848.42	15.4	8.2	6.5 %
L. luteus	M6	Algeria	52	2,42	2366,76	1.15	0.49	5.07 %
L. micran-thus	B12	Algeria	52	1,07	1147	1.13	1.06	10.46 %

*According to (Naganowska, 2003) and (Mahé, 2009)
** Using 1 pg = 978 Mb according to (Dolezel *et al.*, 2003)
***Repeat Explorer (Novák *et al.*, 2010; Novak *et al.*, 2013)

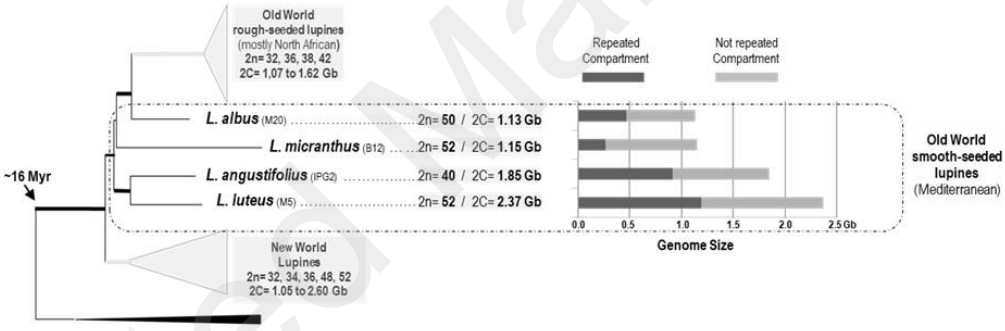

Figure 13.2: Condensed phylogenetic tree of the lupins (on the left) redrawn from (Mahé *et al.*, 2011), showing the position and relationships of the four Mediterranean smooth-seeded species subjected to a RepeatExplorer analysis of their repetitive genomic compartment. The Old World rough-seeded lupins and the New World lupin clades are presented. The mean age of the lupin stem node is indicated (according to Hughes and Eastwood, 2006). Chromosome numbers (2*n*) and genome size (2C in Gb) of the taxa are given on the right of the figure, together with a histogram showing the genomic proportion of the repeated compartment in the four Mediterranean smooth-seeded species analyzed.

Accepted Manuscript

13.3.1 Composition of the repetitive compartment in lupin genomes.

Identification and distribution of the main elements of the repetitive compartment have been determined in the four targeted genomes. As summarized in Table 13.3 and illustrated in Figure 13.2, the repetitive compartment (including transposable elements and tandem repeat satellites; excluding nuclear ribosomal DNA sequences or nrDNA) represents a large part of the genomes and varies from 23.27% of the small genome of *L. micranthus* to 50.36% in the largest genome of *L. luteus*, regardless of their same chromosome number (2n=52). While *L. albus* shares a close chromosomes number (2n=50) and a similar genome size with *L. micranthus*, it contains a much larger proportion of repeats (41.10%). In turn, the accession of *Lupinus angustifolius* analyzed here, which has a lower chromosome number (2n=40), and a relatively large genome (2C=1.85Gb), exhibits a high repeats proportion (49,63%), which is underestimated compared to the 54% reported for the NLL cultivar Tanjil sequenced genome (Hane *et al.*, 2017).

In *L. albus*, *L. luteus*, and *L. micranthus*, the repetitive compartment is mainly composed of transposable elements, which are essentially represented by variable proportions of LTR retrotransposons (33,27%, 41,10% and 13,43% of the genome, respectively), whereas LINEs and DNA transposons are present at less than 2% in each ge-

Accepted Manuscript

nome (Table 13.3; Figure 13.3). Apart from the indeterminate repeats (around 2 to 3%), satellites (tandem repeats) are present at a low proportion in the three genomes, ranging from 3.37% in *L. albus*, to ~5-6% in *L. luteus* and *L. micranthus*. Whilst similar repeat categories were detected in *L. angustifolius*, it exhibited a noteworthy different pattern, with a repetitive compartment made up of a little more than half by satellites (around 26% of the genome). The remaining part is mainly composed of LTR retrotransposons (approximately 21% of WG). Although, the RepeatExplorer-based proportion of LTR retrotransposons was lower than that estimated from the NLL (narrow-leafed lupin) sequenced genome (~28%), the above observations already demonstrate that not only TEs but also satellites may account for high proportions in lupine genomes where they may reach substantial amounts, ranging from ~38Mb and ~65Mb in the small genomes (of *L. albus* and *L. micranthus*, respectively) to ~126Mb and ~481Mb in the larger genomes of *L. luteus* and *L. angustifolius*, respectively. Otherwise, the nrDNA varies from 1.5 to 2% in the genomes of *L. albus*, *L. luteus*, and *L. micranthus*, while it displays a significant increase in *L. angustifolius* (~3%), which suggests the occurrence of different nrDNA evolutionary patterns among the Mediterranean lupins (Wolko and Weeden, 1989; Kroc *et al.*, 2014).

Accepted Manuscript

Table 13.3: Proportions of the main DNA repeats categories (as % of the genome) in four Old World lupins. Repeats are classified is according to RepeatExplorer annotation

Genomic proportion of the different DNA repeat categoties				
Repeats annotation	***L. albus***	***L. angustifolius***	***L. luteus***	***L. micranthus***
LTR retrotransposons (*copia+gypsy*)	**33,27**	**20,44**	**41,10**	**13,43**
LINE	0,07	0,04	0,03	0,08
DNA transposons	1,52	1,05	1,24	0,88
Satellites *sensu lato*	3,37	**25,97**	5,29	5,74
Unknown	2,94	2,13	2,69	3,15
nrDNA/45S	1,95	2,91	1,53	1,67
Repetitive Compartment (nrDNA excluded)	**41,18**	**49,63**	**50,36**	**23,27**

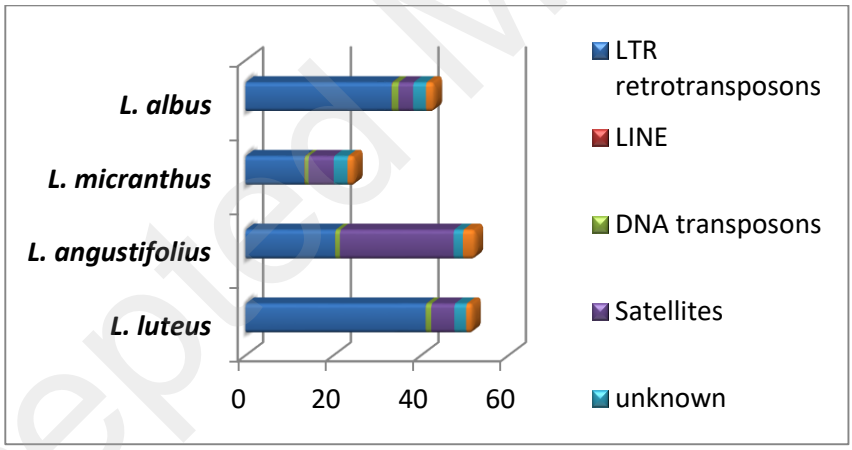

Figure 13.3: Histogram showing the genomic proportions of the main DNA repeats categories (as % of the genome) in four Old World lupins. Repeats are classified according to RepeatExplorer annotation outputs (Novák *et al.*, 2010).

Accepted Manuscript

13.6.2 Identification and distribution of LTR retrotransposons in the lupin genomes. A more accurate analysis of the repetitive compartment shows that the *copia* and *gypsy* superfamilies of LTR retrotransposons are well represented in all species, at various proportions and different relative ratio (Table 13.4; Figure 13.4). *Copia* elements constitute 5.48% (62.9 Mb) to 11.73% (216.8 Mb) of the genomes, with highest proportions in large genomes (216.8 Mb for *L. angustifolius* and 231.5 Mb for *L. luteus*). In turn, the *gypsy* elements exhibited a wider range, from 3.73% (42.8 Mb) in the small genome of *L. micranthus* to 31.31% (741 Mb) in the largest genome of *L. luteus*, with however no correlation with GS regarding the substantial proportion of 20.31% (230.4 Mb) in *L. albus* (with a small GS) as compared to that of *L. angustifolius* (8.7%; 160.8 Mb) which has a larger GS. Accordingly, this observation reveals two distribution patterns of the LTR retrotransposon superfamilies. The first one is characterized by a *gypsy/copia* ratio lower than 1, where *copia* elements are ~1.3-1.5 times more abundant than the *gypsy* ones, such as in *L. micranthus* and *L. angustifolius*. The second pattern is defined by a *gypsy/copia* ratio much higher than 1, where *gypsy* elements clearly represent the prominent part of the LTR elements and are 2.6 and 3.2 times more abundant than the *copia* ones in *L. albus* and *L. luteus*, respectively.

Accepted Manuscript

Table 13.4: Proportions of the LTR-retrotransposon *copia* and *gypsy* families (as % of the genome) detected in four Old World lupins (annotated according to the nomenclature of (Wicker *et al.*, 2007).

TE Superfamily & Family	***L. albus***	***L. luteus***	***L. micranthus***	***L. angustifolius***
copia - AleI/Retrofit	0,09	0,03	0,07	0,03
copia - AleII	0,05	0,38	0,17	0,09
copia - Angela	1,13	1,41	0,92	1,49
copia - Ivana/Oryco	0,21	0,42	0,09	0,22
copia - TAR	0,44	0,50	0,37	0,54
Copia - Tork	0,77	0,96	0,70	1,30
copia - Maximus/SIRE	5,06	6,06	3,15	8,06
Subtotal *copia* (% of the genome)	**7,75**	**9,78**	**5,48**	**11,73**
gypsy - Athila	1,09	16,12	0,97	3,30
gypsy - Ogre/Tat	0,65	4,58	1,47	1,27
gypsy - Chromovirus	18,57	10,62	1,29	4,14
Subtotal *gypsy* (% of the genome)	**20,31**	**31,31**	**3,73**	**8,70**
***gypsy/copia* ratio**	2,62	3,2	0,68	0,74

Accepted Manuscript

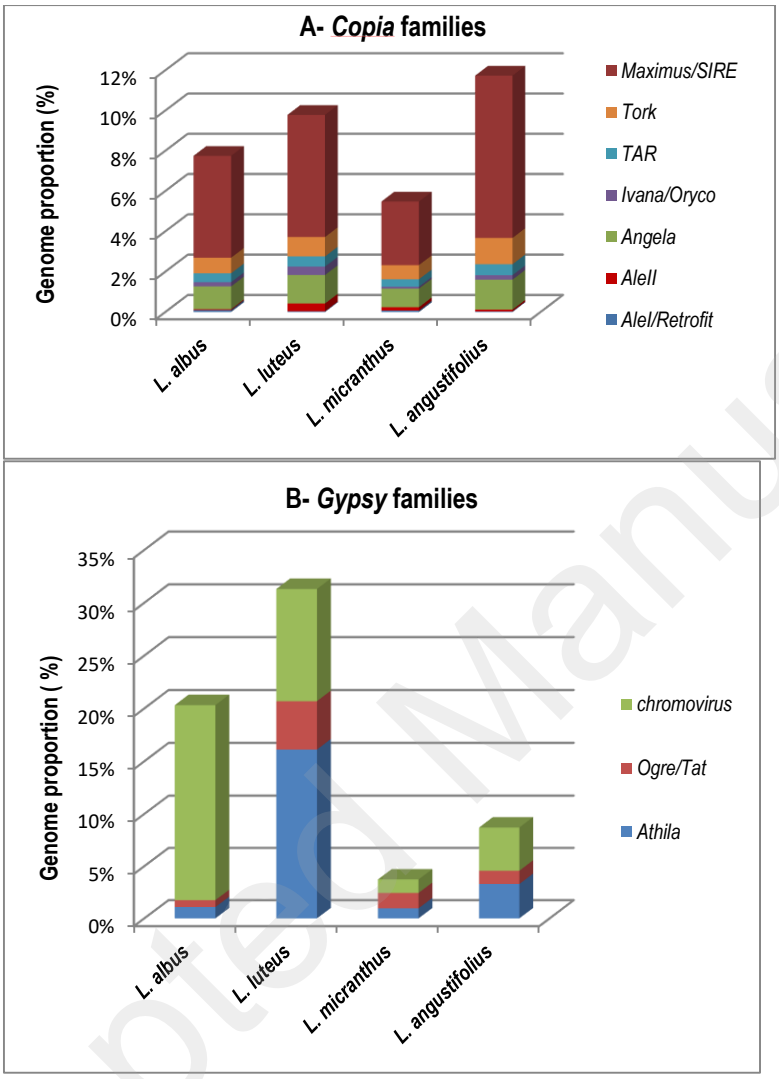

Figure 13.4: Histogram showing the genomic proportions of the LTR-retrotransposon *copia* (A) and *gypsy* (B) families (as % of the genome) detected in four Old World lupines. Retrotransposons are annotated according to the nomenclature of (Wicker *et al.*, 2007).

Accepted Manuscript

A thorough annotation revealed that each of the LTR superfamilies is characterized by a fairly homogeneous and similar profile of *copia* and *gypsy* TE families in the lupin genomes surveyed, regardless of their variable proportions. Indeed, seven different *copia* (*Alel/Retrofit, AlelI, Angela, Ivana/Oryco, TAR, Tork*, and *Maximus/SIRE*) and three *gypsy* (*Athila, Ogre/Tat, Chromovirus*) families were identified in all species (Table 13.4; Figure 13.4). The *Maximus/SIRE* family is the best represented in the *copia* superfamily with 3 to 8.06% of the genomes (in *L. micranthus* and *L. angustifolius*, respectively), followed at a lower level by the *Angela* (0.92 to 1.49%) and *Tork* (0.7 to 1.3%) families. Together, the latter three families represent 86 to 92% of the *copia* elements of each genome, while the remaining families (*TAR, Tork, AleI, Allel/Retrofit, Ivana/Oryco*) are poorly represented, each at less than 0.6% of the nuclear genome. In the gypsy superfamily, the *Athila* family alone represents 16.2% (~382Mb, *i.e*. half of the repetitive compartment) of the large *L. luteus* genome, whereas the *Chromovirus* family makes up 10.62 (~252Mb) and 18.57% (210Mb) of the genomes of *L. luteus* and *L. albus* (a small genome). The *Ogre/Tat* family is much less represented throughout the lupine genomes (less than 5%), with however a substantial amount (4.58%, *i.e*. ~109Mb) in the large *L. luteus* genome. It is interesting to notice here: (i) that the amplification of *Athila* and *Chromovirus* elements played a decisive role in genome size increase in *L. luteus* (together representing 26.74% of the genome) compared to its counterpart *L. micranthus* (2.26%), which has the same chromosome number and a smaller ge-

Accepted Manuscript

nome; (ii) that the latter elements were either only moderately amplified (or amplified then partly deleted *via* removal mechanisms; (Devos, 2002)), such as in the other large genome of *L. angustifolius* (7.44%); but also, (iii) that *gypsy* elements may significantly proliferate in the small genomes, such as Chromovirus (18.57%) in *L. albus*.

13.3.3 Phylogenetic analysis on LTR retrotransposons RT domains.

In order to refine the annotation of LTR retrotransposons and to get insights into their diversity and dynamics in the Mediterranean lupine genomes, phylogenetic analyses were performed on RT (reverse transcriptase) domains extracted from clusters of reads generated by the RepeatExplorer analyses.

For each species, reads of each cluster (annotated as *copia* or *gypsy*) were assembled independently with Mira4 (Chevreux *et al.*, 1999), and the consensus sequences obtained were submitted to BLASTx v. 2.6.0+ (Altschul *et al.*, 1990; Camacho *et al.*, 2009) against a public database of RT nucleotide sequences (Repbase v. 23.08; (Bao *et al.*, 2015)). Sequences translated in protein which showed homology with RT domains, and that have at least 130 amino acids in length, were kept for further analyses. RT sequences from six angiosperms species (*Glycine max*, *Medicago truncatula*, *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, *Triticum monococcum)* were selected in Repbase and added to the dataset. Only potentially functional se-

Accepted Manuscript

quences without stop codon were retained. Each of the *gypsy* or *copia* RT protein sequences were aligned separately using Clustal Omega (Sievers *et al.*, 2014). Informative blocks in multiple alignment were selected with the GBlocks package (Castresana, 2000) prior to perform phylogenetic analyses with IQ-TREE (Nguyen *et al.*, 2015). The LG+R6 and the LG+R7 protein evolution models were respectively retained (*via* ModelFinder; (Kalyaanamoorthy *et al.*, 2017) for phylogenetic reconstruction of *gypsy* and *copia* trees using the maximum likelihood method. The robustness of branches was estimated after 10,000 Ultrafast Bootstraps (Hoang *et al.*, 2017). Annotation of *copia* and *gypsy* elements was determined according to the classification of Wicker *et al.*, 2007.

The *copia* tree was built with 71 lupin sequences (43 from *L. angustifolius*, 2 from *L. micranthus*, 13 from *L. albus* and 13 from *L. luteus*) and 244 sequences from other taxa (Figure 13.5). Interestingly, all the most conserved RT sequences detected in *L. albus*, *L. luteus* and *L. micranthus*, and about half of those detected *L. angustifolius*, belong to the *Maximus/SIRE* family, which agrees with the prominence of this *copia* family in the lupin genomes. Moreover, this suggests that these elements, displaying well conserved RT domains, most likely result from recent amplification events experienced by each species, as this seems corroborated by some specific groups of poorly divergent sequences with short branches (indicated in Figure 13.5). All the other remaining conserved RT sequences represented diverse *Angela*, *TAR*, *Tork* and *Ale* elements detected in

Accepted Manuscript

L. angustifolius, which indicates that it is the only Mediterranean lupine species containing conserved copies of these *copia* families that are potentially able to proliferate. In particular, a distinct monophyletic group of *Angela* RTs suggests a lineage-specific amplification of one *Angela* line during the recent evolutionary history of this species. Although, *AleII/Retrofit* and *Ivana/Oryco* elements were detected in all lupins, indicating their common and ancient origin, no conserved RTs were found, which suggests that these poorly represented elements have undergone degenerative processes that tend towards their elimination from the genomes.

Accepted Manuscript

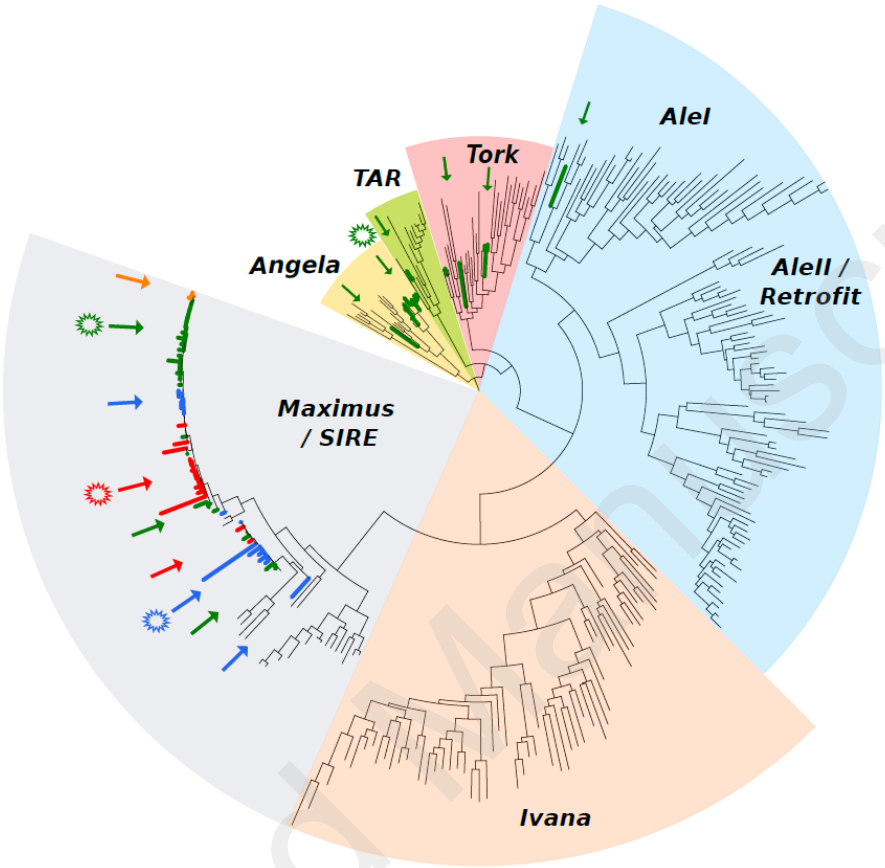

Figure 13.5: Phylogenetic analysis of lupin LTR-retrotransposon *copia* elements based on amino-acid sequences of their conserved RT domains, using the maximum likelihood method. The tree was built with 71 lupin sequences (43 from *L. angustifolius*, 2 from *L. micranthus*, 13 from *L. albus* and 13 from *L. luteus*) and 244 reference sequences from databases. Annotation of *copia* families (colored clades named in black and bold) were determined following the classification of (Wicker *et al.*, 2007). Each terminal branch is colored according to its species origin: green for *L. angustifolius*, orange for *L. micranthus*, blue for *L. albus*, red for *L. luteus*, and black for reference taxa. Radiated/irregular circles likely represent recent species-specific amplification of particular *copia* lines.

Accepted Manuscript

The *gypsy* tree was constructed with 72 lupine sequences (7 from *L. angustifolius*, 1 from *L. micranthus*, 29 from *L. albus* and 35 from *L. luteus*) and 236 reference sequences from a set of plant genomes (Figure 13.6). All the most conserved RT sequences detected belong to the three *gypsy* families identified *via* RepeatExplorer, *Chromovirus*, *Athila* and *Ogre/Tat* (Table 13.4). The distribution of the conserved RTs among species appears correlated with the relative proportions of the *gypsy* families in the genomes. Conserved RTs of *Athila* elements were mostly found in *L. luteus* (13) and few in *L. angustifolius* (3) and *L. albus* (1). Few conserved RTs (1 to 3) of the *Ogre/Tat* elements were detected in lupins (with none in *L. albus*). With regard to C*hromovirus* elements, conserved RTs were mostly extracted from *L. luteus* and *L. albus*, the richest genomes in *gypsy* elements, and only three from *L. angustifolius*. Among the wide range of known *Chromovirus* elements, the phylogeny allowed to refine the classification of the lupin ones into two subfamilies, most of them as *Tekay* homologs and the few others as *CRM* homologs (following RepBase annotation). Interestingly, the *gypsy* phylogeny reveals that *L. luteus* and *L. albus* most likely experienced recently independent and specific proliferation of *gypsy* elements, as this is illustrated by noteworthy monophyletic and monochromatic groups of poorly divergent RTs (with short branches) of *Tekay* and *Athila* retrotransposon lineages in Figure 13.6. The other conserved RTs are minority lineages of *gypsy* elements represented in the Mediterranean lupin genomes that seem, however, yet potentially functional and able to proliferate.

Accepted Manuscript

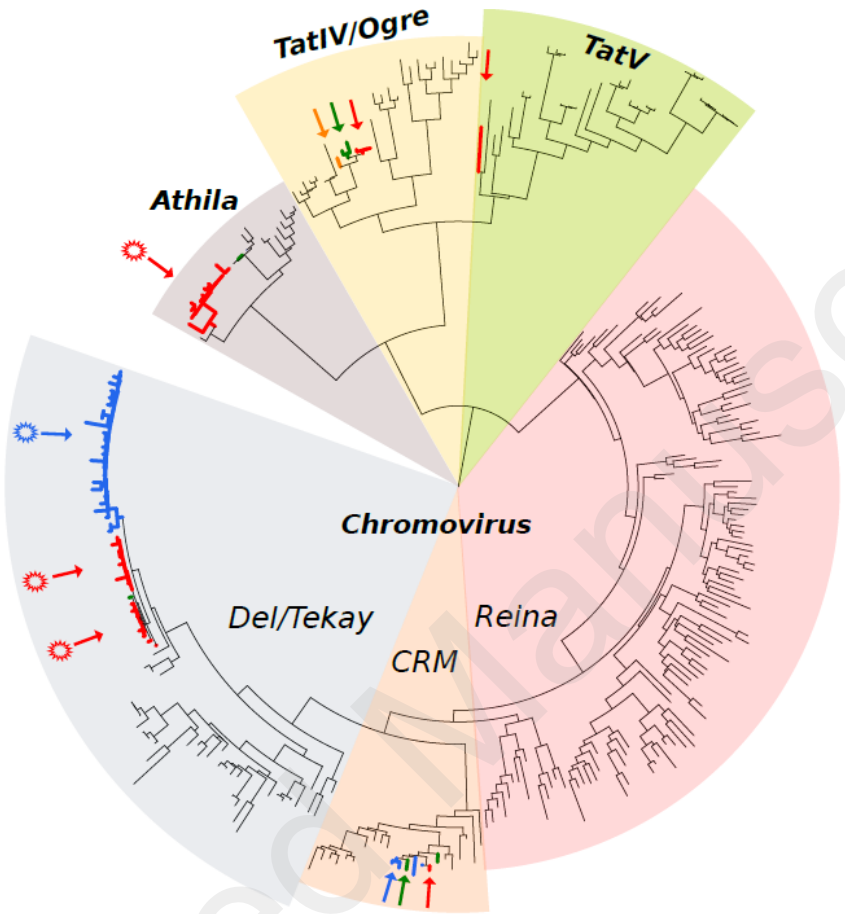

Figure 13.6: Phylogenetic analysis of lupin LTR-retrotransposon *gypsy* elements based on amino-acid sequences of their conserved RT domains, using the maximum likelihood method. The tree was built with 72 lupin sequences (7 from *L. angustifolius*, 1 from *L. micranthus*, 29 from *L. albus* and 35 from *L. luteus*) and 236 reference sequences from databases. Annotation of *gypsy* families (colored clades named in black and bold) were determined following the classification of (Wicker *et al.*, 2007). Each terminal branch is colored according to its species origin: green for *L. angustifolius*, orange for *L. micranthus*, blue for *L. albus*, red for *L. luteus*, and black for reference taxa. Radiated/irregular circles likely represent recent species-specific amplification of particular *gypsy* lines.

Accepted Manuscript

13.3.4 Diversity and abundance of Tandem Repeats in lupin genomes

As shown from the above RepeatExplorer-based analysis, the proportion of tandem repeats (excluding nrDNA) in the Mediterranean lupin genomes, varies from 3.37% in *L. albus* to a tremendous value of 26% in *L. angustifolius* (Table 13.3). In the latter species, TRs were even revealed more abundant than TEs. For each species, the reads contained in the clusters annotated as TRs were together analyzed using the TRF program v.4.09 (Tandem Repeat Finder; (Benson, 1999)) in order to identify the TR motives (k-mers < 50bp) and their statistical distribution. Among the best represented SSRs (with k-mer motives < 10bp), three k-mers (AGGAT, GATGAG and GTTTAGG) were almost always present at a low level (less than 0.6%) in the four genomes, with however an exceptional accumulation of the 6-mer GATGAG estimated at 15.24% of the genome in *L. angustifolius*. (Table 13.5; Figure 13.7). Tandem repeats with k-mers>10bp may constitute substantial amounts in lupin genomes and represent the main TR fraction in *L. albus* (1.64%) and *L. luteus* (2.78%). Interestingly, complementary analyses of the latter TR fraction (using TAREAN program; (Novák *et al.*, 2017)) allowed identification of one major 28-mer minisatellite in *L. luteus*, one major 170-mer satellite and one 38-mer minisatellite in *L. albus*, as well as two 165-mer and 629-mer satellites in *L. micranthus*.

Accepted Manuscript

Table 13.5: Proportion of the main types of tandem repeats (as % of the genome) detected in four Old World lupins.

TR motifs	*L. albus*	*L. angustifolius*	*L. luteus*	*L. micranthus*
AGGAT (5 bp)	0,16	0,33	0,37	0,60
GATGAG (6 bp)	0,00	15,24	0,25	0,57
GTTTAGG (7 bp)	0,13	0,16	0,06	0,20
Others <=10 bp	0,20	0,93	0,06	0,19
Others >10 bp	1,64	2,26	2,78	0,53

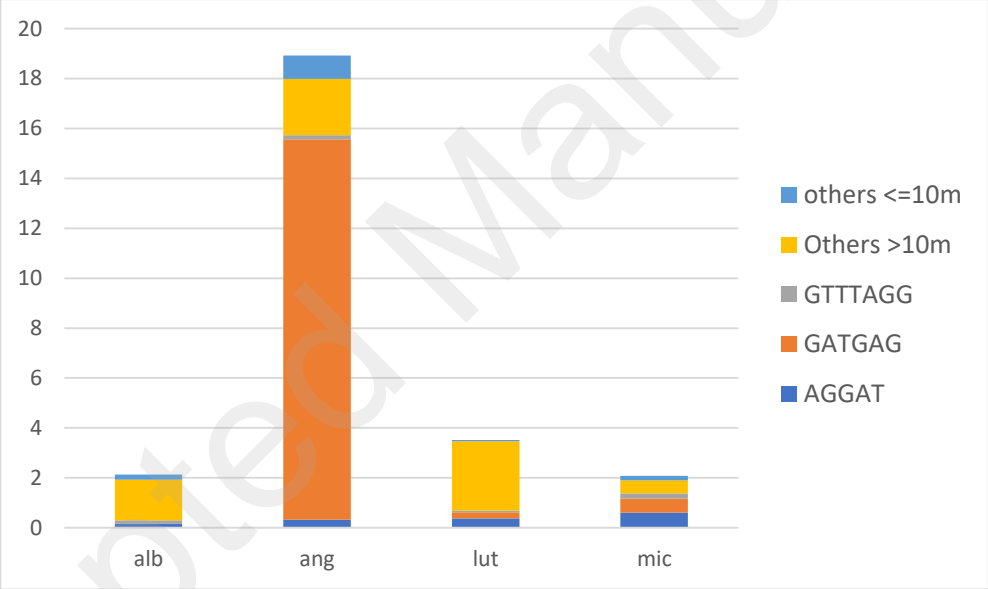

Figure 13.7: Histogram representing the diversity and proportion of the simple sequence repeats (as % of the genome) detected in four Old World lupins using Tandem Repeat Finder program (Benson, 1999; Lim *et al.*, 2013): from *L. albus*(alb), *L. angustifolius* (ang), *L. luteus* (lut), and *L. micranthus* (mic).

Accepted Manuscript

Moreover, taking advantage of the availability of a reference genome (*L. angustifolius* NLL cultivar. Tanjil; (Hane *et al.*, 2017)), the twenty annotated pseudochromosomes were screened with TRF in order to identify, localize and estimate the distribution of microsatellites (as per cent of 100-kb). Almost all tandem repeats found in coding sequences are 2- or 3-mers, of which the 3-mer "CTT" is the most commonly distributed. However, they only represent a total of 24,000 bp (*i.e*, 0.03% of the assembled genome). Interestingly, the presence of the other abundant SSRs (5-, 6- and 7-mers) detected above in our *L. angustifolius* accession (IPG2 from Morocco) were confirmed in the Tanjil genome, but were rather localized outside of the coding sequences. The density and localization of the SSRs relative to the distribution of the genes are summarized in Figure 13.8 (using a Circos representation; (Krzywinski *et al.*, 2009)). The SSRs are distributed in all the genome and didn't exhibit any chromosome specificity. The 6-mer SSR$(GATGAG)^n$ previously identified in the IPG2 accession is confirmed as the major SSR in the NLL genome cv. Tanjil, with pics of density mainly distributed in gene-poor regions. A thorough survey reveals that 1,143 genes include SSRs with the 3-mer (CTT) repeated at least four times. For example, a microsatellite with 65 perfect tandemly repeated (CTT) monomers was found in a putatively functional gene encoding a cytosolic oligopeptidase (ID:109349122).

Accepted Manuscript

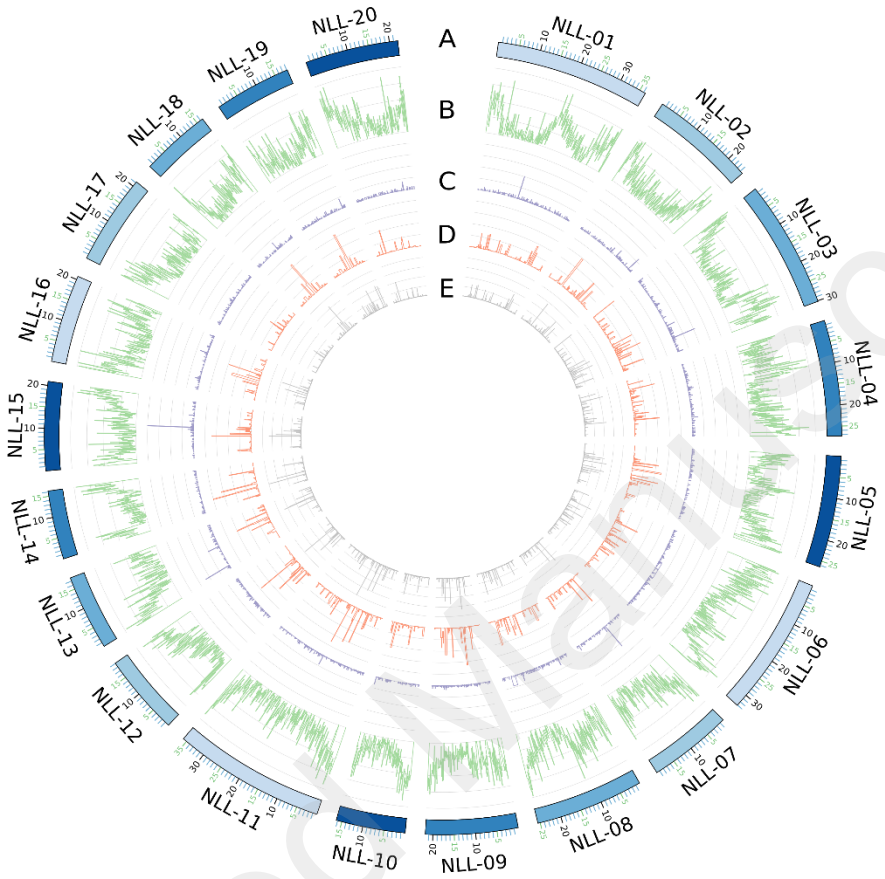

Figure 13.8: Microsatellites distribution along pseudochromosomes of the first WG lupin sequenced, *L. angustifolius* (NLL var. Tanjil). The five consecutive circles from the outside to the inside of the figure represent: A, the 20 chromosomes (named NLL-01 to NLL-20) (in blue); B, Genes distribution and proportions per 100kb (in green); C, proportions per 100 Kb of the 3-mer SSRs $(CTT)^n$ (x10 to be readable; in mauve); proportions per 100 Kb of 6-mers SSRs (in orange); and proportions per 100Kb of all microsatellites (in grey). The higher the peaks are the higher is the proportion of genes or SSRs

Accepted Manuscript

13.4 Repetitive compartment of lupin genomes

In this chapter, we present the first detailed evaluation of the repetitive compartment in genomes of four smooth-seeded Mediterranean lupin taxa, based on the analysis of low-depth NGS genomic resources, using various bioinformatics programs to identify and estimate the repetitive sequences (Benson, 1999; Novák *et al.*, 2010, 2017; Novak *et al.*, 2013). This approach already proved its usefulness to detect and to evaluate repeats in several taxa, which represent at least 0.01% of the genome, based on a genome coverage of >0.5% (*Pisum sativum*, (Macas *et al.*, 2007); *Musa acuminata*, (Hřibová *et al.*, 2010); *Nicotiana tabacum*, (Renny-Byfield *et al.*, 2011) ; Orobanchacea, (Piednoël *et al.*, 2013); *Genlisea*, (Vu *et al.*, 2015). Our estimate of the repetitive DNA in one accession of *L. angustifolius* (IPG2), based on a reduced sample of randomly selected reads (1C genome coverage = 3.25%) following the RepeatExplorer strategy, resulted in a proportion of 52.54% (including nrDNA) which is fairly close to the proportion of 57% found in the whole genome sequenced of the NLL cultivar Tanjil (Hane *et al.*, 2017). Also this was underlined by the studies cited above, which supports the robustness and reliability of this approach to investigate and compare non model species. Accordingly, this study yielded major information and insights on the composition, characterization, distribution and dynamics of the repetitive sequences in lupin genomes.

Accepted Manuscript

13.4.1 The repetitive compartment represents a significant fraction of lupin genomes

As frequently observed in other angiosperms (Bennetzen, 2000, 2005; Piegu *et al.*, 2006; Hu *et al.*, 2011; Bennetzen & Wang, 2014; Vu *et al.*, 2015; Wendel *et al.*, 2016), the repetitive compartment represents a large proportion of the genomes (23 to 51%, excluding nrDNA) in the Mediterranean smooth-seeded lupins. The highest proportions were found in the largest genomes, regardless of their chromosome number, 50.36% in *L. luteus* (2*n* = 52; 2C= 2.37 Gb) and 49.63% in *L.angustifolius* (2*n* = 40; 2C= 1.85 Gb), whereas the two lupins with small genomes and fairly similar chromosome numbers exhibited very contrasted proportions of repeats in their genomes, 23.27% in *L. micranthus* (2*n* = 52; 2C= 1.15 Gb) *versus* 41.10% in *L. albus* (2*n* = 50; 2C= 1.13 Gb). Therefore, the proportion of the repetitive compartment in the smooth-seeded Mediterranean lupins is overall neither correlated to chromosome numbers nor to GS, although large genomes are associated with a strong accumulation of repeated sequences (but not only, regarding the example of *L. albus*).

13.4.2 *Gypsy* and *copia* retrotransposons significantly contribute to genome size variation.

The repetitive compartment is mainly composed of transposable elements (~43 to ~85%) in the lupins surveyed and they significantly

Accepted Manuscript

contribute to the variation of their genome size. Moreover, the overwhelming majority of TEs is composed of Class I *gypsy* and *copia* LTR-retrotransposons (ranging from 93.9% of TEs in *L. micranthus* to 97% in *L. luteus*), which *in fine* are the main repeats fraction involved in GS differences (but see later). Together the other Class I (such as LINEs) and Class II elements (DNA transposons) only represented a minor fraction of the lupin genomes (less than 1.6%). This is in general accordance with estimates from other angiosperms, albeit some taxa exhibited a much higher proportion of Class II elements (11 to 16.5%), such as in *A. thaliana*, *G. max*, wheat and rice (Hawkins *et al.*, 2006; Oliver *et al.*, 2013).

The analyses based on random amplified RT domains and on Illumina HiSeq sequence data sets revealed in both lupin and Genistoid genomes a wide diversity of shared *copia* and *gypsy* LTR-retrotransposons families. The thorough evaluation of LTR-retrotransposon elements (*via* the RepeatExplorer strategy) highlighted the occurrence of a typical general profile of *copia* and *gypsy* families and subfamilies in the smooth-seeded Mediterranean lupin genomes, each species displaying its specific profile characterized by its own relative proportions of these elements. Additionally, a remarkable difference in the *gypsy/copia* ratio was observed among these species, regardless of their genome size as well as of their phylogenetic relationships, which is well exemplified by the prevalence of *copia* elements (~1.4 times more than *gypsy*) in *L. micranthus* and *L. angustifolius* and conversely by the over-accumulation of *gypsy*

Accepted Manuscript

elements (2.6 to 3.2 times more than *copia*) in *L. albus* and *L. luteus*. It is noteworthy that few individual *gypsy* (Chromovirus and Athila) and *copia* (Maximus-SIRE) families alone have been remarkably accumulated in the lupins and hence strongly contributed in shaping their LTR-retrotransposon profiles and in their GS differences, as shown in *L. luteus* (26.74% of Athila + Chromovirus), *L. micranthus* (18.57% of Chromovirus) and *L. angustifolius* (12.2% of Maximus/SIRE + Chromovirus).

13.4.3 Evolutionary considerations on the dynamics of transposable elements in lupins.

Altogether the above observations provide interesting insights on the dynamics of the repetitive sequences in lupin genomes, particularly of their major component, LTR-retrotransposon elements. Overall, the same types of elements have been retrieved in both lupins and Genistoids (Mahé, 2009), which supports their ancient origin from the common ancestor of the Genistoid alliance (and earlier). Nevertheless, it is obvious that the lupin genomes experienced divergent evolutionary dynamics, as demonstrated by the remarkable variability of the species-specific profiles of elements observed among the few representatives of the closely related Old World lupins investigated. Some LTR retrotransposon families appear to have actively proliferated and accumulated in some species (e.g., *Athila*, *Chomovirus*, *Maximus-SIRE* elements, or even *Ogre/Tat*) while they

Accepted Manuscript

have been maintained at a low level in others. Most other families remained poorly represented throughout species. This strongly suggests that different processes and mechanisms regulating amplification, proliferation and clearance of these repeats (Lippman *et al.*, 2004; Ma & Bennetzen, 2004; Hawkins *et al.*, 2006, 2009; Slotkin & Martienssen, 2007; Lisch, 2009; Yaakov & Kashkush, 2012) have differentially operated in these species over the last ~10 Myr of their diversification. This was also shown in other plant systems (e.g.: (Hawkins *et al.*, 2006; Charles *et al.*, 2008; Hu *et al.*, 2011; Estep *et al.*, 2013; Piednoël *et al.*, 2013).

Accordingly, phylogenetic analyses of the most conserved RT sequences (which presumably represent the most recent and potentially yet functional LTR retrotransposons) provided substantial clues which support recent (after species divergence, likely < 8-10 Myr) and independent amplifications and accumulations (bursts) of the major *gypsy* and *copia* elements (*Athila*, *Chomovirus*, *Maximus-SIRE* and even *Ogre/Tat*) in the lupin genomes. The other less common retrotransposons (such as *Angela*, *TAR*, *Tork* and *Ale*), which seem still potentially able to proliferate in *L. angustifolius*, would represent families that either have low transposition rates or that have been specifically subjected to rapid purging processes following their expansion (Ma & Bennetzen, 2004; Bennetzen, 2005; Hu *et al.*, 2011; Renny-Byfield *et al.*, 2014; Vu *et al.*, 2015). This leaves open the way to different evolutionary trajectories for the later families. Moreover, some weakly represented *copia* families, such as

Accepted Manuscript

AleII/Retrofit and *Ivana/Oryco*, seem to have lost their ability to transpose. The yet recognizable but degenerated RTs found for these elements would likely represent the witnesses of ancient transposition events experienced by these families, which are ultimately prone to be erased from the DNA repetitive compartment of the smooth-seeded Mediterranean lupins. Another important evolutionary insight derived from the phylogenetic analysis of conserved RTs is that, not only various LTR retrotransposons families or subfamilies have been differentially accumulated among the different lupin species, but also that particular lineages of these families or subfamilies have been differentially amplified within each species, leading to the emergence of species-specific lineages of elements. For example, the major repeats in *L. luteus* essentially result from the recent proliferation of three species-specific *gypsy* lines (one from the *Athila* family, and two from the *Tekay* subfamily). Similarly, the prominent fraction of *gypsy* elements in *L. albus* results from the massive amplification of another specific lineage of the *Tekay* subfamily. Also, there are some evidence of likely recent lineage-specific amplification of *Maximus-SIRE* and *Athila* elements in *L. angustifolius*. Besides, a quick screening (results not shown) of available raw transcriptomic data sets from roots of *L. albus*, *L. luteus* and *L. mariae-josephae* (Keller *et al.*, 2018) provided some clues indicating a transcriptional activity for various TEs (including for some weakly represented families and ClassII elements). However, deeper investigations of more complete transcriptomic data sets are needed before making any reliable conclusion.

Accepted Manuscript

13.4.4 Tandem repeats may also greatly contribute to genome obesity and dynamics in lupins.

In the Mediterranean lupin genomes, the proportion of tandem repeats (excluding nrDNA) remarkably varies from 3 to 6% in *L. albus*, *L. luteus* and *L. micranthus*, to 26% in *L. angustifolius*. In contrast to the three former lupins and to the general trend in plants (Oliveira *et al.*, 2006; Barghini *et al.*, 2014; Heitkam *et al.*, 2015; Satović *et al.*, 2018), the proportion of tandem repeats is not only tremendous, but also is higher than that of transposable elements and represents more than half of the repetitive compartment in *L. angustifolius*. Also, it is noteworthy that even a low proportion of TRs may constitute a substantial fraction, equivalent to ~125 Mb in the large genome of *L. luteus*, for example. Among the best represented SSRs in the smooth-seeded Mediterranean lupins, three were almost always detected in the genomes (AGGATn, GATGAGn and GTTTAGGn). This is in agreement with the so-called "library hypothesis" evolution model which predicts that closely related species inherit from a common ancestor a same pool of satellites that are then independently amplified or lost in genomes (Fry & Salser, 1977; Oliveira *et al.*, 2006; Plohl *et al.*, 2012; Garrido-Ramos, 2017). Accordingly, our results revealed different SSR patterns which reflect the differential evolutionary dynamics experienced by these repeats in the lupin genomes. This is particularly well illustrated by the TR profile of *L. angustifolius*. In the latter species, the mi-

Accepted Manuscript

crosatellites *k*-mer < 10 bp) have been much more accumulated (16.66 %) than TRs with *k*-mers > 10 bp (2.26 %) in the genome, compared to its close Mediterranean relatives and to the lower frequencies reported for most other plants surveyed in the literature (Oliveira *et al.*, 2006; Barghini *et al.*, 2014; Heitkam *et al.*, 2015; Satović *et al.*, 2018). Even more striking, only one SSR (the 6-mer $GATGAG^n$) has been highly amplified and accumulated in *L. angustifolius* (estimated at 15.24% of the genome), whereas it is maintained at less than 0.6% in *L. luteus* and *L. micranthus*, and seem to have been erased from *L. albus*. Such contrasted frequencies of particular SSRs among genomes could be partially explained by divergences in the DNA repair system, as suggested by Oliveira *et al.*, (2006).

Alternatively, while SSRs are yet mostly ubiquitous in the smooth-seeded Mediterranean lupins (regardless of their various proportions), few distinct families of minisatellites and satellites have been each differentially and specifically amplified in either *L. albus*, *L. luteus* or *L. micranthus*. This suggests that they most likely results from dynamic and complex molecular processes and mechanisms that operated in the repetitive compartment following the diversification of the smooth-seeded Mediterranean lupins, which yielded species-specific satellite families (see: (Garrido-Ramos, 2015; Ávila Robledillo *et al.*, 2018). It has been suggested that differences in satellites types and abundance would play a role in speciation through the establishment of reproductive barriers between species, as

Accepted Manuscript

demonstrated in *Drosophila* (Ferree & Barbash, 2009). It is likely that the dramatic expansion of some satellites (alone and/or in conjunction with transposable elements) contributed to isolation and speciation processes among the Mediterranean lupins, as could be suggested by the striking divergent evolutionary dynamics observed following the separation of the closely related *L. luteus* (2*n* = 52; which preferentially accumulated a specific minisatellite and *gypsy* element) and *L. angustifolius* (2*n* = 42; which rather accumulated a remarkable amount of a particular hexamer SSR and *copia* elements). Additionally, these species-specific satellites represent an important basis for the development of cytogenetic markers to identify chromosomes, and to help understanding genome organization in lupins.

Another interesting observation highlighted from the screening of the available reference genome of *L. angustifolius* (NLL cv. Tanjil) is that all satellites *sensu lato* detected in our NLL accession (IPG2) were retrieved throughout all the twenty pseudochromosomes. Two different distribution patterns were observed. On one side, almost all the 5-, 6- and 7-mer SSRs observed in IPG2 are localized outside of the coding sequences in the gene-poor regions, with (GATGAG)n confirmed as the major SSR in this species. On the other side, the tandem repeats found in the coding regions are almost all SSRs with short monomers (*k*-mer < 4bp), of which the SSR $(CTT)^n$ is the most abundant and widespread throughout the pseudochromosomes. Such prevalence of trinucleotide SSRs in the coding regions indicate that

Accepted Manuscript

the other types with larger k-mers, which have a greater likelihood to induce frameshift mutations, are subjected to a counter-selection (Metzgar *et al.*, 2000; Toth, 2000). The screening of the NLL cv. Tanjil genome identified 1143 genes which contain a (CTT)n SSR with n equal to or greater than 4, which raises important questions to be addressed in order to evaluate their molecular, functional and evolutionary impact.

13.5 Conclusion and perspectives

This paper represents the first study on the repetitive compartment in lupin genomes, using low-depth high-throughput sequencing, reads clustering and annotation. The detailed analyses performed in four smooth-seeded Mediterranean lupins revealed a wide diversity of repeat types and allowed identification of the most abundant categories involved in shaping their genomes. In particular, only few *gypsy* (*Tekay*, *Athila*, *Ogre*) and *copia* (*Maximus-SIRE*) LTR retrotransposon families make up the prominent fraction of the repeats, which significantly contributes to genome size variation among species, regardless of their chromosome numbers and phylogenetic relationships. Interestingly, the results revealed that, not only retrotransposons, but also tandem repeats, such as microsatellites, may greatly contribute to genome obesity and dynamics in lupins, as demonstrated in *L. angustifolius*. Additionally, it has been shown that differential lineage-specific accumulation of transposable elements and/or tandem repeats occurred in lupins, which strongly supports that different processes and mechanisms regulating amplification, prolifera-

Accepted Manuscript

tion and clearance of repeats have differentially operated within the same genus and among closely related Mediterranean species over the last ~10-12 Myr.

Further extension of such evaluation to representatives of the different lupin clades circumscribed in the genus will undoubtedly provide a more accurate and enhanced overview of the repetitive components and their evolutionary dynamics following diversification, evolution and adaptation to diverse environmental conditions in both the Old and the New World. Additionally, the annotated raw material generated by this work represents a valuable basis to start building a repeats database specifically dedicated to the genus: (i) to accompany and facilitate assembly and annotation of novel lupin genomes; and (ii) to develop potentially useful genetic (e.g., microsatellites) and cytogenetic markers (e.g., specific minisatellites, satellites and TEs). This will help understanding structure, organization, repeats distribution and localization), variability, and evolution of the genomic landscape of lupins, and will enable comparative analysis with other legumes. Furthermore, the development of such database of repeats, using and combining genomic resources from both rapid low-depth high-throughput sequencing of various taxa and deep WGS of targeted species or accessions of particular interest, are of great importance to investigate and evaluate their structural, functional and evolutionary impact on genes, such as, for example, those responsive for important agronomical, adaptive and defense features.

Accepted Manuscript

Acknowledgements:

We are grateful to INEE-CNRS (France) and to the University of Rennes for their support to this work as part of the research program of the International Associated Laboratory "Ecological Genomics of Polyploidy" involving the University of Rennes (France) and the Iowa State University (Ames, USA). We thank Prof. Barbara Naganowska (Institut of Plant Genetics/PAS, Poznan, Poland) for kindly providing *L. angustifolius* seeds (IPG2 accession).

REFERENCES

Aïnouche A, Bayer RJ. 1999. Phylogenetic relationships in *Lupinus* (Fabaceae: Papilionoideae) based on internal transcribed spacer sequences (ITS) of nuclear ribosomal DNA. *American Journal of Botany* 86(4): 590-607.

Ainouche A, Bayer RJ, Misset M-T. 2004. Molecular phylogeny, diversification and character evolution in Lupinus (Fabaceae) with special attention to Mediterranean and African lupines. *Plant Systematics and Evolution* 246 (3–4), 211–222.

Alix K, Heslop-harrison JS. 2004. The diversity of retroelements in diploid and allotetraploid *Brassica* species. *Plant Molecular Biology* 54: 895–909.

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

Atnaf M, Yao N, Martina K, Dagne K, Wegary D, Tesfaye K. 2017. Molecular genetic diversity and population structureof Ethiopian white lupin landraces: Implicationsfor breeding and conservation. *PLoS ONE* 12(11):e0188696.

Ávila Robledillo L, Koblížková A, Novák P, Böttinger K, Vrbová I, Neumann P, Schubert I, Macas J. 2018. Satellite DNA in Vicia faba is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports* 8:5838.

Axtell MJ. 2013. Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology* 64: 137–159.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6.

Barghini E, Natali L, Cossu RM, 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 Biology and Evolution* 6: 776–791.

Bennett MD. 2005. Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. *Annals of Botany* 95: 45–90.

Accepted Manuscript

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* 42: 251–269.

Bennetzen JL. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115: 29–36.

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

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology* 65: 505–530.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27: 573–580.

Biémont C, Vieira C. 2006. Genetics: Junk DNA as an evolutionary force. *Nature* 443: 521–524.

Biscotti MA, Olmo E, Heslop-Harrison JS. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Research* 23(3): 415-420.

Cabello-Hurtado F, Keller J, Ley J, Sanchez-Lucas R, Jorrín-Novo JV, Aïnouche A. 2016. Proteomics for exploiting diversity of lupin seed storage proteins and their use as nutraceuticals for health and welfare. *Journal of Proteomics* 143: 57–68.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics* 14: 100–112.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* 17: 540–552.

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O, *et al.* 2008. Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat. *Genetics* 180: 1071–1086.

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* 509: 7–15.

Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* 99: 45–56.

Conterato IF, Schifino-Wittmann MT. 2006. New chromosome numbers, meiotic behaviour and pollen fertility in American taxa of Lupinus (Leguminosae): contributions to taxonomic and evolutionary studies. *Botanical Journal of the Linnean Society* 150: 229–240.

Devos KM. 2002. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Research* 12: 1075–1079.

Dolezel J, Bartos J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. *Cytometry* 51A: 127–128.

Accepted Manuscript

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603.

Eastwood RJ, Drummond CS, Schifino-Wittmann MT, Hughes CE. 2008. Diversity and evolutionary history of lupins–insights from new phylogenies. *Lupins for health and wealth: 12th International Lupin Conference*: 10.

Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity* 110: 194–204.

Ferree PM, Barbash DA. 2009. Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila (MAF Noor, Ed.). *PLoS Biology* 7: e1000234.

Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A. 1992. *Ty1–copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. *Nucleic Acids Research* 20: 3639–3644.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in *de novo* annotation approaches (Y Xu, Ed.). *PLoS ONE* 6.

Fry K, Salser W. 1977. Nucleotide Sequences of HS-a Satellite DNA from Kangaroo Rat *Dipodomys ordii* and Characterization of Similar Sequences in Other Rodents. *Cell* 12: 1069–1084.

Garrido-Ramos MA. 2015. Satellite DNA in plants: more than just rubbish. *Cytogenetic and Genome Research* 146: 153–170.

Garrido-Ramos M. 2017. Satellite DNA: an evolving topic. *Genes* 8: 230.

Gladstones JS, Atkins CA, Hamblin J (Eds.). 1998. *Lupins as crop plants: biology, production, and utilization*. Wallingford, Oxon, UK ; New York, NY, USA: CAB International.

Grandbastien M-A, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa A-PP, Le QH, Melayah D, Petit M, Poncet C, *et al.* 2005. Stress activation and genomic impact of *Tnt1* retrotransposons in Solanaceae. *Cytogenetic and Genome Research* 110: 229–241.

Gregory TR. 2005. The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Annals of Botany* 95: 133–146.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest Angiosperm Genomes Found in Lentibulariaceae, with Chromosomes of Bacterial Size. *Plant Biology* 8: 770–777.

Hane JK, Ming Y, Kamphuis LG, Nelson MN, Garg G, Atkins CA, Bayer PE, Bravo A, Bringans S, Cannon S, *et al.* 2017. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. *Plant Biotechnology Journal* 15: 318–330.

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* 106: 17811–17816.

Accepted Manuscript

Heitkam T, Petrasch S, Zakrzewski F, Kögler A, Wenke T, Wanke S, Schmidt T. 2015. Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest Camellia japonica. *Chromosome Research* 23: 791–806.

Hoang DT, Chernomor O, von Haeseler A, Quang Minh B, Sy Vinh L. 2017. Ufboot2: Improving The Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 32: 518–522.

Hosaka A, Kakutani T. 2018. Transposable elements, genome evolution and transgenerational epigenetic variation. *Current Opinion in Genetics & Development* 49: 43–48.

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. *BMC Plant Biology* 10.

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.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* 421: 163–167.

Jiang N, Feschotte C, Zhang X, Wessler SR. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Current Opinion in Plant Biology* 7: 115–119.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. 2000. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences* 97: 6603–6607.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.

Kamphuis LG, Hane JK, Nelson MN, Gao L, Atkins CA, Singh KB. 2015. Transcriptome sequencing of different narrow-leafed lupin tissue types provides a comprehensive uni-gene assembly and extensive gene-based molecular markers. *Plant Biotechnology Journal* 13: 14–25.

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

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Keller J, Imperial J, Ruiz-Argüeso T, Privet K, Lima O, Michon-Coudouel S, Biget M, Salmon A, Aïnouche A, Cabello-Hurtado F. 2018. RNA sequencing and analysis of three *Lupinus* nodulomes provide new insights into specific host-symbiont relationships with compatible and incompatible *Bradyrhizobium* strains. *Plant Science* 266: 102–116.

Kroc M, Koczyk G, Święcicki W, Kilian A, Nelson MN. 2014. New evidence of ancestral polyploidy in the Genistoid legume *Lupinus angustifolius* L. (narrow-leafed lupin). *Theoretical and Applied Genetics* 127: 1237–1249.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* 19: 1639–1645.

Accepted Manuscript

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

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

Lerat E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* 104: 520–533.

Levinson G, Gutman G. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203–221.

Li Y-C. 2004. Microsatellites Within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution* 21: 991–1007.

Lim KG, Kwoh CK, Hsu LY, Wirawan A. 2013. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Briefings in Bioinformatics* 14: 67–81.

Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, Lavine K, Mittal V, May B, Kasschau KD, *et al.* 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430: 471–476.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. *Annual Review of Plant Biology* 60: 43–66.

Lisch D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics* 14: 49–61.

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

Lönnig W-E, Saedler H. 1997. Plant transposons: contributors to evolution? *Gene* 205: 245–253.

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics & Development* 49: 70–78.

Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F, Bauersachs S, *et al.* 2015. Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Reports* 10: 551–561.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences* 101: 12404–12410.

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*. *BMC Genomics* 8.

Mahé F. 2009. Phylogénie, éléments transposables et évolution de la taille des génomes chez les lupins.

Mahé F, Pascual H, Coriton O, Huteau V, Navarro Perris A, Misset M-T, Aïnouche A. 2011. New data and phylogenetic placement of the enigmatic Old World lupin: *Lupinus mariae-josephi* H. Pascual. *Genetic Resources and Crop Evolution* 58: 101–114.

Accepted Manuscript

Mayer KFX, Martis M, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, *et al.* 2011. Unlocking the Barley Genome by Chromosomal and Comparative Genomics. *The Plant Cell* 23: 1249–1263.

Metzgar D, Bytof J, Wills C. 2000. Selection Against Frameshift Mutations Limits Microsatellite Expansion in Coding DNA. *Genome Research*: 9.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* 37: 997–1002.

Naganowska B. 2003. Nuclear DNA Content Variation and Species Relationships in the Genus Lupinus (Fabaceae). *Annals of Botany* 92: 349–355.

Naganowska B, Wolko B, Śliwińska E, Kaczmarek Z, Schifino-Wittmann MT. 2005. 2C DNA variation and relationships among New World species of the genus Lupinus (Fabaceae). *Plant Systematics and Evolution* 256: 147–157.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32: 268–274.

Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research* 45.

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

Novak 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.

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC. 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology* 29: 294–307.

Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to Angiosperm evolution and diversity. *Genome Biology and Evolution* 5: 1886–1901.

Orgel LE, Crick FH, Sapienza C. 1980. Selfish DNA. *Nature* 288: 645–646.

O'Rourke JA, Yang SS, Miller SS, Bucciarelli B, Liu J, Rydeen A, Bozsoki Z, Uhde-Stone C, Tu ZJ, Allan D, *et al.* 2013. An RNA-Seq Transcriptome Analysis of Orthophosphate-Deficient White Lupin Reveals Novel Insights into Phosphorus Acclimation in Plants. *Plant Physiology* 161: 705–724.

Parra-González LB, Aravena-Abarzúa GA, Navarro-Navarro CS, Udall J, Maughan J, Peterson LM, Salvo-Garrido HE, Maureira-Butler IJ. 2012. Yellow lupin (Lupinus luteus L.) transcriptome sequencing: molecular marker development and comparative studies. *BMC Genomics* 13: 425.

Pellicer J, Oriane Hidalgo, Steven Dodsworth, Ilia Leitch. 2018. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes* 9: 88.

Accepted Manuscript

Petes TD. 1980. Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. *Cell* 19: 765–774.

Piednoël M, Carrete-Vega G, Renner SS. 2013. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. *The Plant Journal* 75: 699–709.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, *et al.* 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research* 16: 1262–1269.

Plohl M, Mestrovic N, Mravinac B. 2012. Satellite DNA Evolution. In: Garrido-Ramos MA, ed. Genome Dynamics. Basel: S. KARGER AG, 126–152.

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W & Wurm Y. 2015. Sequenceserver: a modern graphical user interface for custom BLAST databases. *biorxiv* doi: 10.1101/033142.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology* 1: 166–175.

Raman R, Cowley RB, Raman HD, Luckett DJ. 2014. Analyses using SSR and DArT molecular markers reveal that Ethiopian accessions of white lupin (*Lupinus albus* L.) represent a unique gene pool. *Open J Genet.* 4:87–98.

Renny-Byfield S, Chester M, Kovarik A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novak P, Chase MW, *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, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, Wang X, Paterson AH, Wendel JF. 2014. Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. *Genome Biology and Evolution* 6: 559–571.

Renny-Byfield S, Wendel JF. 2014. Doubling down on genomes: polyploidy and crop plants. *American Journal of Botany* 101: 1711–1725.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports* 6.

Satović E, Vojvoda Zeljko T, Plohl M. 2018. Characteristics and evolution of satellite DNA sequences in bivalve mollusks. *The European Zoological Journal* 85: 94–103.

Schmuths H. 2004. Genome Size Variation among Accessions of *Arabidopsis thaliana. Annals of Botany* 93: 317–321.

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

Shi J, Huang S, Fu D, Yu J, Wang X, Hua W, Liu S, Liu G, Wang H. 2013. Evolutionary Dynamics of Microsatellite Distribution in Plants: Insight from the Comparison of Sequenced *Brassica*, Arabidopsis and Other Angiosperm Species (BA Vinatzer, Ed.). *PLoS ONE* 8: e59988.

Accepted Manuscript

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, *et al.* 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539–539.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* 8: 272–285.

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

Streelman JT, Kocher TD. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiological Genomics* 9: 1–4.

Thomas CA. 1971. The Genetic Organization of Chromosomes. *Annual Review of Genetics* 5: 237–256.

Toth G. 2000. Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. *Genome Research* 10: 967–981.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* 13: 36–46.

Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. 2017. Comparative genome-wide analysis of repetitive DNA in the genus Populus L. *Tree Genetics & Genomes* 13: 96.

Vicient CM, Suoniemi A, Anamthawat-Jónsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH. 1999. Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *The Plant Cell* 11: 17.

Vu GTH, Schmutzer T, Bull F, Cao HX, Fuchs J, Tran TD, Jovtchev G, Pistrick K, Stein N, Pecinka A, *et al.* 2015. Comparative Genome Analysis Reveals Divergent Genome Size Evolution in a Carnivorous Plant Genus. *The Plant Genome* 8(3): 1-14.

Wajid B, Serpedin E. 2012. Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers. *Genomics, Proteomics & Bioinformatics* 10: 58–73.

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. *Genome Biology* 17: 37.

Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences* 103: 17600–17601.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R, Mayer KFX, Paux E, Choulet F. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* 19(1):103.

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 Reviews Genetics* 8: 973–982.

Wolko B, Weeden NF. 1989. Estimation of *Lupinus* genome polyploidy on the basis of isozymic loci number. *Genetica Polonica* 30: 165-171.

Accepted Manuscript

Wu DD, Ruban A, Fuchs J, Macas J, Novak P, Vaio M, Zhou YH, Houben A. 2019. Nondisjunction and unequal spindle organization accompany the drive of *Aegilops speltoides* B chromosomes. *New Phytologist* 223:1340–1352

Yaakov B, Kashkush K. 2012. Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Molecular Biology* 80: 419–427.

Accepted Manuscript