University of Leicester

Nuclear and chloroplast genome diversity in apomictic microspecies of *Taraxacum*

A thesis submitted to the University of Leicester

for the degree of Doctor of Philosophy

By

Rubar Hussein M.Salih M.Sc.

February 2017

Department of Genetics

Nuclear and chloroplast genome diversity in apomictic microspecies of Taraxacum

Rubar Hussein M.Salih

Abstract

Whole genomic survey sequences were obtained for *Taraxacum obtusifrons* Markl. (O978); *T. stridulum* Trávniček ined. (S3); and *T. amplum* Markl. (A978), three apomictic triploid (2n=3x=24) dandelions from the *T. officinale* agg. (Asteraceae) Retroelement-based markers and chloroplast data showed that S3 and O978 are genetically the most similar microspecies. Genomic diversity in *Taraxacum* and also *Hieracium* was high, discriminating species but not showing phylogeny; major groups of retroelements were abundant in both genera. The chloroplast genomes of accessions O978 and S3 were identical. Repetitive DNA including transposable elements (TEs) are dynamically evolving in genomes, but their variability and abundance make them challenging to study using molecular biology. In the current study, we used the whole genomic sequences to investigate the repetitive structure, diversity and components of the three closely related *Taraxacum* accessions. Analysis of about 45Gb sequence (10x to 20× genome coverage) of three closely related *Taraxacum* microspecies, were analysed by graph-based clustering of the raw reads (using the program RepeatExplorer) and frequency analysis of all DNA motifs possible for various motif lengths (k-mer analysis). Different DNA motif lengths were evaluated and complemented the graph-based results. Graph-based clustering showed that many of the *Taraxacum* microspecies repeats consist of Ty1-copia (13-16%) and Ty3-gypsy (10-14%) family retroelements, while DNA transposons were rare. Unclassified repetitive DNA sequence clusters were investigated. *In situ* hybridization was used to localize major repetitive DNA families on chromosomes. Apart from 5S and 45S rDNA and telomere sequences, few tandemly repeated DNA motifs were found, although a 49bp repeat was found at some centromeres. There were differences between the three *Taraxacum* microspecies in genomic proportions and locations for repetitive DNA types suggesting many sequence motifs are evolving rapidly with increasing or decreasing copy numbers. A class of repetitive DNA has been recognized as Passively Amplified DNA Sequences, PADS.

Declaration

I hereby declare that no part of this thesis has been previously submitted to this or any other University as part of the requirements for a higher degree. The content of this thesis is the result of my own experimentation and data analysis unless otherwise acknowledged in the text or by reference.

The work was conducted in the Department of Genetics, University of Leicester, during the period January 2013 to December 2016.

Signed .....

Rubar M.Salih

Dedication

A special feeling of gratitude to my loving parents, my Father and Mother whose words of encouragement and push for tenacity ring in my ears. My sisters (Bahar, Kazhal, Nigar, Nian) and Brothers (Gharib, Abdullah, Mzafar, Daban) have never left my side and are very special. I dedicate this dissertation to my many nieces and nephews who made my life beautiful.

This thesis work also is dedicated especially to my husband, Amed, who has been a constant source of support and encouragement. I am truly thankful for having you in my life. Thank you to been my only friend and best cheerleader.

Acknowledgements

After an intensive period of 4 years and 7 months, today is the day: writing this note of thanks is the finishing touch on my thesis. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level. Writing this thesis has had a big impact on me. I would like to reflect on the people who have supported and helped me so much throughout this period.

*First and foremost, my praises and thanks to the **Allah**, the Almighty, whom I owe my very existence, and who gave me the strength and courage to reach this stage and helped me at every step in my life.*

*This thesis could not be done and been possible without the funding provided from the Higher Committee for Education Development in Iraq (**HCED**), which I am very grateful.*

*I would like to express my deep and sincere gratitude to my research supervisor **Prof. J.S (Pat) Heslop-Harrison**, for giving me the opportunity to do research and providing invaluable guidance throughout this research. It was a great privilege and honour to work and study under his guidance, and thankful for his help and critical comments about writing of my thesis and scientific papers.*

*I thank to **Dr. Richard Gornall** for his assistance, expert knowledge of dandelions and valuable comments and discussion to understand agamospecies life cycle. I am thankful also to Associate **Dr. John Bailey**, for his expert on preparations of metaphase chromosome slides and for taking time to read my thesis and offer suggestions. I thank **Dr. Luboš Majeský** from Palacký University, Czech Republic for providing seeds of plants necessary for my research. I thanks to **Dr. Trude Schwarzacher** for her assistance in lab working and providing the plant materials I need for my study.*

*My special thanks go to my **parents** for their love, caring and sacrifices for educating and preparing me for my future, your prayer for me was what sustained me thus far words cannot express how grateful I am for having you in my life.*

*I am very much thankful to my **husband** for his love, extraordinary patience, understanding, and continuing support to complete this research work, without which the life would not be so glamorous.*

*I would also like to thanks my fellows in the lab and research colleagues, **Mr. Ramesh Patel**, **Dr. Niaz Ali**, **Dr. Nauf Alsayied**, **Dr. Stuart Desjardins**, **Dr. Chetan Patokar**, **Dr. Jotyar Muhammed**, **Osamah Alisawi, Noorhariza Mohd Zaki** for nice and friendly environment.*

It is not possible to mention the names of all those people who contributed to this piece of work but I fully appreciate your valuable contributions. I therefore wish to sincerely thank all such people.

Thank you very much, everyone!

Rubar M.Salih

Leicester 20-02-17

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1. Apomictic reproduction in plants

1.1.1. Sexual vs asexual reproduction

In plants, the standard sexual life cycle involves an alternation between sporophytic (2n) and gametophytic (n) generations (Figure 1.1). In the angiosperm, flower male pollen mother cells (2n) in the anther and diploid female embryo sac mother cells in the ovule undergo meiosis to produce male microspores and female megaspores, each with a haploid chromosome complement. Both male microspore and female megaspores undergo mitosis to produce gametophytes that consist of a defined number of cells. Each pollen grain contains a vegetative nucleus and one or two sperm cells at anthesis (depending on the species). Inside the ovule, the megagametophyte with its egg cell and dikaryotic central cell are produced by a variety of developmental patterns, depending on the species (Maheshwari 1950; Dumas and Mogensen 1993). So, sexual reproduction (amphimixis) is taking place by fusion of the egg cell with one sperm cell and fusion of the dikaryotic central cell with the second sperm cell, namely double fertilization, which leads to a diploid embryo and triploid endosperm respectively (Koltunow 2012).

Although, in some flowering plants, seed production undergoes some modification and development to produce viable seeds without fertilization by a process known as apomixis s.s. (apomixis *sensu stricto* = agamospermy = asexual reproduction through seed), (Nogler 1984), and it is almost synonymous with asexual

reproduction. Asexual reproduction (apomixis) through seeds occurs in flowering plants through different mechanism.

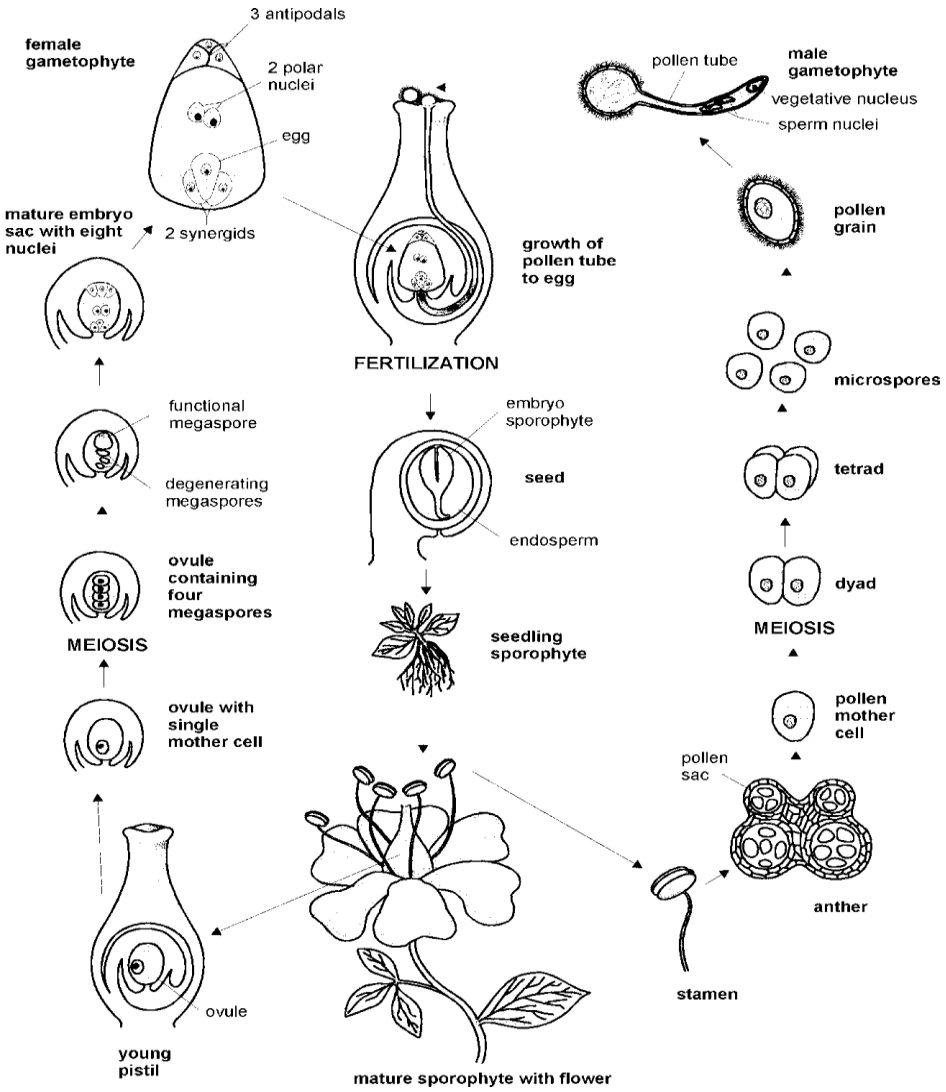


Figure 1.1. The sexual (amphimixis) life cycle in flowering plants (Singh 2003).

Apomixis can be defined as asexual (agamic) reproduction that results in seed production (as distinct from vegetative propagation) without fertilization, and gives rise to offspring identical to the maternal parent (Carneiro *et al.* 2006) which are called metromorphous offspring (Nogler 1984). Female meiosis fails at an early stage of the division so that it leads to the production of a diploid spore, which then grows into a diploid megagametophyte that gives rise to a diploid egg. This egg develops through parthenogenesis to produce an embryo without fertilization (Solntseva 1976; Nogler 1984; Asker and Jerling 1992; Koltunow 1993; Richards 1997; Savidan 2000; Carneiro *et al.* 2006; Ozias-Akins 2007). The term apomixis is synonymous with the term

agamospermy, which specifically refers to the production of seeds without the sex involvement, and term “agamic complex”, derived from agamospermy define mixed population of sexual species and apomictic clones (Stebbins 1950; Nogler 1984).

In 1869, the first known genetic study of an apomictic plant done by Gregor Mendel. He chose *Hieracium* as one of the plants for experiments on the laws of inheritance (Nogler 2006). In contrast with *Pisum,* the F1 hybrids that he observed from *Hieracium* showed clear extensive segregation. In contrast, the F2 progeny was uniform and did not segregate (there is no genetic variation in the second generation because of the maternal type progeny generated in the first generation). Mendel noted that in these two systems, *Pisum* and *Hieracium* “almost opposed behaviour both of them has the outcomes of a higher universal law”. After Mendel, Ostenfel along with Rosenberg (1906, 1907), observed that in *Hieracium* there is the expression of apomixis, and they published their notes 40 years after Mendel’s experiments (Bicknell and Koltunow 2004). According to Whitton *et al*. (2008), Smith (1841) described apomixis for the first time.

1.1.2. Types of asexual reproduction (apomixis)

Apomixis types share one or more of this three mechanisms: 1- Apomeiosis: failure of the meiotic division which leads to production of an unreduced spore that germinates to produce an unreduced megagametophyte; 2- parthenogenesis: embryo development from an unreduced egg-like cell without fertilization; 3- production of endosperm without fertilization (autonomous endosperm formation) or by normal fertilization (Koltunow 1993, Carman 1997; Koltunow and Grossniklaus 2003).

Apomixis can be divided into two distinct pathways according to the type and fate of the tissues that produce the embryo, (Koltunow and Grossniklaus 2003). The passway will be sporophytic apomixis type if the unreduced cell gives rise directly to an embryo, or will be gametophytic apomixis if an unreduced cell gives rise to a diploid embryo sac. Figure (1.2) shows the differences between sexual and asexual (apomictic) passways in angiosperm ovules, it shows the most important feature is meiosis division, which leads to reduce the ploidy level of the cells from diploid to haploid.

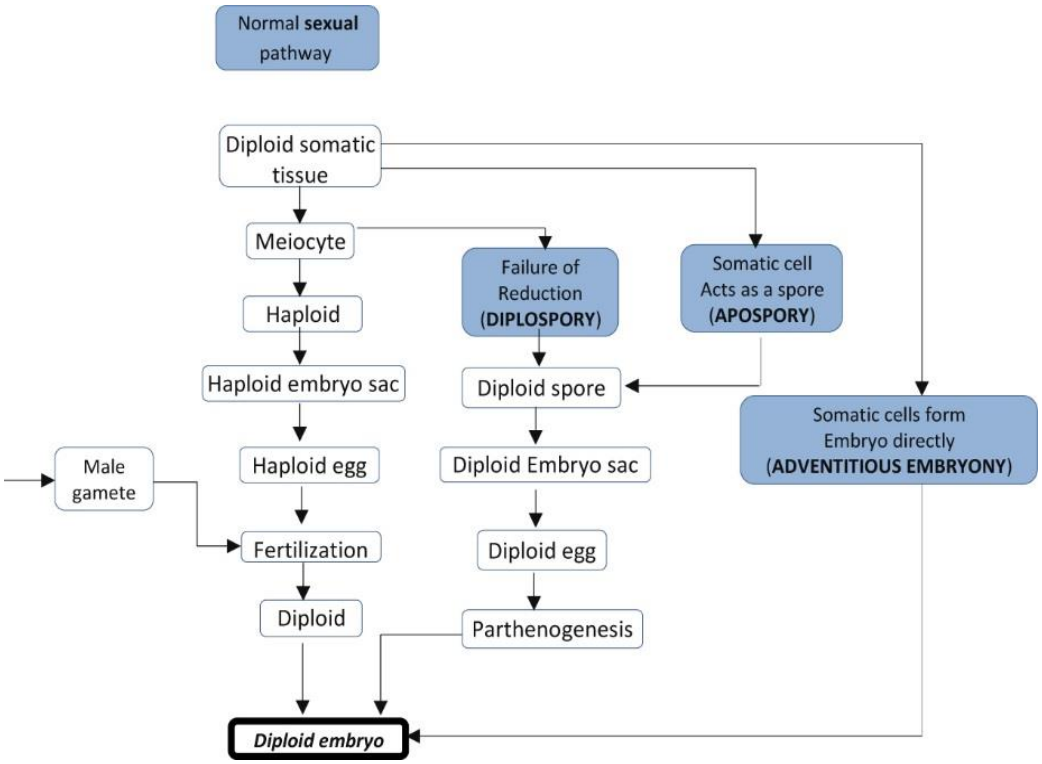


Figure 1.2. Comparison between sexual and apomictic reproduction (Heslop-Harrison 1972).

1.1.2.1. Gametophytic apomixis

Embryos develop from the gametophytic phase, unreduced gametophytic tissue (Van Dijk *et al.* 2009). The most important feature in gametophytic apomixis is apomeiosis or the avoidance of female meiosis (Asker and Jerling 1992; Nogler 2006). The diploid egg cell develops autonomously without any genetic recombination or fertilization, and results in a seed genetically identical to its maternal parent (Carneiro *et al.* 2006). Endosperm formation occurs by fertilization rarely without fertilization (Koltunow and Grossniklaus 2003). Gametophytic apomixis can be further subdivided into two categories, diplospory and apospory, depending on the origin of the cells that generate the diploid female gametophyte (megagametophyte initiation cells), (Koltunow and Grossniklaus 2003; Singh 2003). Figure 1.3 shows classification of apomixis reproduction depending on the source tissue that produce diploid embryo sac.

1.1.2.1.1. Diplospory

In diplospory (also known as generative apospory - Singh 2003), a normal meiosis reductional division is replaced by a non-reductional division (Van Dijk *et al.* 2009).

Diplosporus apomixis can be subdivided into two types according to whether the dominant and active type of cell division is mitotic or meiotic. They are also named after the genera in which they were first discovered: *Antennaria* type (mitotic) and *Taraxacum* type (meiotic), (Singh 2003), (Figure 1.3).

In meiotic diplospory, the megaspore mother cell undergoes meiotic prophase-1, but the chromosomes fail to separate during anaphase, resulting in the formation of a diploid nucleus. This is followed by a mitotic division and produces the usual dyad cell (chalaza dyad) but with an unreduced chromosome number which ends after division three with the formation of a polygonum-type embryo sac. Meiotic diplospory can be seen mostly in Asteraceae in many genera like *Taraxacum* spp. (Van Dijk *et al.* 1999), *Erigeron annus* (Noyes and Rieseberg 2000; Noyes *et al.* 2007), (Figure 1.3).

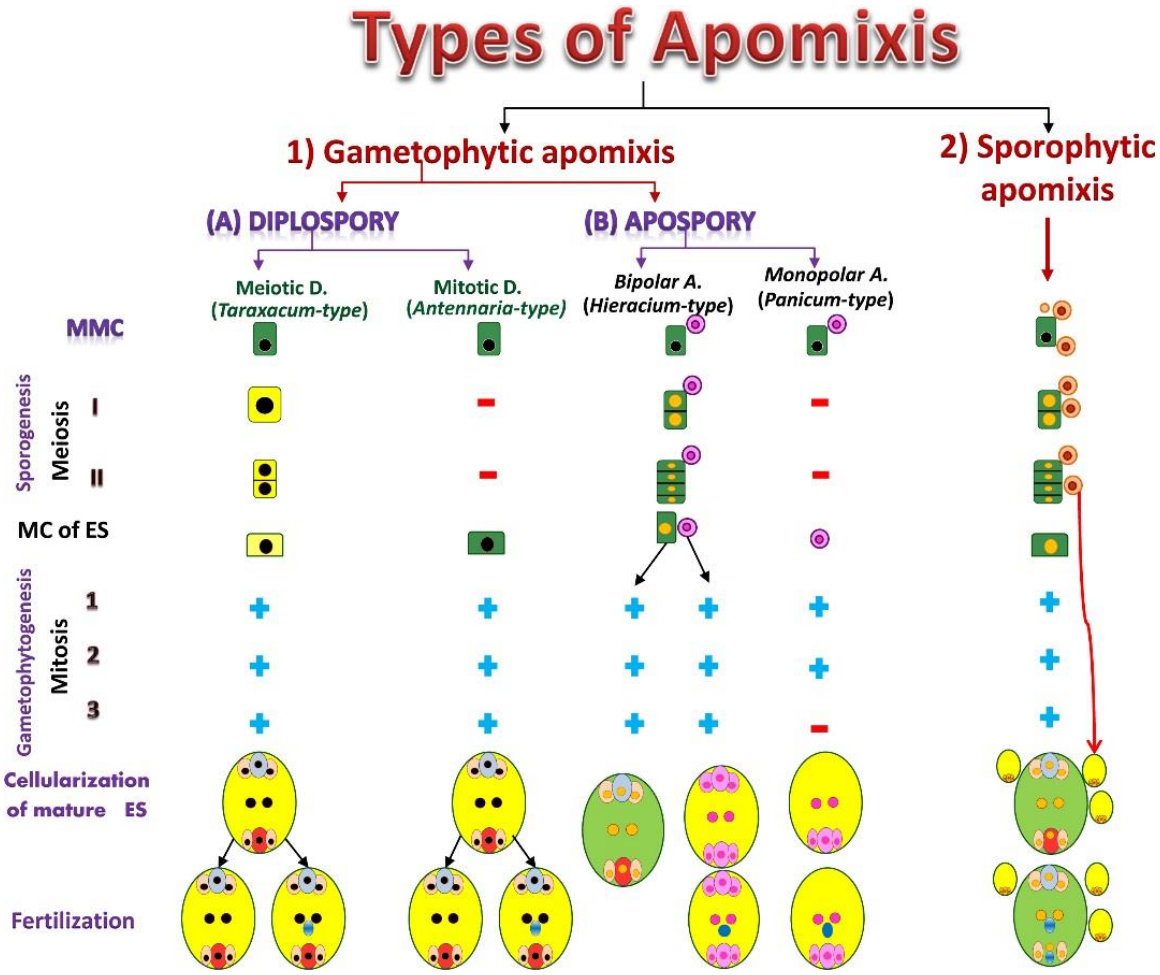


Figure 1.3. Asexual reproduction classifications and reproductive pathways in sexual and some asexual variants. Green color: sexual pathway; yellow: asexual pathway; purple colour: Aposporous Initial-AI; orange colour: diploid somatic cells (integument or nucellus); MMC: mega spore mother cell, ES: embryo sac.

In mitotic diplospory or *Antennaria*-type, the megaspore mother cell does not enter meiosis but directly enters three mitoses to form a polygonum-type diploid eight-nucleate embryo sac. This type of apomictic diplospory is widely distributed taxonomically (Singh 2003) such as in some *Hieracium* spp. and some *Antennaria* spp. (Van Dijk *et al.* 2009), (Figure 1.3).

There are other types of apomictic diplospory derived from the *Antennaria* and *Taraxacum* types, such as, *Eragrostis* (Poaceae) in which there is a reproductive pathway like the *Antennaria* type, but the third mitotic division does not occur and, as a result, the embryo sac contains four nuclei.

Furthermore, the *Ixeris* (*Ixeris dentata*) type is similar to the *Taraxacum* type, but the megaspore mother cell undergoes a second meiotic division and avoids cytoplasmic division to produce two unreduced nuclei in place of the four nuclei in the *Taraxacum* type. Finally, after two mitotic divisions, an eight-nucleate embryo sac is formed.

In some onion species (*Allium nutans*) the somatic chromosomes in the nuclei undergo an extra round of DNA replication before starting meiosis, which results in doubling of the somatic chromosome number so that a mitotic S phase is duplicated prior to meiosis and 4x G1-phase megaspore mother cells are produced (Asker and Jerling 1992), following by normal meiosis forming tetrads that keep the same set of maternal chromosomes; finally, two mitoses in the chalaza dyad form an eight-nucleate embryo sac (Grimanelli *et al.* 2001a; Singh 2003; Van Dijk *et al.* 2009). This process is known as automixis in parthenogenetic animals (Suomalainen *et al.* 1987).

So, as a result of all types of diplospory, megaspore mother cells fail to undergo the meiotic division, so that crossovers between chromosomes are avoided, resulting in development of Polygonum-type embryo sacs that avoided meiosis (apomeiosis) including egg-like cells that initiate embryogenesis without fertilization (2*n*+0) (Noyes 2007; Tucker and Koltunow 2009) and that are genetically identical to the mother plant (Nogler 1984). Endosperm commonly develops by fertilization of the central cell (pseudogamy) to produce viable seed in diplospory, however in other cases

endosperm develops autonomously (Koltunow *et al.* 1995; Koltunow and Tucker 2008; Tucker and Koltunow 2009).

Apomeiotic apomixis, pollen grains have a reduced ploidy level in comparison with the egg cells of the same plant, because the apomeiotic mechanism is specific to the female sex function (Van Dijk *et al.* 2009).

1.2.1.1.2. Apospory

Previously named "somatic apospory" (Nogler 1984), is similar to diplospory except that the unreduced embryo sac forms directly from a somatic cell of the ovule wall (Nucellar or chalaza cell of the ovule) near the megaspore mother cells (Nogler 1984; Asker and Jerling 1992; Koltunow and Grossniklaus 2003).

In apospory within the ovule, besides the normally reduced megagametophyte (n), from a somatic cell (namely Aposporous Initial-AI), an unreduced (2n) megagametophyte is generated. The 2n aposporous gametophyte forms an unreduced egg cell, which develops through parthenogenesis into an embryo being genetically identical to the mother plant (Figure 1.3; Van Dijk *et al.* 2009).

Aposporous apomixis comprises two types: the bipolar *Hieracium*-type and the monopolar *Panicum*-type (Singh 2003).

In the bipolar or *Hieracium*-type, one or more somatic cells (nucellar) in close proximity to the megaspore mother cell, or megaspores differentiate into aposporous initial (AI) cells, bypass meiosis and directly undergo three mitoses (apomeiosis) to give rise to eight-nucleate unreduced embryo sacs; concurrently, such ovules may also contain a normal reduced polygonum-type embryo sac. The unreduced embryo sac survives while the reduced embryo sac may degenerate (Khokhlov 1976), (Figure 1.3). The *Hieracium* apospory type can be seen in the *Hieracium piloselloides* (Koltunow *et al*. 1998), *Hypericum*, *Poa*, *Ranunculus auricomus*, *Crepis*, *Hierochloe* and *Beta* (Nogler 1984).

In the monopolar *Panicum*-type the embryo sac is four-nucleate after only two mitotic division (Grimanelli *et al.* 2001a), (Figure 1.3). This type was observed in *Panicum maximum* (Warmke 1954), subfamily Panicoideae, tribe Andropogoneae

(Nogler 1984; Asker and Jerling 1992), *Bothriochloa, Dichanthium, Capillipedium*, *Cenchrus*, *Chloris*, *Digitaria*, *Eriochloa*, *Heteropogon*, *Hyparrhenia*, *Paspalum*, *Pennisetum squamulatum* (Roche *et al*. 1999), *Sorghum*, *Themeda*, and *Urochloa* (Nogler 1984).

In aposporous apomixis, as in diplospory, egg-like cells within the unreduced embryo sacs initiate embryogenesis without fertilization and generate offspring that are genetic clones of the maternal plant. Depending on the species, the endosperm can initiated with or without fertilization.

In aposporous apomicts, apomixis initial cells can appear at different times and with different frequencies during ovule development, and the presence of multiple apomixis initial cells (AI) cells, and in some cases sexual and aposporous pathways coexist. If the formation of embryos is successful in both pathways then it can lead to polyembryonic seed in ovules of aposporous apomicts.

1.1.2.2. Sporophytic apomixis

In the sporophytic pathway, somatic embryos are formed directly from diploid somatic cells within the sporophytic tissue, nucellus or rarely integument cells, adjacent to or surrounding the gametophyte (a reduced embryo sac) (Koltunow and Grossniklaus 2003). Because somatic embryos usually arise in parallel with sexually formed embryos, this type of apomixis also called adventitious embryony (Naumova 1993). Thus, according to Koltunow (1993) and Nogler (1984), the sporophytic pathway can occur side by side with normal sexual reproduction so that the seed comprises both sexual and apomictic embryos (polyembryony; Figure 1.3). Sporophytic agamosperms are mostly diploid and sexually fertile (Richards 2003). In sporophytic apomixis, no parthenogenesis is involved (Van Dejk 2009). Thus, it excludes the alternation of generations pathway (Grant 1981). The embryo can survive if it is accompanied by an endosperm (Koltunow and Grossniklaus 2003).

Sporophytic apomixis is little known and poorly studied. However it is the most widespread form of agamospermy, and has been recorded in more than 250 species of 121 genera belonging to 57 families of flowering plants (Naumova 1993; Carman

1997), mostly in tropical trees, often fruit trees, including *Citrus*, *Euphorbia*, *Mangifera*, *Malus*, *Ribes*, *Beta* and several genera of grasses, cacti, and orchids.

Sporophytic apomixis differs from apospory in that cell (the nucellus and integuments) do not enter a gametophytic phase, meaning that no egg cell is formed, but remain at the diploid level and produce an embryo (somatic embryo) directly.

1.1.3. Facultative and obligate apomixis

Some natural apomicts have the ability to reproduce both sexually or asexually so they classified as a facultative apomictic (Hand and Koltunow 2014), meaning that in facultative apomixis sexual reproduction is not completely eliminated; females can produce offspring either sexually or via asexual reproduction (Bell 1982). Facultative parthenogenesis is extremely rare in nature, with only a few examples of animal taxa, such as triploid lizards, being capable of facultative parthenogenesis. Facultative parthenogenesis is believed to be a response to a lack of a viable male. A female may undergo facultative parthenogenesis if a male is absent from the habitat or if it is unable to produce viable offspring. Obligate apomixis occurs if 100% of the offspring are identical to the maternal parent in organism which reproduces exclusively through asexual reproduction (Stelzer *et al.* 2010).

The majority of apomictic plants produce functional pollen grains, meaning that apomictic plants are not cutoff completely from sexual reproduction. Therefore, in a geographical region in a mixed population of sexual and apomictic plants, sexual plants can produce apomictic offspring, and vice versa for apomictic plants, resulting in a sexual–apomixis cycle.

Apomixis in *Taraxacum officinale* is obligate, and facultative apomixis has rarely been reported in other *Taraxacum* species (Richards 1973, Van Dijk *et al.* 2009).

1.1.4. Species concept in apomicts

Determination of the genus or family of plants that undergo apomixis is easy. However, it is not easy to determine the species that belong to each apomictic genus because the taxonomy of such agamic complexes is difficult and contentious

(Dickinson 1998). These difficulties are caused by morphological polymorphism. There are arguments between scientists about the classification of apomictic plants. Some scientists classify morphologically distinct clones as full species, whereas others treat them as microspecies, and some others as individual genotypes. According to Van Dijk *et al.* (2009) the clone forms a level above the individual in asexual organisms. It was observed that many different morphologically clones can be found growing together in one population. These differences in classification of apomictic plants caused in different record of species that undergo asexual reproduction have been published so far. It might be a few hundred or thousands of taxa, due to the species concept used by different taxonomists. It is more difficult when trying to classify facultative apomict individual taxonomically. Because of outcrossing, genetically different evolutionary lines will be continually produced, mixed with constantly renewed and unique genotypes, which leads to the breakdown of the species concept otherwise used to classify morphological individuals in the case of obligate apomicts.

Because *Taraxacum* is one of the most widespread plants that undergoes apomictic reproduction, its classification is of interest to scientists. There is disagreement between scientists over how to treat *Taraxacum* lineages as species and thus how to investigate evolutionary pathways. To try to ameliorate this situation, scientists have recognised a number of sections, where each section consists of closely related agamospecies; the section replaces the species as the basic taxon.

1.1.5. Taxonomic occurrence of apomixis

Apomixis is widespread across the plant kingdom from algae to angiosperms (Asker and Jerling 1992; Singh 2003); however, it is apparently absent from gymnosperms (Bicknell and Koltunow 2004, Carman 1997). Mogie (1992) estimated that about 0.1% of all flowering plants are apomictic and, according to APGII (2003), it is prevalent among ca. 457 angiosperm families. In addition, Carman (1997) made a list of more than 330 genera in which apomixis occurs; two-thirds of them exhibited sporophytic apomixis (adventitious embryony) and ca. 126 genera exhibited gametophytic apomixis. Among angiosperms, it occurs in magnoliids, monocotyledons and eudicotyledons. Three plant families in particular show high frequencies of apomixis:

Poaceae, Asteraceae and Rosaceae. These three families together include 75% of all apomictic plants, representing 15% of all angiosperm species (Nygren 1954; Richards 1986; Asker and Jerling 1992; Ramulu *et al.* 1999; Whitton *et al.* 2008).

Table 1.1. Occurrence of apomixis in common plant families (Nygren 1967).

#	Family	Sporophytic apomixis	Gametophytic apomixis		Total
			Aposporous	Diplosporous	
1	Asteraceae	-	18	51	69
2	Poaceae	-	68	27	95
4	Rosaceae	-	65	3	68
4	Liliaceae	6	-	1	7
5	Rutaceae	5	2	-	7
6	Urticaceae	-	2	7	9
	28 other families	33	5	15	53
	Total	44	104	104	308

From the table, it seems that members of Rosaceae and Poaceae mostly have aposporous apomixis, while in the Asteraceae diplosporous apomixis is more common, and sporophytic apomixis in these three families is not reported (Asker and Jerling 1992; Naumova 1993). Despite being the largest plant family, the Orchidaceae appears rarely to reproduce apomictically (Grimanelli *et al.* 2001b).

The presence of apomixis among different clades and higher taxa makes it obvious that apomixis has evolved independently and repeatedly (Majeský *et al.* 2012), and the occurrence of apomixis in different forms and in unrelated families suggests that apomixis originated multiple times during flowering plant evolution (Koltunow and Grossniklaus 2003).

Whilst apomixis is much rarer in animals than in plants, some animals, such aphids, show apomictic reproduction (Suomalainen *et al.* 1987; Moran 1992).

1.1.6. Variation in apomictic plants

Apomixis has been considered as an evolutionary dead end (Chapman *et al.* 2003; Hörandl and Hojsgaard 2012). Theoretically, the offspring produced by apomictic plants are presumed to be identical to their maternal parent. Nevertheless, there are

some indications that there are low level of genetic variation in apomictic plants, such as variation in morphology and aneuploidy (Sørensen and Gudjónsson 1946) and variation in size and morphology of pollen grains (Chiguryaeva 1976). Recently more genetic variations have been discovered using molecular technique. There has been a number of studies focussed on genetic variation within apomictic plant populations (e.g. *Taraxacum*), using isozyme or allozyme markers (Menken *et al.* 1995) and DNA markers (King and Schaal 1990; Van der Hulst *et al.* 2000).

The source of these variations present in apomictic species may be due to three factors. (1) somatic mutation, especially when the egg-like cells are produced from somatic cells, leads to increase in the possibility of incorporation of somatic mutations into the gametes, such as found in rDNA intergenic spacers and *Adh*1 gene within asexual lineages in *Taraxacum* (King and Schaal 1990), and somatic microsatellite mutations within apomictic lineages of *Ranunculus auricomus s.l.* (Paun and Hörandl 2006). Whilst, slightly deleterious mutations purged in the sexual gene pool, deleterious mutations cleansed in the haploid genome of sexual egg cells.

(2) The possible origin of variation could come from hybridization. Indeed, in comparison with the variation caused by mutation it causes the greater variations between individual apomictic plants. Hermaphrodite apomict plants can act as pollen donors, and when sexual and apomict plants occur in the same population, hybridization can occur between them. The haploid egg cells from sexual diploid plants are fertilized by diploid pollen from triploid or tetraploid facultative apomicts, which results in the generation of triploid neo-apomictic lineages (Richards 1973; Tas and Van Dijk 1999). In such neo-apomictic clones part of the genetic variability will be transferred from the sexual mother and this will increase total variability of the apomictic genetic pool (Van Dijk 2003; Verduijn *et al.* 2004). This kind of hybridization process has been demonstrated in nature when rare allozymes were found to be shared between mixed sexual and apomictic populations (Menken *et al.* 1995). King (1993) characterized genetic variation in rDNA and chloroplast DNA, and came up with the result that multiple hybridization events were a more important source of genetic variation than mutation in the asexual polyploids.

However, when sexual plants are absent, it is more difficult to explain the clonal diversity in those regions. One of the possible explanations could be the migration of new triploid apomictic lineages, generated by hybridization between diploid and apomictic populations in the sympatric regions, to purely apomictic regions (King 1993). In these regions, new apomictic clones can arise after hybridization with facultative apomicts.

(3) Auto segregation is another possible source of genetic variation within apomictic plants in the absence of sexual plants (Gustafsson 1935). Autosegregation comprise two processes: chromosome gain or loss (Sørensen and Gudjonsson 1946) and subsexual reproduction (Nogler 1984). Subsexual reproduction involves crossing-over between a heterozygous locus and the centromere without reduction, which leads to homozygosity of genes distal to the crossovers. For example, in the *Taraxacum*-type of apomixis, meiotic prophase I is initiate but fails to reach the first meiosis reductional division so there is the possibility of some recombination activity. This suggests that subsexual reproduction is a potential source of variation in *Taraxacum* (Darlington and Mather 1950). In other kinds of apomixis, such as mitotic diplospory or apospory, meiotic prophase I is absent and so subsexual reproduction cannot occur in these cases. The occurrence of subsexual recombination may be an advantage for apomictic lineages because it limits accumulation of deleterious mutations (Baarlen *et al.* 2000).

For example, in apomictic *Taraxacum* there is a possibility of generating heritable variation by means of increased transposon activity by somatic recombination (Richards 1989; King and Schaal 1990). Epigenetically variations lead to diversitry among apomictic species including inheritance of the level of methylation induced by different stress factors, and *de novo* methylation variation due to the hybridization process (Verhoeven *et al.* 2010).

Whatever its source, studies have demonstrated the existence of variation within apomictic microspecies. These variations may help prevent apomictic lineages from becoming extinct due to their restricted potential to adapt.

1.1.7. Causes of apomixis

There are several different hypotheses that explain the origin and causes of apomixis (Koltunow and Grossniklaus 2003). The earliest is Ernstis hypotheses, which proposed that apomixis might be caused by hybridization and polyploidization (Ernst 1918).

Another hypothesis is that apomixis might be caused by asynchronous expression of genes or genome duplication, which is termed the hybridization-derived floral asynchrony (HFA) theory, i.e. the events of polyploidization (Carman 1997, 2007). However, apomixis cannot be induced just by polyploidization alone, because not all polyploids are apomicts. Heslop-Harrison (1959) considered that environmental conditions were the caused transition from asexual to sexual reproduction within lower plants. Similar processes have been observed in animals that are apomictic during favourable conditions and sexual during stressful periods, such as aphids and water fleas (Suomalainen *et al.* 1987).

There are different traits belonging to plant life cycle provide insight into the ecological role and nature of supporting occurrence of the apomixis in plants. Asker and Jerling (1992) reviewed that apomixis occurs in plants that display physiological self-incompatibility (autogamy), dioecy or heterostyly (Philipson 1978; Gadella 1991; O'Connell and Eckert 1999; Heenan *et al.* 2002; Bicknell *et al.* 2003).

1.1.8. Genetic control and inheritance of apomixis

Much interest has recently been shown in genes involved in the control of apomixis due to using apomixis as a tool in agriculture and plant breeding (Ozias-Akins 2006).

Nijs and Van Dijk (1993) reported that apomixis is under the control of a single dominant gene. Singh (2003) suggested that apomixis can be controlled by qualitative traits, recessive and dominant genes, and in some facultative species polygenes may be involved, for example, apomeiosis, parthenogenesis, and endosperm formation are controlled by a separate genetic loci (Grimanelli *et al.* 2001a; Grossniklaus *et al.* 2001). Thus, possibly, three specific loci may be affecting apomixis (Koltunow and Grossniklaus 2003). The *DIPLOSPOROUS* (*Dip*) gene controls the avoidance of meiotic

reduction and *PARTHENOGENESIS* (*Par*) controls parthenogenesis of the embryo, are two dominant loci in apomixis. The *Dip* function is sex determined, leads to generating unreduced egg cells without affecting pollen meiosis in the same plant. However, perhaps, the apomixis allele have ability to reverse from apomixis to sexuality (Sørensen and Gudjonsson 1946; Sørensen 1958), so that in triploid plants, suppression of the sexual reproductive pathway cannot caused by just a single dominant allele.

Environmental factors have effect on apomixis including temperatures and changes in light duration (Bashaw 1980; Hanna and Bashaw 1987; Hanna 1995; Asker and Jerling 1992; Ramachandran and Raghavan 1992; Nijs and Van Dijk 1993; Lutts *et al.* 1994).

Van Dijk *et al.* (2009) studied dominant apomixis genes by using co-dominant genetic markers and classified populations of apomictic *Taraxacum* in to three structures in three hierarchical levels: the individual plant, the clone and the apomixis gene. This suggests that in some apomicts the single apomixis locus might contain several genes linked with different function.

According to Bicknell and Koltunow (2004) crosses between apomictic and closely related sexual species can improve the inheritance of most of apomixis.

The inheritance of gametophytic apomixis can be associated with transferring of a single locus or a small number of loci (Bicknell and Koltunow 2004) or it can be controlled by 1-5 dominant genetic loci (Ozias-Akins and Van Dijk 2007).

In aposporous, apomixis is inherited and controlled by a single dominant locus that co-segregates with parthenogenesis such as in apomictic *Panicum* spp. (Savidan 1983), *Ranunculus* spp. (Nogler 1984), *Hieracium* spp. (Bicknell *et al.* 2000), *Pennisetum* (Sherwood *et al.* 1994) and *Brachiaria* (Valle *et al.* 1994). Whereas, according to Vijverberg *et al.* (2010), because there is a low recombination frequency around apomixis loci formation of unreduced embryo sac and fertilization-independent embryogenesis has not been genetically separated for example in aposporous *Pennisetum*.

In the Asteraceae, parthenogenesis and apomeiosis (diplospory) are distinct and inherited as separate genetic factors or unlinked loci (Noyes and Rieseberg 2000), which can be seen for example in *Taraxacum* (Tas and Van Dijk 1999; Van Dijk and Bakx-Schotman 2004), *Hieracium* (Catanach *et al.* 2006) and *Erigeron annuus* (Noyes and Rieseberg 2000). In *Hieracium praealtum* it was observed that meiosis and fertilization are avoided by the action of two dominant independent genetic loci: Loss of Apomeiosis (LOA) and Loss of Parthenogenesis (LOP) (Koltunow *et al.* 2011). LOA has sporophytic actions and is responsible for the differentiation of aposporous initial (AI) cells then avoidance of sexual pathway. LOP has gametophytic functions and is necessary for the development of autonomous embryo and endosperm formation (Koltunow *et al.* 2011). Absence of either LOA or LOP results in sexual reproduction.

1.1.9. Transposable elements and apomixis

Major proportion of the genomic DNA of living organism is comprised of transposable elements, and this kind of repetitive DNA has a great role in genome evolution.

When the transposable elements become active they can cause slightly deleterious mutations. In asexual organisms, accumulation of deleterious mutations could cause extinction (Ozias-Akins and Van Dijk 2007) because of the rare of meiosis and sexual reproduction, which prevent selection against deleterious alleles (Dolgin and Charlesworth 2006). In *Taraxacum* transposable elements substitution rates at non-synonymous sites were much lower than at synonymous sites (Docking *et al.* 2006). Examples of active transposons like Ty1-copia, Ty3-gypsy and LINE-like retroelements, which are possibly still functional in the apomictic *Taraxacum* which may leads to reduced fitness of clone mates consequently decline the number of clone mates then extinction of the apomictic clone (Van Dijk *et al.* 2009). In sexual populations transposable elements could spread to all individuals through sexual reproduction and outcrossing, however in asexual populations this spreading prevented, because of the absence of a high frequency of horizontal spread (Hickey 1982).

After introducing active transposable elements into sexual and asexual lines, indicate that spreading and ability of increase in frequency of genomic parasites in asexual populations lower compared with their sexual relatives. Because asexuality is a

derived state in higher eukaryotes, so there is the possibility that transposons transferred from sexual ancestors to asexual species because it reduce the need for initial spread.

The sequence of the Apomixis Specific Chromosome Region (ASCR) in *Pennisetum s*howed high numbers of transposable elements duplications and insertions (Conner *et al.* 2008; Calderini *et al.* 2006). ASGR in *Pennisetum* and the LOA locus in *Hieracium* located on a chromosome linked with substantial repetitive sequence and transposon-rich regions (Akiyama *et al.* 2004; Okada *et al.* 2011).

The hemizygous apomictic controlling locus (ACL) of *Paspalum*, incompare with syntenic region in rice (Calderini *et al.* 2006) which shows strong suppression of recombination, has undergone large-scale rearrangements due to transposable elements, suggesting that repetitive chromosomal structure may have a functional role in apomixis. A hypothesis revealed that repetitive sequences may act as a sink to sequester factors involved in the sexual reproductive pathway, thereby altering the expression of sexual reproductive processes, and possibly causing apomixis (Koltunow and Grossniklaus 2003).

The vast majority of sequence diversity and evolution of TEs in eukaryotes studied so far have been carried out in sexually reproducing organisms, and a few surveys and experimental studies have been performed to test the relationship between transposable element abundance and mode of reproduction.

There are not enough study on comparison of transposable element activity and expression in sexual and asexual species in plants, however such studies with animals has shown transposable element diversity in asexual taxa is much lower when compared with sexual taxa (Docking *et al.* 2006). Docking *et al.* (2006) investigated diversity of transposable elements (Ty1-copia, Ty3-gypsy and LINE) in four asexual plant species (*Taraxacum, Hieracium, Antennaria, Vittaria*), suggesting the possibility of recent evolving of relative asexual reproduction within these taxa, so that the loss of transposable elements cannot be detected through analysis of sequence diversity. Further, observation of retroelements in every group of animal phyla tested by broad PCR-based survey, except for the anciently asexual Bdelloid rotifers (Arkhipova and Meselson 2000). In addition, study of Ty3-elents in yeast and homing endonuclease genes (Zeyl *et al.* 1996; Goddard *et al.* 2001) have demonstrated that both types of

elements tend to spread in sexual populations but not in asexual populations. In bacteria, DNA elements have been spread through experimental populations, but the spread appears to be tied to specific beneficial mutations caused by element insertion (Cooper *et al.* 2001; Edwards *et al.* 2002; Docking *et al.* 2006).

1.2. Repetitive DNA

Eukaryotic genomes consist of large amounts of repetitive DNA ranging from sequence motifs of di-nucleotides to more than 10 kb length. Repetitive DNA sequences can be classified based on their organization (tandem arrays or dispersed) throughout the genome, their chromosomal location, and function. Today, with the advancement of next generation sequencing (NGS) several eukaryotic genomic DNA have been sequenced by low cost in short time, made study of repetitive DNA sequences in the genome easer which changed the concept of repetitive DNA sequences (Heslop-Harrison and Schmidt 2012; Lopez-Flores and Garrido-Ramos 2012). The major repetitive DNA classes (Tandem and dispersed repeat) are further sub divided in to different superfamilies of repetitive DNA. An abundant, ubiquitous and diversity of repetitive DNA in the genome caused difficulty in genomic assembly and annotation. Figure 1.4 shows the diagram of plant nuclear genome compositions, and classifications.

1.2.1. Tandem repeated DNA

Tandem repeat arrays of the repeat unit occur where individual copies of the DNA fragment are repeated adjacent to each other (Kubis *et al.* 1998). Tandemly repeated DNAs have been characterized and localized on chromosomes. Schmidt and Heslop-Harrison (1998) suggested that tandem repetitive DNA sequences in plant genome are pericentromeric, sub-telomeric, telomeric, intercalary, and centromeric regions on most or all chromosomes, as shown in Figure (1.5). Tandemly repeated DNA include ribosomal RNA (rRNA), protein-coding gene families, microsatellite and satellite DNAs, and centromeric DNA (Figure 1.4), (Lopez-Flores and Garrido-Ramos 2012).

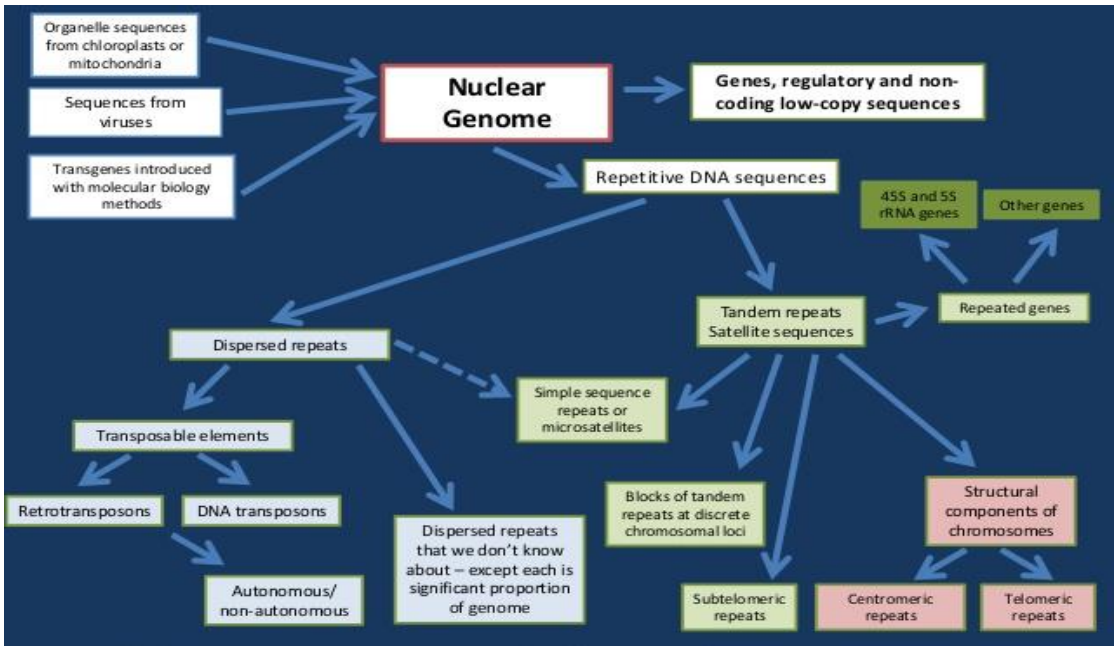


Figure 1.4. DNA sequence components of the nuclear genome after Heslop-Harrison and Schmidt 2012.

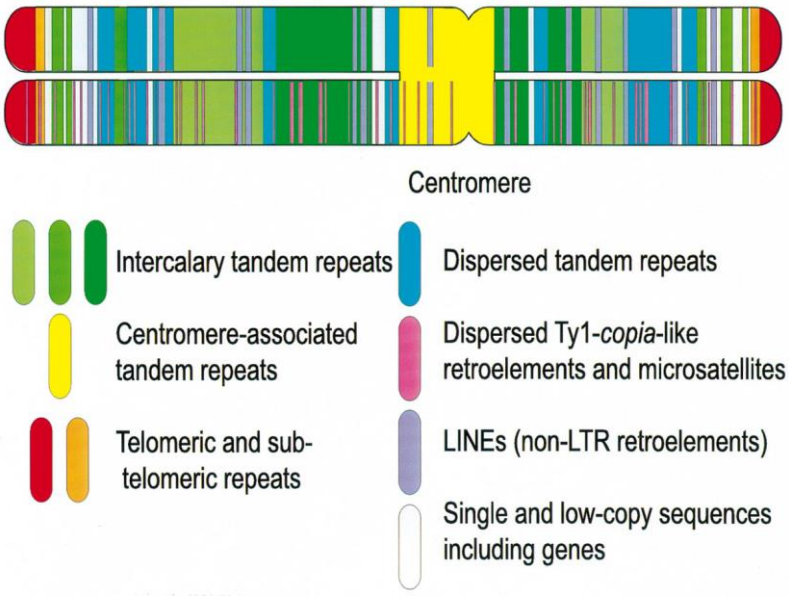


Figure 1.5. A model of plant chromosome, which shows characteristic genomic distribution of different classes of repetitive DNA (Schmidt and Heslop-Harrison 1998).

1.2.1.1. Ribosomal RNA genes

Ribosomal RNA genes (rDNA) are a multigene family, consist of tandem array repeat units, containing 3 of the 4 genes encoding nuclear rRNA, located in the nucleolar organizer region (NOR) on one or more chromosomes per haploid genome. Each repeat unit consist of 26S large subunit, 18S small subunit, 5.8S gene, two external transcribed spacers (ETS), two internal transcribed spacers (ITS1 and ITS2), and a non-transcribed spacer (NTS). The ETS along with NTS form inter genic spacers (IGS). The 26S, 5.8S, and 18S rRNAs are encoded by a 45S transcription unit. The rDNA differ in copy number and varies between eukaryotes, from 39-19300 in animals and from 150-26000 in plants (Prokopowich *et al.* 2003). Figure 1.6 represent the arrangement of the ribosomal genes that encoding for 45S rRNA along with the component of spacer region, and the 45S rDNA tandemly repeated sequence in the genome.

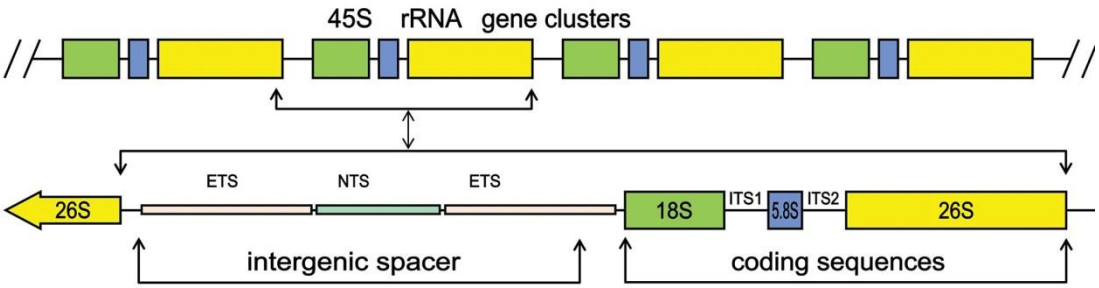


Figure 1.6. Arrangement of ribosomal genes coding for 45S.

The rDNA components are evolve at different rates. The 18S rDNA is the slowest-evolving genes in contrast to the IGS which evolve rapidly with the NTS and even faster than ITSs (Long and Dawid 1980).

The 18S and 28S rRNA genes give the interpretation of phylogenetic history across a broad taxonomic range, while the ITSs is useful in determining the relationships between closely related species, intraspecific relationships and population studies. Studies showed that there is no difference between repeats of ITSs nucleotide sequence of the same species but they have wide range of differences between species. Whereas, nucleotide sequences of the rRNA coding regions have great similarity between closely related species even among distantly related species (Dover 2002).

Also, 5S rRNA gene encodes rRNA, form a minor rDNA family, consists of multiple tandem repeated DNA of the gene separated by NTS. In most eukaryotes, the 5S rRNA located in another region of the nuclear genome. In protozoa, fungi, and algae the 5S rRNA genes are located within the IGS between the 28S and the 18S genes. Ribosomal RNA characteristics are important in evolution, and taxonomy. *In situ* hybridization has made the rDNA loci valuable marker and easy screening for chromosomes evolution examination (Heslop-Harrison and Schwarzacher 1996).

1.2.1.2 Microsatellite, Minisatellite and satellite DNA

The microsatellites are tandem repeats with motifs of 2-6 bp found in arrays up to 1kb. They are also known as simple sequence repeats (SSRs) or short tandem repeats (STRs). SSRs are ubiquitous in plants which can be found in both protein coding and non-coding regions. Di-nucleotides are the major type of SSRs for many species. The most common dinucleotide repeat in plants is $(GA)_n/(CT)_n$ and $(AT)_n/(TA)_n$ repeats (Tóth *et al.* 2000). They evolve rapidly, so they are valuable as molecular markers and for fingerprinting (Kubis *et al.* 1998). The SSRs repeats localized at the ends of chromosomes or centromeres (Figure 1.5).

Minisatellites are tandem repeated DNA with a unit size more than 9 bp up to 40 bp. Micro- and minisatellite is different on their distribution and potential function in eukaryotic genomes. The *in situ* hybridization study showed that the minisatellites DNA in the plant are located in the pericentromeric region such as in *A. thaliana* (Vergnaud and Denoeud 2000).

Satellite DNAs (satDNAs) are highly repetitive DNA sequences that constitute the large part of eukaryotic genomes. SatDNAs consist of a series of identical repetitive monomers. They are not encoding any protein. Repeat units are tandemly repeated sequences sized over 200 bp in length and organized in long repetitive arrays in the genomes. They differ from micro- and minisatellite by size, array length, genome location, and the dominant mechanisms to increase their number. Satellite DNA families were found to be localized in regions of the centromeric, pericentromeric, and subtelomeric. Satellites have been characterized for many species in the plant (Traut 1991) by *in situ* hybridization.

1.2.1.3. Telomeric DNA

Telomeric DNA consisting of conserved seven bp repeats (CCCTAAA/ TTTAGGG), is added to the physical ends of the linear chromosome of most plants and animals by an enzyme telomerase activity. This unusual enzyme is a reverse transcriptase, incorporating an RNA template (Schwarzacher and Heslop-Harrison 1991), and the average length is typically a few thousand base paires. The telomeric sequences preserve a linear replication unit, protect chromosome ends, and control the 'end replication problem'. Telomeric arrays can be visualized by *in situ* hybridization, also intercalary arrays of the sequence can be observed (Fuchs *et al.* 1995), (Figure 1.5).

Telomeres and rDNA are an ancient element of genomic DNA because they are found in all animals and plants, and might be considered as early derivatives of the 'RNA world' from which DNA-based organisms evolved.

1.2.2. Dispersed repeats - Transposable elements (TEs)

Transposable elements (TEs) are also known as mobile genetic elements, jumping genes, transposons or retrotransposons, they are DNA fragment that replicate and move from one chromosomal position to another within the same genome, resulting in mutation and alteration of the cell's genome size (Jurka *et al.* 1992; Flavell *et al.* 1994). The Nobel Prize winner Barbara McClintock first discovered TEs in the early 1950s, she discovered the AC element that was the first transposable element described in maize (McClintock 1952), and the simplest transposons have been discovered in bacteria called insertion sequences (IS).

TEs are present in copy numbers ranging from a few up to millions of copies per genome, and are a major component of all eukaryotic genomes as well as comprising the bulk of higher plant genomes (Schmidt 1999; Heslop-Harrison and Schmidt 2012). The main TE groups discovered in most living organisms make them to be considered as an ancient genomic component (Kidwell 2002). TEs are variable and abundance between different species (Hua-Van *et al*. 2005). SanMiguel and Bennetzen (1998) and Morgante (2005) suggested that TEs in plants constitute >80-85% of the total genomic DNA, including maize (Schnable *et al.* 2009; Heslop-Harrison and Schmidt 2012), wheat (Tenaillon *et al.* 2011), 30% in rice

(http://rice.plantbiology.msu.edu/), 15% in *Arabidopsis*, and >90% in Liliaceae (Vitte and Panaud 2005). However, TEs are less abundant in fungi (3-20%) and metazoans (3-45%) in contrast to plants (Daboussi and Capy 2003; Hua-Van *et al.* 2005), and 3% of yeast genome comprises TEs (Kidwell 2005). Some parasitic apicomplexa and 20% of prokaryotic genomes are excluding of TEs (Hua-Van *et al.* 2011; Bringaud *et al.* 2006).

1.2.3 Transposable elements classifications

Finnegan (1989) classified TEs into two major classes according to their mode of transposition; class I transpose by RNA intermediate, and class II their transposition does not need RNA intermediate. Hansen and Heslop-Harrison (2004) classified the retroelements to four classes Non-viral retroelements, viral retroelements, the *envelope* gene, and replicative cycle of retroelements. Figure (1.7) from Hansen and Heslop-Harrison (2004) shows manually drawn of the retroelements including a LINE, copia and gypsy elements, pararetroviruses and retroviruses, with the scale in base pairs of possible repetitive sequence length in the genome. Aalso, figure (1.8) shows another different classification of repetitive DNA component in nuclear genomic DNA, with their repetitive components arrangement, and their popularity in various taxa from Wicker *et al.* (2007), through making a hierarchical system for TEs classification, which consists of six levels (class, subclass, order, superfamily, family and subfamily). This system classified TEs to two classes (Class I and II) according to their transposition mechanism (Figure 1.8).

Kapitonov and Jurka (2008) considering enzymology, structural similarities and sequence relationships, thus, they classified all eukaryotic TEs into two major types, retrotransposons and DNA transposons which composed of five major classes: long terminal repeat retrotransposons, non-LTR retrotransposons, cut-and-paste DNA transposons, rolling-circle DNA transposons (Helitrons) and self-synthesizing DNA transposons (Polintons). Lisch (2013) classified TEs in plant genomes into three classes (retrotransposons, DNA transposons, and Helitrons).

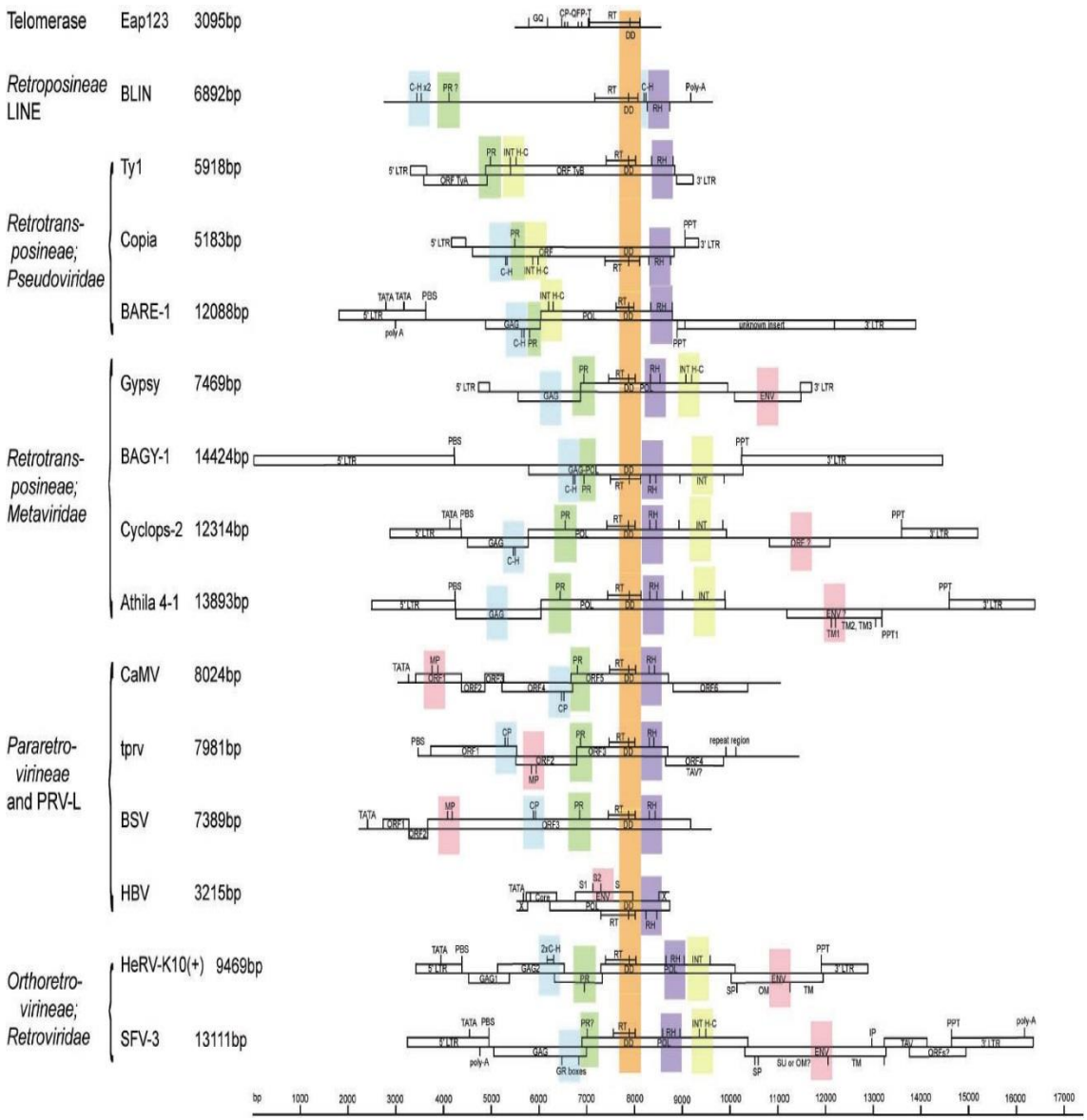


Figure 1.7 Drawings of retroelements including a LINE, copia and gypsy elements, pararetroviruses and retroviruses. A scale in base pairs is shown at the bottom. The elements are manually aligned over the amino acids DD from RT (Hansen and Heslop-Harrison 2004). LTR Long terminal repeat; PBS Primer binding site; gag group specific antigen; CP Coat protein; C-H Cysteinehistidine repeat motif; pol polyprotein; PR Aspartic protease; RT Reverse transcriptase; RH Ribonuclease H/ RNase H; INT Integrase; Env Envelope gene; MP Movement protein; TAV Transactivator; PPT Polypurine tract; Colour code: orange, DD site of RT; purple, RNase domain (RH); yellow, integrase domain (INT); blue, cysteine-histidine motif (C-H); green protease domain 25 (PR); pink, envelope/movement protein domain (ENV/MP).

Classification		Structure	TSD	Occurrence
order	superfamily			
Class I. Retrotransposons				
LTR	Copia	GAG AP INT RT RH	4–6	P, F, M, O
	Gypsy	GAG AP INT RT INT	4–6	P, F, M, O
	Bel-Pao	GAG AP INT RT INT	4–6	M
	Retrovirus	GAG AP INT RT INT ENV	4–6	M
	ERV	GAG AP INT RT INT ENV	4–6	M
DIRS	DIRS	GAG AP INT RT YR	0	P, F, M, O
	Ngaro	GAG AP INT RT YR	0	P, F
	VIPER	GAG AP INT RT YR	0	O
PLE	Penelope	RT EN	var.	P, F, M, O
LINE	R2	RT EN	var.	M
	RTE	APE RT	var.	M
	Jockey	ORF1 APE RT	var.	M
	L1	ORF1 APE RT	var.	P, F, M, O
	I	ORF1 APE RT RH	var.	P, F, M
SINE	tRNA		var.	P, F, M
	7SL		var.	P, F, M
	5S		var.	M, O
Class II. DNA transposons. Subclass 1				
TIR	Tc1-Mariner	Tase'	TA	P, F, M, O
	hAT	Tase'	8	P, F, M, O
	Mutator	Tase'	9–11	P, F, M, O
	Merlin	Tase'	8–9	M, O
	Transib	Tase'	5	P, F
	P	Tase	8	P, M
	PiggyBac	Tase	TTA	M, O
	PIF-Harbinder	Tase' ORF2	3	P, F, M, O
	CACTA	Tase ORF2	2–3	P, F, M
Crypton	Crypton	YR	0	F
Class II. DNA transposons. Subclass 2				
Helitron	Helitron	RPA Y2 HEL	0	P, F, M
Maverick	Maverick	C-INT ATP CYP POLB	6	F, M, O

Designations. long terminal repeat LTR; coding region; noncoding region; terminal inverted repeats (TIRs); hallmarks in the noncoding region; region containing one or more additional ORFs.
AP: asparagine protease; ENV: envelope protein; POL B: DNA polymerase B; Tase: transposase (': with a DDE motif); APE: apurinic endonuclease; GAG: capsid protein; RH: RNase H; ATP: packaging ATPase; HEL: helicase; RPA: replication protein A (only in plants); YR: tyrosine recombinase; C-INT: C integrase; CYP: cysteine protease; ORF: open reading frame of unknown function; RT: reverse transcriptase; Y2 YR with a YY motif; EN: endonuclease; P: plants; F: fungi; M: Metazoa; O: other taxa.

Figure 1.8 Classification system for mobile genetic elements and their popularity in various taxa from Wicker *et al.* (2007).

1.2.4. Retrotransposons (Class I elements)

Retrotransposons or retroposons (Slotkin and Martienssen 2007) transpose via a 'copy-and-paste' mechanism via an RNA intermediate. The mRNA is transcribed from the element by RNA polymerase II, next converted by reverse transcription into a complementary DNA (cDNA), then integrated via an integrase at a new location in the same genome (Wicker *et al.* 2007, Lopez-Flores and Garrido-Ramos 2012 and Lisch 2013), (Figure 1.9). As a result of their life cycle, they have a dispersed distribution along chromosomes (Figure 1.5), (Heslop-Harrison *et al.* 1997). At the end of each cycle, the elements undergo duplicative transposition, as their total number increases leading to genome size expansion (Slotkin and Martienssen 2007; Kumar and Bennetzen 1999; SanMiguel *et al.* 1996; Lopez-Flores and Garrido-Ramos 2012).

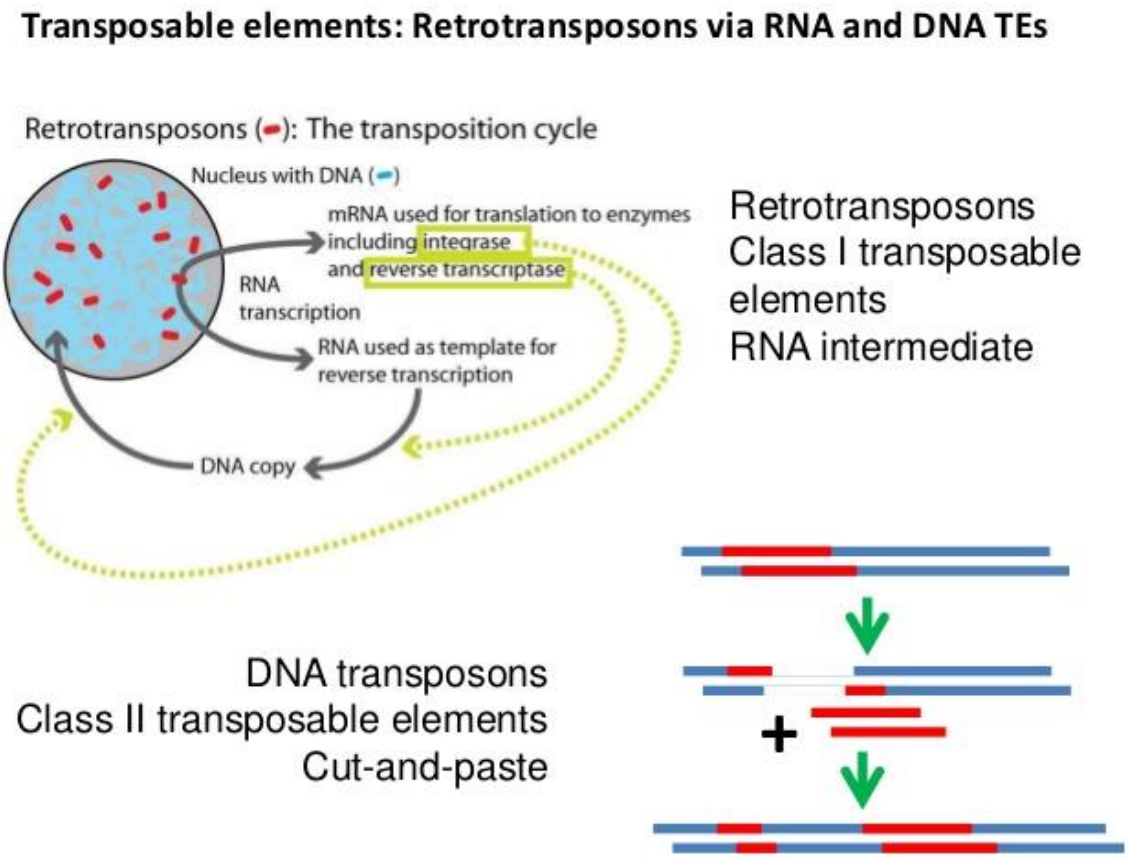


Figure 1.9 Retrotransposon and transposons life cycle (Heslop-Harrison 1998).

Retrotransposons are further divided based on the presence or absence of direct repeats at the ends of the element, known as long terminal repeats (LTRs): LTR-retrotransposons and the non-LTR retrotransposons (Figure 1.7), they are considered as the main retrotransposon order in plants, in contrast to animal genomes contain less LTR retrotransposons (Wicker *et al.* 2007; Lopez-Flores and Garrido-Ramos 2012).

1.2.4.1. LTR-retrotransposons

LTR-retrotransposons reach up to 25 kb in length (Neumann *et al.* 2003). In plants, LTR retrotransposons comprises a small percentage of the genomic component, it constitute <10% in rice, 5% in Arabidopsis, 54.5% in sorghum and 50-80% of the maize genome (Sanmiguel and Bennetzen 1998; Kapitonov and Jurka 1999; Mao *et al.* 2000; Meyers *et al.* 2001; Paterson *et al.* 2009).

Figure 1.10 shows structural component of complete length of LTR retrotransposons. The LTRs bound an internal domain that encodes the proteins required for retrotransposition (Schulman and Kalendar 2005), these proteins are present as two main Open Reading Frames (ORFs). First, *gag* polyprotein encodes the structural protein for a virus-like element, and proteins required for genome integration. *Pol* gene is a longer ORF and most conserved than the *gag*. Second, p*ol* has a polyprotein and it is auto-processed by aspartic proteinase (AP) domain, the most conserved domain present in all the retrotransposons reverse transcriptase (RT), for transposition mechanism RNase H (RH), and catalyzes the transposition and integration integrase (INT), (Suoniemi *et al.* 1998).

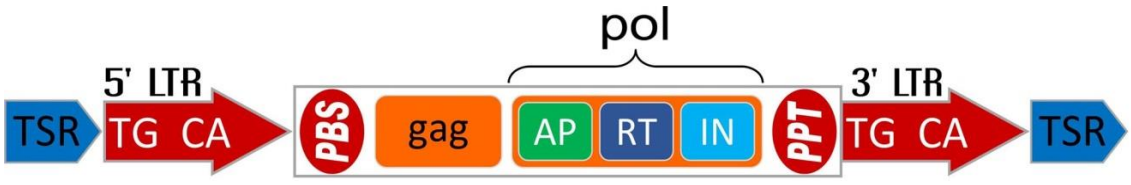


Figure 1.10 Basic structure of a full-length LTR retrotransposons, from Sabot *et al.* (2006). TSR: Target Site Repeat; PBS: Protein Binding Sequence; PPT: Polypurine Tract; ORFs: Open Reading Frames; *gag*: group-specific antigen; *pol*: polymerase; AP: Aspartic Proteinase; RT: Reverse Transcriptase; RH: RNase H; INT: Integrase; TSD: Target Site Duplication.

LTR retrotransposons are further sub-classified into 5 well-known superfamilies (Kumar and Bennetzen 1999; Wicker *et al.* 2007) according to the order of genes within the internal domain, *Ty3*-gypsy group (Metaviridae); *Ty1*-copia group (Pseudoviridae), *Retroviruses* (vertebrate retroviruses); Endogenous retroviruses (ERVs); and Bel-Pao group. Although, Lopez-Flores and Garrido-Ramos (2012) divided LTR retrotransposons into 3 major superfamilies (copia, gypsy, and Bel-Pao) based on their degree of sequence similarity and the order of encoded gene products.

1.2.4.1.1. Ty3-gypsy (Metaviridae)

The Ty3-gypsy is one of the LTR retrotransposons major superfamilies (Wicker *et al.* 2007), classify under the families Metaviridae, have a wide distribution among fungi, plants, and animals. Ty3-gypsy generates 4-6 bp TSDs and flanked by LTR. The *gag-pol* genes encode for protein domains PBS and PPT towards downstream and upstream of 5' and 3' LTR respectively. The name Ty3-gypsy derived from the *Ty3* retrotransposons in the genome of *Saccharomyces cerevisiae* (Hansen *et al.* 1988) and *Drosophila* melanogaster for a gypsy (Marlor *et al.* 1986). The INT domain in this element is located downstream of RT and RH, as found in retroviruses (Figure 1.7, 8). Some of the Ty3-gypsy elements show an ORF3 thus they have similarity to retroviruses.

1.2.4.1.2. Ty1-copia (Pseudoviridae)

The *Ty1*-copia is another abundant LTR retrotransposons superfamilies, classified under the families Pseudoviridae. Found in most of the living organisms genome including plants (Manninen and Schulmann 1993; White *et al.* 1994; Bennetzen 1996; Wicker *et al.* 2007). The Ty1-copia group, named after the *Ty1* retrotransposons in the genome of *Saccharomyces cerevisiae* (Clare and Farabaugh 1985) and *Drosophila melanogaster* copia (Mount and Rubin 1985; Boeke and Corces 1989; Grandbastien *et al.* 1989). The INT domain is located upstream of the RT (Figure 1.7, 8). They are flanked by LTRs and displayed the PBS and PPT towards downstream and upstream of 5′ LTR and 3′ LTR respectively. The *Ty1*-copia elements show a lower sequence divergence in plants in contrast with fungi or insects, as well as a large number of subfamilies of divergent elements (Flavell *et al.* 1992).

Structurally, there are significant differences between copia and gypsy in the order of the three protein domains (INT, RT, and RH) present in a *pol* gene. In gypsy, the INT is located downstream to RT and RH, whereas in copia it is located upstream to RT and RH domains (Figure 1.7, 8).

1.2.4.2. Non LTR- retrotransposons

Non-LTR retrotransposons (retroposons) are terminated by a very short LTR and they are transcribed from an internal promoter (Slotkin and Martienssen 2007). They are sub-divided to long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) according to their size and internal region encoding the domains (Figure 1.7, 8). Both elements are found in plants (Kubis *et al.* 1998) whereas SINEs are more abundant in animals (Schmidt 1999; Jurka *et al.* 2007). They are considered as ancient genome components and ancestors of LTR retroelements.

1.2.5. DNA transposons (Class II elements)

DNA transposons are cut-and-paste transposons. In contrast to retrotransposons, their transposition does not involve RNA intermediate, and they have the ability to transpose by moving the genomic DNA copies from one chromosomal location to another. Their transpositions occur by the protein encoded transposition known as "transposase" recognize as the Terminal Inverted Repeats (TIRs). TIRs flank the retrotransposons. Transposition starts by excising the element from double stranded DNA of the donor position, then integrate it into a new location in the genome (Figure 1.9). Figure 1.11 shows the illustration of the transposition mechanisms of retrotransposons and DNA-transposons and rolling circle transposition of helitrons.

DNA transposons can increase in copy number in the host genome. As shown in figure (1.12) the mechanism that DNA transposons use to increase their number in the host genome and the only possibility of increasing DNA-transposons number occur when they transpose in the time of chromosome replication. In S phase of the cell cycle, from a position that has been replicated earlier to another location prior to the replication fork pass (Greenblatt 1962; Skipper *et al.* 2013). Alternatively, the gap that left behind at the donor position either can be repaired without element replacement

in cut-and-paste transposition, or filled with creation of an extra copy at the donor site (Slotkin and Martienssen 2007; Capy *et al.* 1998; Nassif *et al.* 1994). This mechanism can also results in gene duplication, which plays an important role in evolution (Figure 1.12).

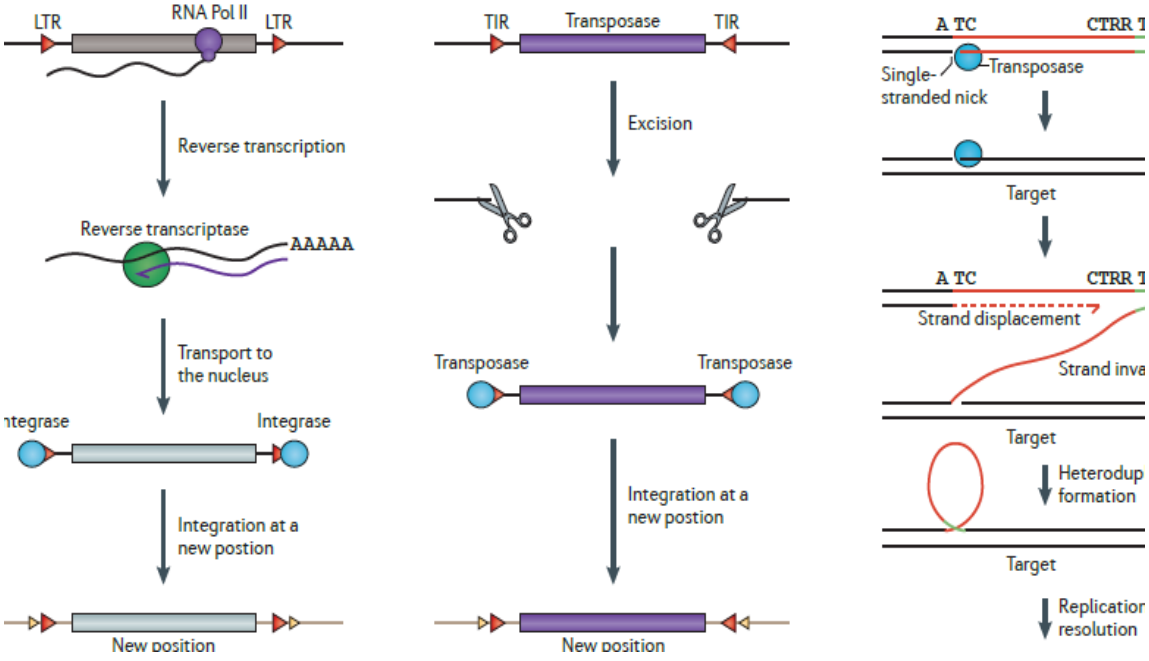


Figure 1.11. Illustrating the transposition mechanism of Class I, Class II elements and *Helitrons* (Lisch 2013).

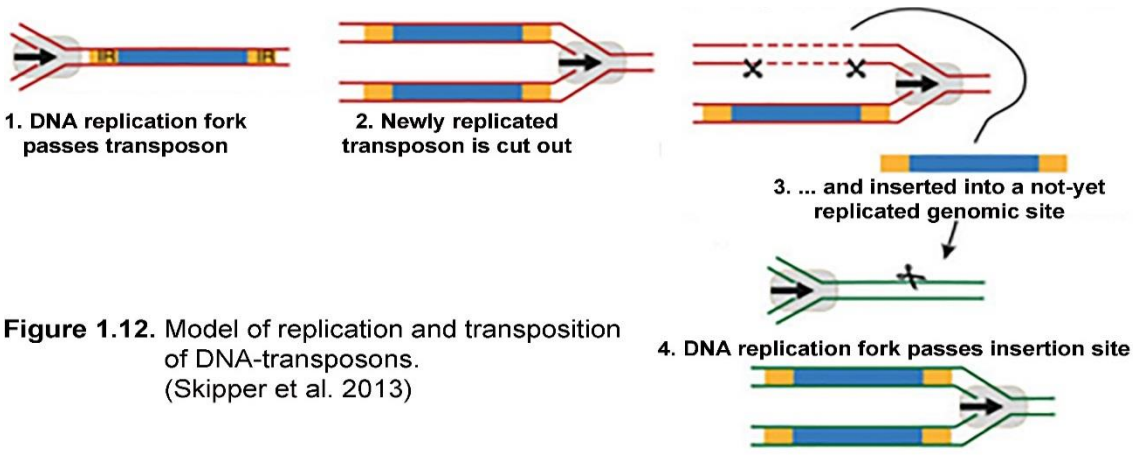


Figure 1.12. Model of replication and transposition of DNA-transposons. (Skipper et al. 2013)

DNA transposons can be sub-divided into 3 subclasses according to the number of DNA strands that are cut during transposition (Feschotte and Pritham 2007; Kapitonov and Jurka 2008; Bao *et al.* 2009), Subclass 1- cut-and-paste DNA transposons; Subclass 2- rolling-circle DNA transposons (*Helitrons*); Subclass 3- self-synthesizing DNA transposons (*Polintons*). However Wicker *et al.* (2007) classified DNA-transposons as two subclasses, according to the same criteria; subclass 1 comprises 'cut-and-paste' is composed of Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger, CACTA and Crypton, subclass 2 comprises DNA TEs entails replication without double-stranded cleavage, including Helitron and Meverick elements (Figure 1.8). These DNA transposons mentioned above just few of them common in plants such as Tc1-Mariner, hAT, CACTA, PIF-Harbinger and Mutator superfamilies (Wicker *et al.* 2007).

1.2.5.1. Subclass-1 DNA transposons

They are flanked by an inverted orientation of TIRs repeats. One ORF present to encode a transposase that recognizes the TIRs and cuts both strands at each ends.

This subclass is divided into two orders, TIR, characterized by the presence of different length of TIRs, and *Crypton*, contains a tyrosine (TY) recombinase and RT domain (Wicker *et al.* 2007). They are represented by 17 superfamilies, classified depending on the transposase which is superfamily-specific (Bao *et al.* 2009), although in (Wicker *et al.* 2007) they classified to 12 superfamilies based on TIR sequences and TSD size (Figure 1.8).

1.2.5.2. Subclass-2 DNA transposons

In contrast to subclass 1, they are DNA TEs that undergo a transposition process without a double-stranded cut. The most important example of this class is helitrons. They are DNA transposons that duplicate by a rolling-circle mechanism. Autonomous *Helitrons* contain a DNA helicase protein, as well as a replicate protein similar to replicon protein A (RPA) (Slotkin and Martienssen 2007). The transposons replicate itself to a new target site via cleavage of one strand on each terminal site by a rolling-circle mechanism (Kapitonov and Jurka 2001) involving replicative transposition (Figure 1.11). *Helitron* does not introduce TSDs and lack of TIRs. It has hairpin structures at the

end. *Helitron* ending with TC or CTRR motif, and have been shown to have a major role in genome restructuring through their capture and duplication of gene fragments (Heslop-Harrison and Schmidt 2012).

1.2.6. Autonomous and non-autonomous TEs

Transposition can be classified as either "autonomous" or "non-autonomous" in both Classes I and Class II TEs, according to the presence or absent of the genes that encode the enzymatic mechanism needed for their transposition. Autonomous TEs can move by themselves while non-autonomous TEs depend on the other autonomous TE to transpose because they are lake the proteins require for transposition (Sabot and Schulman 2006).

Activator element (Ac) is an example of an autonomous TE, and dissociation element (Ds) is an example of non-autonomous TE. Without Ac, Ds is not able to transpose. Non-autonomous elements are often mutated identical of autonomous family members, but sometimes have only limited sequence similarity to their autonomous counterparts. Non-autonomous DNA transposons often consist of a pair of TIRs surrounding non-transposon DNA (Slotkin and Martienssen 2007).

Figure (1.13) shows the mechanism of the impact of autonomous and non-autonomous transposable elements on corn kernel colour by TEs activator (Ac) in Maize, transposition of Ds, and chromosome breakage controlled by Ac control. C gene is responsible of expression of kernel colour. When Ds is insert in the C gene, it creates colourless cells, and when Ds remove from the C gene or transpose suppression effects release aleurone-colour gene of the Ds changed into the active form.

The large retrotransposon derivatives (LARDs), terminal repeat retrotransposons in miniature (TRIMs), miniature inverted-repeat TEs (MITEs) and SINEs are groups of TEs that are clearly non-autonomous. Each of LARDs (Kalendar *et al.* 2004) and TRIMs (Witte *et al.* 2001) describe large (more than 4 kb) and small (less than 4 kb) non-autonomous LTR retrotransposons derivatives, respectively.

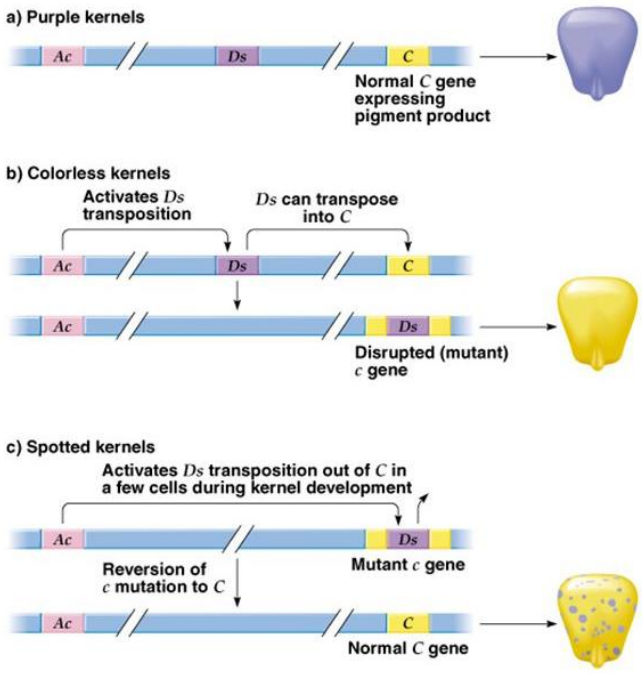


Figure 1.13. Transposon effects on corn kernel colour. (http://www.slideshare.net).

1.2.7. Retrotransposons impact on organism's genome

Transposons have been known as "junk" DNA because they have no obvious benefit to their host, and "selfish" DNA because they make many copies of themselves.

Transposition is linked with replication, recombination, and repair. The process of moving from one place to another involves a type of recombination. Mutations could be happen due to transposons insertions. Transposons are generating new copy so they replicate themselves.

Transposable elements might act as mutagens, because they cause mutation in several ways, including the insertion of TE near or into the gene which cause gene deactivations, then leads to disrupt expression of that gene. This kind of mutation caused by TE insertion is not different from the mutation that knocks out gene function, such as Mendel's wrinkled peas, white wine grape varietals, and several strains of seedless apples (Bhattacharyya *et al.* 1990; Kobayashi *et al.* 2004; Cadle-Davidson and Owens 2008; Shimazaki *et al.* 2011; Yao *et al.* 2001). Moreover, mutations can happen in the location that filling the gap did not successful, this gap left behind after cut and paste transformation.

Transposable elements contribute directly or indirectly to genome size variation. The accumulation of TEs in the genome leads to lacking association between the size of genome and number of functional genes. DNA content (C values) can vary greatly between different species. However, even between related species, there can be large differences in TE content, for example, genome size differences of the *Takifugu* and *Tetraodon* genomes (Jaillon 2004), *Oryza sativa* and its wild relative *Oryza australiensis* (Piegu *et al.* 2006), *Arabidopsis thaliana* and *Arabidopsis lyrata* (Hu *et al.* 2011).

Gene movement is a common feature of plant genomes which facilitate by TE activity. Then, it mediates changing chromosomal architecture in the large scale. As well as gene movements may alter genes regulations caused by movement of these genes by TE mediations into new chromosomal contexts (Woodhouse *et al.* 2010; Yang *et al.* 2008; Bhutkar *et al.* 2007). The neighboured genes also can affect by TEs through y altering splicing and polyadenylation patterns or by act as enhancers or promoters (Slotkin and Martienssen 2007).

TEs can cause deletions or inversions of DNA, and transposition can move DNA sequences to new locations that are not part of a TE. It happening when transposition results in two copies of the same sequence in the same orientation, recombination can delete the DNA between them. If the two copies are in the opposite orientations, recombination will invert the DNA between them, and it can be moved together with them when they move, so additional DNA sequences can be mobilized as a part of transposition mechanism.

1.2.8. Repetitive DNA and TEs identification

The biological impact of repetitive DNA makes them important to study them such as genomic structure, gene regulation, and genomic evolution (Makałowski *et al.* 2012).

So far, molecular analysis and cytological approaches was the useful method for characterization and isolation of different repetitive DNA sequence elements from different plant genomes to build up a picture of the DNA of the whole genome and find all the repetitive sequences that are present. These methods including cloning of

restriction satellites whole genomic DNA digestive on gels, making a clone library and probe with genomic DNA to find abundant clones, using degenerated primer to amplify RT domain of different repetitive DNA motifs, and microdissecting regions of chromosomes with repeats/heterochromatin and clones. Nouroz (2012) has been used several methods to identify transposable elements such as studying their mobility and insertions/deletions in comparisons of homologous or homoeologous chromosome sequences, characteristic sequence properties such as repeats and short duplications, and homology to known elements. Also, distinguishing of repetitive DNA motifs by their organization in the genome and their chromosomal localization is an important method.

Likewise, nowadays, computational approaches and tools are advanced technique and methods to annotate genomic sequences and identify all repetitive DNA sequences in the genomes. Especially it became useful method since the advent of next generation sequencing (NGS), which leaded to analysis of highly repetitive sequences in the genomes of several angiosperm species such as banana, pea, soybean, barley, tobacco have been finished.

Consequently, side by side to development of NGS, different program categories have been used according to their methodology to identify repetitive DNA and transposable elements, and these programs are differing according to the repeat type that they can identify (Lerat 2010). In two recent reviews, Bergman and Quesneville (2007) and Saha *et al.* (2008) technical and algorithmic aspects of the majority of these programs have been described in detail, however, Lerat (2010) focused on describing the programs and practical way to use each program (Lerat 2010, their Table 1).

Nowadays, with the improvement of many advanced computer software, and developing facilities of speed and storage capacity of the computer beside of the facilitate of next-generation sequencing, it is easy to search for organism transposable elements and come up with the result that transposable elements are an important components of all eukaryotic genomes and how they play a major affects in their evolution (Wicker *et al.* 2007).

1.3. Aims and objectives of the study

The overall aims of the study are (1) to define the nature, abundance and large-scale genome organisation of repetitive sequences in *Taraxacum*; and (2) to find the diversity, evolutionary mechanisms and consequences of repetitive DNA sequences for the genome; and the objective of each chapter is:

Chapter 3. Focus on using retrotransposon-based and other DNA markers to investigate diversity in the tribe *Cichorieae* (Asteraceae), by using PCR primers specific for conserved domains of RT genes of copia-, gypsy-like and LINE retroelements, and confirm the diversity of *Taraxacum* microspecies assayed by Majeský *et al.* (2012) using IRAP and other markers. Compare *Taraxacum* and *Hieracium* to confirm that the *Taraxacum* species are most appropriate for whole-genome analysis.

Chapter 4. Sequ ence whole chloroplast genomes (plastomes) of three morphologically well-defined apomictic microspecies from the *Taraxacum officinale* aggregate (dandelions), and investigate features of plastome variation that may be a consequence of apomixis, comparing at taxonomic distances from tribe to Eudicots.

Chapter 5. Using graph based cluster method of the genomic sequences to analyse repeat composition, identification, and quantification of major groups of repetitive DNA family sequences, including transposable elements comparing with Helianthus and angiosperms. Investigating sequence diversity of repeats, distribution, and abundance of transposable elements between the three related *Taraxacum* microspecies in order to understand the nature and consequences of genetic variation between *Taraxacum* microspecies, and role of transposable elements in generating variation in sexual against agamospecies. Also investigate the characterisation with distribution of different repetitive DNA resulting from cluster methods.

Chapter 6. Based on 46 Gb of Illumina raw reads from three *Taraxacum* agamospecies, what is the distribution of short k-mers (short sequence motifs where k is between 10 and 150 bp long)?

1. What is the nature of the most abundant k-mers?
2. Where are the most abundant k-mers located on the chromosomes of *Taraxacum*?

CHAPTER 2

Materials and Methods

2.1. Materials used

2.1.1. Plant material

Eighteen different *Hieracium* agamospecies were obtained from the Leicester Botanic Garden, on 25 March 2013, which they are used in the study of (Thomas *et al.* 2011), (Table 2.1a).

Some, 17 new different lines from five different agamospecies of *Taraxacum* accessions were identify and genotyped by Ľuboš Majeský from Palacký University, Olomouc, Czech Republic (Majeský *et al.* 2012) has been obtained. In addition, seeds of two species were obtained from Dr. R. Vašut in Moravian Silesia in the Czech Republic, diploid dandelion *Taraxacum linearisquameum* and the triploid *Taraxacum gentile* (Musiał *et al.* 2012). Also, six different *Taraxacum* agamospecies plant collected in the different location in Leicester city-UK with pulling out most of the plant roots as much as possible (Table 2.2b).

All species and accessions germinated and replanted in the pot. The plants grew by two duplications with two different place but same conditions, one in Leicester botanic garden and another one in Leicester university department of Genetic in growth cabinet under the optimal environmental condition (25 ˚C, 16 h light/15 ˚C, 8 h dark; humidity 60%; and intensity for light levels >150 µmol/m/s).

2.1.2. Solutions and media

All solution and Media used in this study listed in Appendix 9.1.

Table 2.1. The sexual and agamospecies of (a) *Hieracium* and (b) *Taraxacum* including *T. officinale* agg. (Majeský 2013) species (Asteraceae) used in this study.

(a)

	Taxon	Source	Section	Chromosome No.	Ploidy levels	Accession or reference
1.	*H.amaurostictum* Walter Scott & R.C.Palmer	Semblister	Alpestria	2n=36	Tetraploid	2002.048
2.	*H.attenuatifolium* P.D.Sell & C.West	Laxo Burn	Alpestria	2n=36	Tetraploid	1998.142/143
3.	*H.australis* (Beeby)Pugsley	Burrafirth area, Unst	Alpestria	2n=36	Tetraploid	2006.016
4.	*H.breve* Beeby	Ronas Voe	Alpestria	2n=36	Tetraploid	2002.050
5.	*H.difficile* P.D.Sell & C.West	Okraquoy	Alpestria	2n=36	Tetraploid	2002.041
6.	*H.dilectum* P.D.Sell & C.West	Laxo	Alpestria	2n=36	Tetraploid	2005.001
7.	*H.gothicoides* Pugsley	Lunning	Tridentata	2n=37*	Triploid	2002.046
8.	*H.gratum* P.D.Sell & C.West	Burra Firth, Unst	Alpestria	2n=36	Tetraploid	2005.024
9.	*H.hethlandiae* (F.Hanb.) Pugsley	Mavis Grind	Alpestria	2n=36	Tetraploid	2005.025
10.	*H.lissolepium* Roffey	Eric's Ham, Yell	Tridentata	2n=36	Tetraploid	2002.049
11.	*H.northroense* Pugsley	Burravoe, North Roe	Alpestria	2n=27	Triploid	2006.018
12.	*H.pugsleyi* P.D.Sell & C.West	Whale Firth, Yell	Alpestria	2n=36	Tetraploid	2005.023
13.	*H.scottii* P.D.Sell	Near Windy Scord	Oreadea	2n=36	Tetraploid	2002.047
14.	*H.spenceanum* Qalter Scott & R.C.Plamer	Sandness	Alpestria	2n=36	Tetraploid	2006.013
15.	*H.subscoticum* P.D.Sell	Ronas Voe	Oreadea	2n=27	Triploid	2006.015
16.	*H.subtruncatum* Beeby	Scarvister, West Mainland	Alpestria	2n=36	Tetraploid	2006.019
17.	*H.vinicaule* P.D.Sell & C.West	Whale Firth	Alpestria	2n=27	Triploid	2006.012
18.	*H.zetlandicum* Beeby	Isbister, North Roe	Alpestria	2n=36	Tetraploid	2006.014
19.	*H. umbellatum* L.	-	-	2n=18	Diploid	-

← Table 2.1. Continue…

(b)

	Taxon	Source	Section	Chromosome No.	Ploidy levels	Identifier code
1	*T. obtusifrons* Markl.	Czechia	Ruderalia	2n=24	Triploid	OVS
2	*T. obtusifrons* Markl.	Czechia	Ruderalia	2n=24	Triploid	O978
3	*T. obtusifrons* Markl.	Czechia	Ruderalia	2n=24	Triploid	O914
4	*T. stridulum* ined.	Czechia	Ruderalia	2n=24	Triploid	S3
5	*T. stridulum* ined.	Czechia	Ruderalia	2n=24	Triploid	S983
6	*T. stridulum* ined.	Czechia	Ruderalia	2n=24	Triploid	S933
7	*T. pulchrifolium* Markl.	Czechia	Ruderalia	2n=24	Triploid	PUL943
8	*T. amplum* Markl.	Czechia	Ruderalia	2n=24	Triploid	A976
9	*T. amplum* Markl.	Czechia	Ruderalia	2n=24	Triploid	A978
10	*T. amplum* Markl.	Czechia	Ruderalia	2n=24	Triploid	AK07
11	*T. amplum* Markl.	Czechia	Ruderalia	2n=24	Triploid	A5
12	*T. jari-cimrmanii* ined.	Czechia	Ruderalia	2n=24	Triploid	TP983
13	*T. gentile* Haglund & Railonsala.	Poland	Ruderalia	2n=24	Triploid	T.gen
14	*T. linearisquameum* Soest	Poland	Ruderalia	2n=16	Diploid	T.lin
15	T.sp.	Leicester/UK	-	2n=24	Triploid	Ta
16	T.sp.	Leicester/UK	-	2n=24	Triploid	Tb
17	T.sp.	Leicester/UK	-	2n=24	Triploid	Tc
18	T.sp	Leicester/UK	-	2n=24	Triploid	Td
19	T.sp.	Leicester/UK	-	2n=24	Triploid	Te
20	T.sp.	Herefordshire/UK	-	2n=24	Triploid	Tp

2.2. Methods

2.2.1. DNA extraction

Total DNA including nuclear, mitochondrial and plastome DNA extracted from fresh green young leaves collected from agamospecies *Taraxacum* microspecies and *Hieracium* using Cetyl-Trimethyl-Ammonium Bromide (CTAB) by Doyle and Doyle (1987) with some modification to obtain high quality and quantity of DNA.

Freshly young leaves washed, dried, and removed the midrib. Extraction of total genomic DNA conducted by using 1.5-2.0 g. Leaves wrapped in labelled aluminium foil and placed them inside liquid nitrogen (N_2) in order to quick freeze.

Extraction buffer prepared and preheated at 65 °C for at least 30 minutes, extraction buffer consist of 2% autoclaved CTAB solution, 2% PVP (Polyvinylpyrrolidone –Sigma-Aldrich) powder, and 2% β- merceptoethanol. The proportions of the three components in DNA extraction buffer was as follow under different volume of extraction buffer:

CTAB buffer (ml)	**PVP (gm)**	**β-merceptoethanol (µL)**
0.5 ml	0.01	10
5	0.1	100
20	0.4	400

Leaf tissue placed into a pre-autoclaved mortar or HCl washed with a pinch or one spatula volume of general-purpose grade fine sand (Sand-low iron) and liquid nitrogen. Frozen leaves grounded to a fine powder. The powdered leaf tissue transferred and scraped into 5ml of extraction buffer in a 50 ml tubes, and mixed gently, until the mixture reached a slurry-like consistency. It is very important to avoid leaving dry leaves material around the rim of the tube, and all these steps should do very quickly to avoid melting of the leaf tissue, otherwise will result to share the

extracted DNA. The mixture incubated for 60-90 minute at 60-65 ˚C in a water bath, with inverting gently several times every 15 minute.

The next step was centrifugation of the leaf material with the extraction buffer at 5,000 rpm for 10 min, then clear supernatant solution (DNA contained) collected carefully with the wide-bore pipette (the tips of the blue pipette tips slantingly cut to pipette-out the solution) and transferred into a clean centrifuge tube (15 ml). An equal volume of chloroform : iso-amyl alcohol (24:1) was added and the samples shaken by hand for 5-10 minutes at room temperature, then the mixture centrifuged at 5,000 rpm for 10 min (this step repeated 2 times).

The upper phase transferred to a clean new tube. The DNA was precipitated with 2/3 (0.66) volume of ice-cold isopropanol (100%) and 0.08 volume of 7.5 M cold ammonium acetate, according to the aqueous of the solutions the mount of isopropanol and ammonium acetate was added as follow:

Aqueous phase (ml)	Ammonium acetate 7.5 M (µL)	Isopropanol 100% (µL)
4.0	320	2,160
4.5	360	2,430
5.0	400	2,700

The mixture inverted several times, and then incubated overnight on the ice at the cold room. The DNA pellet pooled out with clean glass sticks, or the solution spin down at 2000 rpm for 2 minute. Then, the pellet dried briefly and transferred to wash buffer for 20 minute before dissolving in 250-500 µl of (1 x TE) overnight.

Finally, DNA was treated with 1 µl of RNAse (10 mg/ml) for 1 hour at 37˚C. Isolated genomic DNA was diluted with ddH_2O (Sigma-Aldrich) by 1:10 then from this concentration to working on, and stored at -20 °C until use. The remaining stock kept at -80 °C for longer storage.

2.2.2. Quantitation of total genomic DNA

2.2.2.1. Gel electrophoresis

The determination of concentration and integrity of DNA extracts were conducted by running the mixture (2 µL genomic DNA, 3 µL sigma water, and 3 µL loading buffer 0.25X) on a 0.8% (v/w) standard agarose gel (Bioline), alongside HyperLadder™ 1kb (Bioline).

2.2.2.2. NanoDrop Spectrophotometry

NanoDrop 8000 Spectrophotometer (Thermo Scientific) was used to assess the quantity (concentration-ng/µl) and quality (purity ratio-A260/A280; A260/A230) of DNA (genomic and eluted). The NanoDrop was first blanked using either elution buffer or ddH_2O (Sigma-Aldrich) according to the material used to elute the DNA. Readings were taking by using 1.5 µl of extracted genomic DNA. The acceptable DNA samples should give high molecular weight with no visible shearing on gels and reading of NanoDrop for the ratio A_{260}:A_{280} should be between 1.8-1.9 of purity; such DNA samples were used for subsequent PCR amplifications, restriction digestion experiments, and Next Generation Sequencing (NGS).

2.2.3. Primer design

The PCR primer pairs were designed using the Primer3 (Rozen and Skaletsky 1999) within Geneious program, which allows to set the location of the primers within the sequences, choosing primer length, melting temperature, and expected product size. Then primers have been purchased from Sigma (www.sigmaaldrich.com/). Further details about primer markers are given in the respective results chapters.

2.2.4. Polymerase chain reaction (PCR) amplification

Different types of repetitive DNA, chloroplast and nuclear markers were amplified from *Taraxacum* (*Hieracium*) total genomic DNA using specific and degenerated primers.

Total genomic DNA was amplified using a T professional Gradient Thermocycler (Biometra) in a 15 µL reaction mixture containing 50–100 ng of template DNA, 1x Kapa Biosystems buffer A [750 mM Tris–HCl pH 8.8, 200 mM $(NH_4)2SO_4$, 15 mM $MgCl_2$, 0.1 % Tween-20], 1.5 mM $MgCl_2$, 200 µM dNTPs (Bioline), 0.4 µM of each primer (0.2 µM forward, 0.2 µM reverse; Sigma-Aldrich), and 0.5 U of Kapa Taq DNA polymerase (Kapa Biosystems, USA), the mixture made up to final volume of 15 µl with ddH_2O. Each reaction was runs with negative control, in which the template DNA was replaced with 2-µl ddH_2O. The cycling conditions for each primer set are given in the relevant chapter.

2.2.5. Agarose gel electrophoresis

Agarose gel electrophoresis used to analyse PCR amplified DNA fragment and restricted enzyme digestions. For each of both TEs and IRAP usually high percentage of agarose gels 1.5-2% W/V were used, because of requiring resolution of bands below 2kb. The rest PCR amplification including chloroplast, other nuclear DNA amplifications, and cloning the concentration of agarose gel used was 1%. For IRAP markers, to obtain a good separation and sharpness of bands 'high resolution' (Super AGTC Agarose, Geneflow, UK) types of agarose were used, as mixture of 1 part high-resolution with 3 parts multipurpose agarose gels, and for other PCR reactions (TEs and cloning PCR) normal agarose used alone.

Agarose gel was dissolved in 1 x TAE by microwaving. 1 µl of ethidium bromide (EtBr) (0.5 µg/ml) were added for every 1 ml TAE. The gel was immersed in a gel electrophoresis tank containing 1 x TAE as the running buffer. Samples were then mixed 4:1 (v/v) with 0.25, 0.5, 1 x loading dye (for IRAP 15 µL PCR product with 5 µL loading buffer 0.25 x and for PCR cloning and other reactions 3 µL PCR product with 2 µL loading buffer 0.25 x). The mixture pipetted into the wells and the gel run at 4-5 V/cm for 45 minutes to 1 hour (for IRAP the gels were run under a reduced voltage for 3-4 hour). The hyperLadder 1 (Bioline) or Q-step 2 (YorkBio), were used to obtain the molecular weight and the concentration of PCR products. Gels were then visualised under UV light in a GeneFlash (Syngene) gel documentation system.

2.2.6. Purification of PCR products

After gel electrophoresis, the PCR products or excised bands from gel purified with the NucleoSpin® Extract II Clean- up Kit and PCR clean-up kit (Machery-Nagel) gel extraction kit following manufacturer's instructions. Purified amplicons were then assayed using NanoDrop 2000, and stored at -20 °C until use.

2.2.7. Cloning

DNA fragment was ligated into pGEM-T easy vectors (available commercially) by using pGEM-T Easy Vector System I (Promega). An over hanged single base 3' thymidine of pGEM-T Easy vector, have ability to ligate with an over hanged single base 3' adenine of PCR products generated by *Taq* DNA polymerase (Figure 2.1).

2.2.7.1. Ligation

Ligation conducted by inserting the PCR amplicons into vectors, in a reaction mixture containing: 2 x Rapid Ligation Buffer (30 mM Tris-HCl, pH 7.8; 10 mM $MgCl_2$; 10 mM DTT; 1 mM ATP; 5% PEG from Promega), 50 ng pGEM®-T Easy vector (Promega), 3 Weiss units of T4 DNA ligase (Promega) and 50-250 ng purified PCR product, made up to 10 µL with ddH_2O. The component mixture has been set up on ice, as follow:

Reagent	Standard reaction
2X Rapid Ligation Buffer	7µL
pGEM®-T Easy Vector	0.9 µL
PCR product	100 ng/ µL
T4 DNA Ligase	1.2 µL
Deionized water to a final volume	? µL
Final volume	15 µL

Ligation reactions performed in a 300 µL tubes. Rapid Ligation Buffer had vortex vigorously before each add. The reaction mixture was mixed well by pipetting or by vortex, then the mixture incubated for one hour at room then kept at 4˚C overnight, to gain the maximum number of transformation.

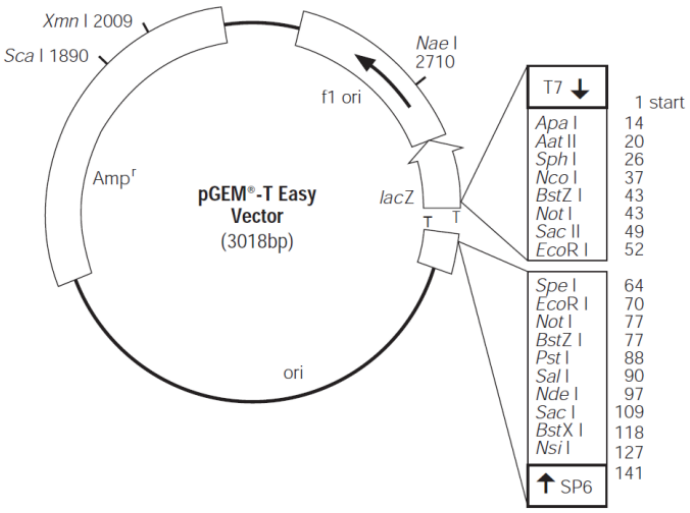


Figure 2.1: pGEM®-T Easy Vector circle map (www.promega.com**).**

2.2.7.2. Transformation

Vectors were transformed into *Escherichia coli*. The competent cells (α-select bronze competent cells - Bioline, kept at -80 °C) were thawed on ice and mixed by gently flicking the tube, then five µL of the overnight ligation mixture mixed with 50 µL competent cells. The transformation mixture was then incubated for 30 minutes on ice, to make the cell membrane poles enclose slowly, and then heat-shocked for 1 min at exact 42 °C, and quickly returned to the ice for 10 minutes. 700 µL SOB (Super Optimal Broth) medium previously warmed in 37 °C were added on the mixture then incubated for cell growing for 3 hour at 37 °C in an orbital shaking incubator (230 rpm; Gallenkamp). In a well sterile flame hood, plating of the cell cultures were conducted by spreading 100, 150, 200 µL of the culture onto LB agar plates containing 100 µg/ml ampicillin, 40 µg/ml X-Gal (5-Bromo-4-chloro-3-indolyl-β-D-galactosisdase) and 500 µM IPTG (Isopropyl-β-A-thiogalacto-pyranoside). Sterile spreader was used to spread the culture on the surface of the plates. The plates incubated overnight at 37 °C.

2.2.7.3. Screening

Blue-white screening method were used to screen recombinant cells, the technique that enables identification of recombinant bacteria (pGEM-T Easy vectors), contained lacZ gene encoding β-galactosidase. The chromogenic X-gal hydrolyse and break down

substrates which leads to turning colonies colour to blue on antibiotics LB agar plate. Where as antibiotic LB agar plate produce white colour colonies when PCR products ligate into the pGEM-T Easy vector, by disrupting the open reading frame of the lacZ gene.

Only white colonies, formed by recombinant cells (plasmid plus insert) have been chosen, with ignoring blue colonies. Sterile toothpick used to collect the colonies from culture plates, then sub-cultured in 5 ml of LB solution with 5µL (40 µg/ml) ampicillin (miniprep) overnight at 37 °C (230 rpm).

2.2.7.4. Verification of insert size and M13-PCR amplifications, purification of plasmid DNA, and storage of E.coli cells

In order to confirm the presence of recombinant plasmids in white colonies and confirm the size of the insert,

M13-PCR amplification was used to carry out colony PCR, in order to confirm presence of the recombinant plasmids in white colonies and confirm the size of the insert. The reaction mixture content as described above (section 2.2.4), with adding 2 µL of overnight culture (recombinant plasmid DNA) as a template or directly inoculate the collected clones to the PCR tubes. Primer sequences: M13 Forward (5' GTA AAA CGA CGG CCA GT 3') and M13 Reverse (5' GGA AAC AGC TAT GAC CAT G 3') were used. PCR cycling conditions: 5 minute at 95 °C plus 30 cycles of (1 minute at 95°C, 30 seconds at 55 °C, 1 minute at 72 °C) with the final extension of 72 °C for 10 minute.

2.2.8. DNA sequencing

2.2.8.1. Sanger sequencing for sequencing of PCR amplicons or cloned PCR products

Purified DNA fragments (IRAP, amplification of RT domain by generated primers, chloroplast and other nuclear amplified DNA) were sequenced commercially at Source Biosciences (Nottingham, UK) or GATC Biotech (London, UK) either by sending the PCR products directly using custom primers or with universal M13 forward or reverse primers, using recombinant plasmid DNA. Sample concentration was 1 ng/µL per 100

bp for PCR products and 100 ng/µL for plasmid DNA. Primer diluted 1:100 times. in most of the cases reverse primers were sent with samples for sequencing.

2.2.8.2. Next generation sequencing

Whole genomic DNAs were sequenced commercially (Interdisciplinary Centre for Biotechnology Research, University of Florida, USA); accession S3 was sequenced with Illumina MiSeq 2x300bp paired-end reads while accessions O978 and A978 were sequenced using Illumina NextSeq500 2x150bp reads. These Illumina sequencing data are submitted under the BioSample accession number (SAMN05300515, SAMN05300516, SAMN05300517).

2.2.9. Dot blot southern hybridization

2.2.9.1. Genomic DNA digestion by endonuclease enzyme

Taraxacum (*Hieracium*) genomic DNA digested with *HaeIII, HindIII, BamHI, Sau3A I, DraI and EcoRI* restriction enzymes (New England Biolabs). 1µL of most restriction enzyme preparations was enough to cut 10 µg of DNA or more, because of high concentrations of restrictions enzymes. Digestions of restriction enzymes were performed in a reaction mixture of: 1 x appropriate buffers following manufacturer's instructions (New England Biolabs), 10 U restriction enzyme (New England Biolabs) and 1 µg high-quality DNA, made up to 10 µL with ddH_2O. The mixing steps conducted on ice, because of sensitivity of restriction enzymes for temperature over 37 °C. Reactions were incubated for 2 hr at 37 °C, and then stopped by adding 3 µL of 1 x loading dye. Restriction fragments were separated and visualised by gel electrophoresis, 1.5% (v/w) standard agarose, alongside HyperLadder™ 1kb. Agarose gels and electrophoresis was carried out at a slow speed of 30 V in 1 x TAE buffer for 2-4 hour.

2.2.9.2. Tailed blunt-end DNA fragments treatments

To ligate the DNA were fragmented by restriction enzymes to Promega T-Vectors successfully, a single nucleotide of adenine (dATP) were added to the 3' DNA fragment. Due to presence of a single 3' terminal Thymidines (T) nucleotide at each end of pGEM®-T Easy vector, which complementing by ligating to a single base

deoxyadenosine (A) to the 3' end of DNA fragment generated by *Taq* polymerase. The components and their amount in tailed blunt-end DNA fragments treatments were as follow:

Components	Final Concentration	Amount (µL)
DNA fragment (from digested DNA with enzyme)	----	7
Taq DNA polymerase reaction 10X buffer	1X	1
MgCl2	2 mM	1
dATP (Bioline)	0.2 µM	1
Taq DNA polymerase (Kapa Biosystems)	10 U	1

2.2.9.3. Selection of plasmid clones for dot blot hybridization

After separating of digested genomic DNA by 1% gel electrophoresis, the gel cut out from 1-2.5 kb size fragments size. Then, the purified DNA fragments processed to cloning as described above. A recombinant DNA library was made from *DraI* and *HaeIII* digest of genomic DNA (*Taraxacum officinalis* and *Hieracium northroense* Pugsley), by choosing 50 white colonies in two replicate LB agar plates, which processed by collecting a white colonies from the main colony plates, using a sterile toothpicks. The colonies inoculated in the same colony number in both replicate previous graded LB agar plates. Two replicate plates were used in order to use one plate for colony transfer and the second plate for selection of potential colonies for plasmid DNA isolation. Library plates were incubated at 37 ˚C overnight.

2.2.9.4. Transfer of bacterial colonies onto charged nylon membrane

The hybridizations steps followed the standard technique of Sambrook *et al.* (2001) with some modifications. Positively charged nylon membrane (Hybond N^+, Amersham Biosciences) of appropriate size (90mm Petri dish) was marked with the pencil at three asymmetric locations to identify the orientation of the membrane in the Petri dish. The membrane placed carefully upside down on the surface of the 50-grade colonies

library in LB-agar plate, and ensured contact the membrane with the bacterial colonies. Four petri dishes used to next washing steps of the membranes using four different solutions (3MM Whatman). Each petri dish contained 5ml of 10% SDS (for 3 minutes), denaturing solution, neutralization solution, and 2 x SSC solutions (for 5 minutes) respectively. With carefully avoidance of coming up the solution over the colony side face. At the end, the membranes air dried (30 minutes) then wrapped with cling film and aluminium foil, then were incubated at 80 ˚C for 2-3 hour, following by incubating overnight at -4˚C.

2.2.9.5. Pre-hybridization of membrane

Next day, the membrane took out from -4˚C to room temperature for 10 minutes. 2 X SSC for 5 minutes and 0.1 X SSC /0.1% (w/v) SDS rehydrated the membranes respectively for 1 minute. The pre-hybridization mixture was prepared (5ml per 100 cm^3 of membrane) as follow:

Component	Final	Amount/5ml
50X Denhardts	5X	0.5 ml
20X SSC	4X	1 ml
10% SDS	0.5%	0.25 ml
Salmon Sperm DNA (denatured) (10ng/µL)	100 ng/ µL	25 µL
EDTA (0.5 M)	10mM	100 µL
Deionized water	----	2.125 ml
		4 ml

The membrane and pre-hybridization mixture placed in a roller bottle and were rotated for 4 hours at 55˚C in Thermohybaid hybridization oven (Ashford, UK).

2.2.9.6. Hybridization of the membrane

The roller bottle removed from Thermohybaid hybridization oven, 1 ml of pre-hybridization solution removed from the roller bottle and mixed with 3-4µl (corresponding to ~150ng) of digoxigenin genomic (*T. officinalis* and *H. northroense*) labelled probe (for more detail about labelling genomic DNA see section 2.2.10 of this chapter). Each of probes and freshly denatured salmon sperm DNA (denatured at 95˚C

for 5 minutes followed by 5 minutes on ice) replaced along with 1ml of 50% (w/v) dextran sulphate and then hybridized at 55 °C for 16-18 hours with constant rotation.

2.2.9.7. Post-hybridization of the membrane

Post-hybridization steps are consisting of several high stringency washing steps. The washing steps started with washing the membranes twice with 2x SSC x 0.1% (w/v) SDS at 56°C for 5 min (64% stringency), then, followed by washing twice with 0.5x SSC x 0.1% (w/v) SDS for 15 min each at 56°C (equivalent to 82% stringency). All washed steps carried out inside Thermohybaid hybridization oven or water bath with rolling by hand continually.

2.2.9.8. Detection

Membrane washed for 5 minutes in 10 ml washing buffer-1, then 10 ml of buffer-2 for 30 minutes followed by incubating for 30 minutes with 10 ml of antibody conjugate solution [anti-digoxigenin conjugated to alkaline phosphatase (Roche Diagnostics)] with final dilution of 150U/ml (1:5000) in buffer 2. Next, the antibody washed out by using 10 ml buffer-1 to wash the membranes twice for 15 minutes, then equilibrated for 5 min with buffer-3.

In the dark room, after draining out the excess of buffer-3, the membranes were incubated with 500 µl of CDP-star solution (Roche Diagnostics) diluted 1:100 in buffer-3 for 5 minutes. Then the membrane drained and wrapped in a cling film, then placed to autoradiographic cassette in complete darkness. The chemiluminescence was recorded by keeping X-ray film (Fuji Medical X-Ray film) of appropriate size below the membrane. Different exposure times from 1-15 min were given to detect all possible signals. X-ray films were developed using the automatic photographic developing machine and scanned with EPSON Expression Pro 1600.

Colonies showing strong hybridisation to either *Taraxacum* or *Hieracium* were selected. The inserts in the plasmids were sequenced commercially from chosen colonies.

2.2.10. Probe labelling

DNA fragments were labelled by indirect fluorophores labelling (Schwarzacher and Heslop-Harrison 2000), in order to detect biotin fluorophores were integrated to avidin or streptavidin, or for detecting digoxigenin anti-digoxigenin used, which was incorporated into probe DNA conjugated to dUTP or dCTP. Additionally, we used some probe labelled directly with fluorophore (ordered directly form Sigma Aldrech company) which in this case no antibodies are needed as the nucleotides have been linked directly with fluorophores.

2.2.10.1. M13-PCR labelling

The DNA fragments smaller than 500 bp in size such as cloned repetitive DNA like 5S rDNA (pTa794) were labelled by using PCR amplification, through universal M13 primers. The reaction mixture was prepared by adding 1 µl of biotin-16-dUTP or digoxigenin-11-dUTP (1mM, Roche Diagnostics) or 1µl of water as the control to the standard PCR mixture and amplified as described above. Amplifications were conducted in a reaction mixture containing: 1 x Buffer A, 1.5 mM $MgCl_2$, 0.4 µM M13 primers (0.2 µM forward, 0.2 µM reverse; Sequences as above), 0.4 mM dNTP mix, 20 µM digoxignin-11-dUTP or biotin-16-dUTP, 0.5 U Kapa Taq DNA polymerase, and 100 ng from the DNA probe, made up to 50 µL with ddH_2O.

2.2.10.2. Labelling DNA fragments by using Random primers

Most of the DNA fragments labelled in current study were labelled with BioPrime® Array CGH Labelling System (Cat. No. 18095-011, www.invitrogen.com), in a final volume of 50 µl reactions, following manufacturer's instruction. Genomic DNA labelling was started with sharing genomic DNA to 3-5 kb pieces through autoclaving the genomic DNA at 110°C for 4 minutes before labelling. Autoclaved DNA gel electrophoresed on 1% agarose gel to estimate fragment size. Consequently labelling procedure were conducted by Random Primer method. 200 ng of amplified DNA fragments or 1 µg of sheared genomic DNA were mixed with 20 µl of 2.5x Random Primer Solution. The mixture denatured by PCR machine for 5 minutes at 95 °C, then immediately incubated on ice for 5 minutes. The to complete the reaction mixture 5 µl

of 10x dNTP Mix and 1µl of 40U Klenow Fragment were added, then incubated at 37 °C for 2 hour or overnight at room temperature. Then 5µl of Stop Buffer (0.5M EDTA pH 8.0) added to stop polymerization reactions. Labelled probes were purified to remove any unincorporated nucleotides, enzyme and salts using BioPrime® Purification Module (Invitrogen) or NucleoSpin® Extract II Kit (MACHERY-NAGEL), following manufacturer's instructions (http://www.mn-net.com/tabid/1452/default.aspx) and stored at -20 °C.

2.2.10.3. Testing the incorporation of labelled nucleotides (dot-blot)

The incorporation of labelled nucleotides (digoxigenin-11-dUTP or biotin-16-dUTP) within probes was tested using a colorimetric dot-blot, according to Schwarzacher and Heslop-Harrison (2000).

DNA labelled were bound to a charged nylon membrane, by soaking a small piece of positive charged Hybond-N^+ membrane (Amersham) in buffer 1 for 5 minutes and semi-dried between two filter papers (Whatman). Place of each probes have been marked on the membrane, and then 1 µL of the probe applied on the nylon membrane. The membrane washed with Buffer 1 for 1 minute and then buffer 2 for 30 min. 0.5 ml antibody solution [1.5 U/ml anti-digoxigenin-AP (Roche) and 2U/ml streptavidin-AP-conjugate (Life Technologies) in buffer 1] were applied to the membrane to expose the probes to alkaline phosphatase conjugates. Plastic coverslip placed on top, and then the membrane incubated for 30 minutes at 37 °C in the dark. The membrane was then washed twice, first in buffer 1 for 15, buffer 3 for 2 minutes consequently. The conjugated alkaline phosphatase was then provided with a substrate by applying 1.5 mL detection solution (0.33 mg INT/BCIP in buffer 3; Roche) to the membrane, and incubating for 10 minutes at room temperature in dark. The incorporation of labelled nucleotides was then visualized by the degree of coloured product. Qualities of the labelled probes were estimated by the strength of the coloured product, darker colour was the best probe labelled.

2.2.11. Chromosome preparations

2.2.11.1. Collection and fixation of root tips

Actively growing *Taraxacum* (*Hieracium*) root tips of small size (about 1-1.5 mm in width) were collected between 9:00-11:30 am from potted plants grown under cabinet growth conditions and light. Then the roots were pre-treated and in the room temperature according to Bailey and Stace (1992), Schwarzacher and Heslop-Harrison (2000) using 0.002 M 8-hydroxyquinoline (BDH Chemicals) either overnight at 4 °C or 2 hours at room temperature then transferred to 4 °C for another 2 hours. The roots washed, cleaned and fixed in 3:1 (v/v) ethanol: glacial acetic acid for 30 minutes at room temperature, then freeze until use.

2.2.11.2. Metaphase chromosomes preparation

Root apical meristems squashes conducted by using aceto-orcein according to Bailey and Stace (1992). Root tips were hydrolysed in 5 N HCl at room temperature for 10 minutes and then transferred to 70% ethanol until use (same day). To prepare metaphase chromosome spread, the root tips were dissected, stained, and squashed in aqueous 2% (w/v) aceto-orcein (Sigma-Aldrigh; Darlington and Lacour 1960). Observation of chromosomes metaphases were conducted under bright field on a Zeiss Universal microscope. At least five well-spread metaphases from different root tips were used to record the number of somatic chromosome.

Air-dried methods were used to prepare metaphase slides for using in florescent *in situ* hybridizations (FISH) and genomic *in situ* hybridizations (GISH). The fixed root tips were washed in destilled water for two times and 10 minutes each, then washed in a 1x enzyme buffer for 2x10 minutes, then digested in enzyme solution for 45-1 hour (according to the species) at 37 °C. After that, the digested root tips were washed twice with 1 x enzyme buffer. Then 45 or 60 % (v/v) acetic acid were used to dissecting, staining, and squashing the digested root tips. To preserve the chromosome preparations directly immersed the slides in to liquid nitrogen for a few seconds then the cover slips removed quickly. The slides air-dried (Conger and Fairchild 1953) in the room temperature or by incubating them in 37 °C over night. Slides qualities were

performed by scanning them under phase contrast on a Zeiss Universal microscope. High quality slides were stored in a dry box at -20 °C with silica gel until use.

2.2.12. Fluorescent *in situ* hybridization (FISH)

Fluorescent *in situ* hybridisation (FISH) was carried out according to Schwarzacher and Heslop-Harrison (2000) with some minor modification, as summarized as follow:

2.2.12.1. Pre-hybridization

Re-fixation of chromosome preparations in fresh 3:1 (v/v) ethanol to glacial acetic acid for 15-30 minute at room temperature. Then, slides were washed 2x10 minutes in 100% ethanol and air-dried in room temperature.

The excess of RNA were removed from chromosomes by incubating the slides with 200 µL RNase solution (100 µg/ml in 2 x SSC; Sigma-Aldrich) with placing a large plastic coverslip (25 × 30 mm), followed by incubation for 1 hours at 37 °C in a humid chamber. After that the coverslips were removed and the chromosome preparations washed twice in 2 x SSC for 10 minute.

A pepsin solution was used for removing excess cytoplasm. The incubation times depend on the density of the cytoplasm around the cells, and cell size. *Taraxacum* chromosome preparations were incubated for 1- 2 minute at room teperature. The chromosome preparations were then washed, first in distilled water for 1 minute and then in 2 x SSC for 5-10 minute.

Chromosome preparation refixed by incubating slides in 4% formaldehyde (Fisher Scientific) for 10 minutes at room temperature, thentwoo times in 2 x SSC for 5-10 minute. Next, washing with an ethanol series (50%, 70%, 85% and 100%) were used to dehydrate chromosome preparations. Then the slides were air-dried completely in the room temperature.

2.2.12.2. In situ hybridisation

A probe mixture was prepared, by mixing: 50% (v/v) deionized formamide (Sigma-Aldrich), 10% (w/v) dextran sulphate (Sigma-Aldrich), 0.125% (w/v) sodium dodecyl sulphate (SDS; Sigma-Aldrich), 1 µg/ml salmon sperm DNA (Sigma-Aldrich), 2 x SSC and 200 ng of each probe, then the mixture made up to 40 µL with ddH_2O. The probe mixture was then denatured for 10 minutes at 85 °C by using a PCR machine, followed by quick transferring them on to ice (≥10 min) to prevent reannealing. 40 µL of probe mixture applied to the chromosome preparations with placing a small plastic coverslip (22 × 22 mm). The chromosome preparations were incubated in either heated block (Thermo Scientific) or PCR machine with placing the metal plate on the block. The incubation were first started with denaturing the chromosome preparations for 5-7 minutes at 70--72 °C, and then the temperature reduced to 37 °C for 16-20 hour to enable the complementary targets on the chromosome preparations hybridise with labelled probes.

2.2.12.3. Post-hybridisation washes

To remove the excess of hybridisation mixture and any unbound probe series of post-hybridisation washing steps were given to the chromosome preparations. Starting with washing the slides twice by 2 x SSC for 2 minutes, then 5 minutes at 42 °C, once in a stringent wash solution (in current study just low stringency was used: 0.1 x SSC) for 10 minutes at 42 °C, then in 2 x SSC for 5 minutes at room temperature, and finally once in detection buffer for 5 minutes at room temperature.

2.2.12.4. BSA block

Blocking solution, 5% BSA (Bovine Serum Albumin) in detection buffer were used to block the non-specific sites which could bind detection reagent. The 200 µL from the blocking solution were applied to chromosome preparations, and covered with a large plastic coverslip, then incubated for 20-30 minutes at 37 °C in a humid chamber.

2.2.12.5. Detection

A detection solution was prepared by mixing the blocking solution with 1 µg/ml of each of FITC (Fluorescein isothiocyanate) conjugated to anti-digoxigenin-fluorescein

(Roche) and Alexa Fluor® 594 streptavidin (Invitrogen). The detection solution then were applied (55 µL) on each slides of the chromosome preparations, a small plastic coverslip (18 × 18 mm) was placed on top, then the slides incubated at 37 °C for 1 hour. Then after, the chromosome preparations were washed 2-3 times in a detection buffer for 8 minutes at 40 °C.

2.2.12.6. Nuclear counterstaining and mounting

The cells were counterstained and mounted by the mixture contained 6 µL of DAPI (stock 100 µg/ml) and 97 µL antifade (Citifluor), and 97 µL ddH_2O, to prevent fading the fluorescent signal under fluorescence. Finally, a large cover glass (No.0, 24 × 40 mm) was placed on top. After all, the slides were stored at 4 °C in the dark for overnight.

2.2.12.7. Microscopy and imaging

Observation of the chromosomes preparations after *in situ* hybridization was performed under immersion oil (Zeiss) on a Nikon Eclipse 80i fluorescent microscope in a dark room. In order to discrete signal detections, three Nikon filters were used to view the preparations, from each of three fluorophores, UV-2E/C (excitation filter wavelengths -340-380 nm, emission filter wavelengths 435-485 nm) for DAPI, B-2E/C (excitation filter wavelengths-465-495 nm, emission filter wavelengths- 515-555 nm) for fluorescein and G-2E/C (excitation filter wavelengths-528-553 nm, emission filter wavelengths-590-650 nm) Alexa Fluor® 594.

Nikon DS-Qi1 digital camera and NIS Elements AR, version 3.2, software were used to take a photograph. Images were later processed with Adobe Photoshop CS3, using only those functions that treat all pixels uniformly and for placing the scale bar for each picture.

2.2.12.8. Reprobing

Occasionally chromosome preparations could use often for the second or even third time, by reprobing them. Firstly, the slides that use to reprobe, should clean with any diffused oil carefully. The slides were placed in 37°C for 10 minute in order to reduce

the viscosity of the antifade, then the cover glass removed by the razor blade. Detection buffer were used to wash the preparation two times for 30-60 minute at room temperature (some times higher temperature should be use to remove the conjugated probes). Then washed twice with 2x SSC for 5 minutes at room temperature. Then the chromosome preparations were dehydrated by using several washing steps with ethanol series 50%, 70%, 85% and 100%, then left to air dry. The *in situ* hybridization steps continued as described in the section (2.2.12.2.) to end of the FISH procedure.

2.3. Molecular analysis

2.3.1. Bioinformatics and computational analysis

In the current study Geneious program were used, as the one of the popular bioinformatics software created by Kearse *et al.* (2012), available online from http://www.geneious.com/. Most of the bioinformatic works were conducted by using Geneious, including the followings:

2.3.1.1. Sanger DNA sequence analysis

The resulting DNA Sanger sequence chromatograms were viewed using the bioinformatics software Geneious version 7.1.4 and later (Kearse *et al.* 2012) on Ubuntu Linux 13.10. The high quality sequences were retained, while sequences with poor quality were removed. Then the DNA sequences were copied and saved in FASTA format.

For the plasmid DNA sequences, the pGEM®-T Easy vector sequences flanking the inserts were identified and deleted from the FASTA file by alignment the reverse and forward M13 primer with the sequence. The homology and differences between the sequences were conducted by aligning them. Multiple sequences were aligned by pairwise and multiple alignment for each gene region using the Geneious alignment algorithm, and always edited and improved manually by eye. Phylogenetic

reconstruction and estimation of nucleotide variability were carried out using Geneious or MEGA 6 program (Tamura *et al.* 2013).

2.3.1.2. BLAST and using NCBI web site

The BLAST (Basic Local Alignment Search Tool - Altschul *et al.* 1990) were used to query a sequence database with a previous studied data base or GenBank database through NCBI (National Centre for Biotechnology Information's) website, to find similarity hits for investigated sequence and to avoid analysing contaminated sequence. Generally, partial or complete matches were detected from the BLAST results matching along the entire length of an analysed sequence.

In current study the BLAST were done either directly from the Geneious program by selecting the query sequence and clicking on the sequence search button in the toolbar in the Geneious program, or by copy and paste the query sequence to NCBI web site (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch).

2.3.1.3. Dot-plot analysis

Geneious program dot-plot tools were used to compare two sequences against each other or one sequence against itself. By using this toll allows identifying and detecting sequence homologous or polymorphism and localizing the region of similarity in the sequence weather it is present from start to end or in a particular region of the sequence, identifying the tandemly repeated array in the same sequence, or observation of the reverse complement for nucleotide comparisons.

2.3.2.2. Setting paired reads

Because in the current study we were interested in analysing repetitive DNA sequences in the whole genomic sequences, so we chose to sequence our data by paired reds. Prior to assembly NGS raw reads, or implementing the genomic sequences to k-mer analysis or RepeatExplorer program the whole data subjected to paired reads (reverse/forward-illumina long read kit) by Geneious program. As the paired-end sequence is the advantage option to sequence whole genome sequences, allowing sequencing the same DNA fragment from both ends to generate high quality sequence

data. Paired-end sequences facilitates detection the repetitive sequence elements, genomic rearrangements such as insertions, deletions and inversions, and produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence.

2.3.2.3. De novo assembly

During the next generation sequencing procedures the whole genomic DNAs have been broken into thousands or millions of DNA fragments (150-300 bp in length in current study). Assembling procedure make all these fragments reassemble into a continuous sequence again. The DNA fragments overlaps after assembly, however, repetitive sequence characteristic can complicate this process. The result of sequence assembly is multiple contigs with different length, and the contigs consensus sequences extracted as the reconstructed sequence from numerous of overlapped DNA sequence reads.

2.3.2.4. Map to reference

It is the procedure of assembling a target sequence and makes it as a reference to the whole genomic sequences to find either the genomic copy number for that reference sequence or find similarity hits.

To assemble complete genome chloroplast, it is easier to assemble the whole genome sequences against known sequence, this procedure known as the map to reference. The result of the map to reference is producing just one contig per reference.

CHAPTER 3

Taxonomic, genomic and genetic diversity in apomictic *Taraxacum* and *Hieracium* (Asteraceae) agamospecies

Abstract

This study covers dandelions (*Taraxacum*) and hawkweeds (*Hieracium*), which reproduce asexually through apomixis. To study the structure, organization and relationship of different types of repetitive DNA sequences in the genome of these plant genomes, different methods were used to study repetitive DNA, such as dot blot hybridization of genomic libraries, amplification of the reverse transcriptase gene of transposable elements using universal primers, IRAPs and *in situ* hybridization. Repetitive sequences are evolving resulted from amplification of reverse transcriptase (RT) genes of different families of repetitive DNA, the IRAP shows some are active, and there are differences between the microspecies, although there are naturally similarities between the different microspecies. Number of polymorphic IRAP bands in *Taraxacum* agamospecies is high. The results from IRAP and whole genome sequences of nuclear gene of 45S rDNA agreed with O978 and S3 being more similar in contrast with A978. However, results from amplification of ITS region and non-coding chloroplast region for both genera resulted in pure polymorphism between agamospecies.

3.1. Introduction

Apomixis is widespread although rare in plants (<1% of species; Asker and Jerling 1992, Whitton *et al.* 2008. Carman (1997) listed more than 330 genera from a wide range of families including apomictic species (applying a broad species concept, and not accepting apomictic clones as species). Many taxa within the genera *Hieracium* and *Taraxacum* (tribe Cichoreae, Asteraceae) are apomictic, giving rise to seedlings without fertilization that are genetically identical clones of the mother, and there are sexual forms in many species (Nogler 1984; Asker and Jerling 1992; Mogie 1992; Koltunow 1993). The members of *Hieracium* form asexual seed by apospory (known in all genera where it occurs as "the *Hieracium*-type"; Nogler 1984; Asker and Jerling 1992) while *Taraxacum* asexual seed formation is through diplospory ("the *Taraxacum*-type"; Richards 1970; 1973; Mogie 1992; Asker and Jerling 1992). Some *Hieracium* and *Taraxacum* members are facultative apomicts where a plant can produce seeds derived from both sexual and apomictic processes. The genus *Taraxacum* Wigg. (Asteraceae) forms a polyploid complex within which there are strong links between the ploidy level and the mode of reproduction: diploids (2n=2x=16) are obligatory sexual, whereas polyploids, mainly triploids (2n = 3x = 24), are usually apomictic. In, *Hieracium* natural populations varies from diploid (2*n* = 2*x* = 18) to octoploid (2*n* = 8*x* = 72); the most frequent cytotypes are tetraploids, pentaploids and hexaploids (Chrtek *et al*. 2007; Mráz *et al.* 2011; Asker and Jerling 1992; Koltunow and Grossniklaus 2003; Suda *et al.* 2007).

The high level of morphological variation in both genera has been recognized since the earliest times (Dioscorides, 50 reprinted 1555). More recently, genetic variation has been found in apomictic populations (Baarlen *et al.* 2000). The agamospecies show extensive morphological variation, not dissimilar in amount and type to that in sexual species, and the source of this variation has been of wide interest. Mendel, following his work on peas, was unable to show genetic inheritance because of their apomictic nature (Iltis 1932); he notes that the two systems in *Pisum* and *Hieracium* were completely different.

Molecular markers have been applied in *Taraxacum* and *Hieracium* and in general discriminate agamospecies and show close genetic similarities (although not identity) within the agamospecies (Majesky *et al.* 2012; de Carvalho *et al.* 2016), although results do not normally give robust phylogenies. Kirschner *et al.* (2003) showed an overall lack of congruence in a comparison of parsimony analysis of *Taraxacum* morphological and chloroplast data, a conflict they suggest is a consequence of reticulate evolution. Large environmental effects (e.g. as in *Potentilla* shown by Clausen *et al.* 1940) on plant morphology have been reported (Asker and Jerling 1992; Van Dijk 2003), confounding genetic studies, and there is now interest in epigenetic induction and inheritance of variation (de Carvalho 2016). Transposon movement has been considered as inducing genetic variation (Kakutani *et al.* 1999; Springer *et al.* 2016), and is under environmental (stress) activation or control. Morphological and genetic variation seen in *Taraxacum* and *Hieracium* contrasts with another asexual triploid, *Crocus sativus*, where minimal genetic differences are seen (Alsayied *et al.* 2015). In triploid bananas, like the *Crocus* also vegetatively propagated, there are genetic differences between the many hundreds of sterile lines (Duren *et al.* 1996; D'Hont *et al.* 2012; Sardos *et al.* 2016), some most likely originating from independent hybridization events but others, such as those within the Cavendish group, being new 'mutations' within a clone (Heslop-Harrison and Schwarzacher 2007).

Plant nuclear DNA includes repetitive DNA motifs, the majority of most nuclear genomes, responsible for variation in genomes (Kubis *et al.* 1997). The most abundant repetitive DNA type, the transposable elements, are divided into retrotransposons flanked by long terminal repeats (LTR) and DNA transposons. Here, we aimed to survey the nature of retroelements in *Taraxacum* and *Hieracium* using universal primers (Konieczny *et al.* 1991; Friesen *et al.* 2001). LTR retrotransposon-based markers have been developed to give multi-locus fingerprints and measure diversity in plant genomes (Kalendar and Schulman 2006) where their mobility or amplification, abundance, ubiquity, and dispersion give extensive polymorphisms (Vicient *et al.* 2001; Alsayied *et al.* 2015; Saeidi *et al.* 2008). IRAPs (InterRetroelement Amplified Polymorphisms) exploit conserved retrotransposon features (long terminal repeat

region (LTR) or reverse transcriptase (RT) domains) for generating primers, and these are conserved across species.

3.2. Aims

This study will focus on using retrotransposon-based and other DNA markers to asses diversity in the tribe Cichorieae (Asteraceae), which contains the agamospermous genera *Taraxacum* and *Hieracium*. The pattern of genetic variation in agamospermous groups leads to understanding of the evolutionary nature, variations, and origin of their apomictic complexes.

We investigate the taxonomic status of agamospecies and analysing the nature of their relationships (reticulate vs divergent) conducted by study of some chloroplast region and nuclear DNA to investigate the relationship between both genera with agamospecies.

I aimed to characterise the retrotransposons of *Hieracium* microspecies by analysing the Ty3-gypsy, Ty1-copia, and LINE retrotransposon and by using PCR primers specific for conserved domains of RT genes of copia-, gypsy-like and LINE retroelements.

I also aimed to use IRAPs here to measure diversity and relationships between *Taraxacum* and *Hieracium* accessions and see how the pattern of genetic variation in agamospermous groups allows understanding of the evolutionary nature, variations, and origin of their apomictic complexes. Using in situ hybridization to chromosomes from triploid agamospecies, we investigated the locations of selected polymorphic IRAP bands to find the dispersion patterns of these sequences, and indicate the auto- or allo-polyploid nature, showing the taxonomic status of agamospecies and the nature of their relationships (reticulate vs divergent). We also also aimed to compare use of retroelement-based markers (IRAP, REMAP) with low-copy markers (ITS and chloroplast) and the AFLPs assayed by Majesky *et al.* (2012).

3.3. Materials and methods

3.3.1. Plant materials

Hieracium agamospecies (Table 2.1a) along with *Taraxacum* agamospecies (Table 2.1b) were used, for more detail and table of plant agamospecies used see section (2.1.1) of material and methods in chapter 2.

3.3.2. Genomic DNA isolation and quantification

DNA was isolated using CTAB method, for more details see (2.2.1 and 2.2.2) of material and methods in Chapter 2.

3.3.3. Amplification of DNA fragments

Different primer and primer combinations were used to amplify DNA from *Taraxacum* and *Hieracium* microspecies (Table 2.1) by PCR. For more detail about primer design, amplification of DNA fragments, gel electrophoresis, and purification of PCR products see section (2.2.3-2.2.6) of material and methods in chapter 2.

IRAP primer amplification was carried out as described in Alsayied *et al.* (2015), and the PCR reaction parameters consisted of: 95°C for 2 minute, followed by 30 cycles of (95°C for 1 minute, annealing at the temperature specified in Table 3.1 for 1 minute, ramp ramp +0.5°C to 72°C for 2 minute and adding +3 second per cycle), a final extension at 72°C for 10 minute.

PCR products were analysed by electrophoresis on 1% (w/v) agarose gel [for IRAP on 2% (w/v) agarose gels (preferably using a 1:3 mixture of high resolution : normal agarose)], and visualised by ethidium bromide, low molecular weight loading buffer (LB) 1x added to the PCR product and Hyper Ladder I (Bioline) added on both sides of the gel. The gel has been run for about 3-4 hours, to visualize separated bands clearly. Selected PCR bands for transposable elements and polymorphic bands for IRAP were excised, purified, and DNA fragments were used in probe labelling, cloned before sequencing or labelling, or directly sequenced as propriate.

Table 3.1. Nuclear gene, chloroplast, transposable element and outward facing retrotransposon (IRAP) primers used in amplification of genomic DNA of *Hieracium* and *Taraxacum*. Primer name, sequence, expected band sizes, annealing temperature, details of sequence/domain amplified, and literature reference to primer are shown. Primers are shown as pairs except for the IRAPs where both single primers and combinations (as indicated) were used for amplification.

Primer name	domain	Sequence (5' ➔ 3')	Annealing Tm (˚C)	Product Size (bp)	Reference
ITS4 ITS5	rDNA gene (ITS1-5.8S-ITS2)	5'-TCCTCCGCTTATTGATATGC 5'-GGAAGGAGAAGTCGTAACAAGG	50	800	White *et al.* 1990
3F-Kim 1R-Kim	Chloroplast *mat*K gene	5'-CGTACAGTACTTTTGTGTTTACGAG 5'-ACCCAGTCCATCTGGAAATCTTGGTTC	50	900	Sang *et al.* 1997
1F 724R	Chloroplast *rbc*L gene	5'-ATGTCACCACAAACAGAAAC 5'-TCGCATGTACCTGCAGTAGC	50	780	
trnH psbA	spacer region between gene *trn*H and *psb*A	5'-ACTGCCTTGATCCACTTGGC 5'-CGAAGCTCCATCTACAAATGG	50	450	
Ty-1-1	TAFLHG (F)/ Psedoviridae/ Ty1-copia	5'ACNGCNTTYYTNCAYGG	42	270	Flavell *et al.* 1992
Ty-1-2	YVDDML (R)/Psedoviridae/ Ty1-copia	5'ARCATRTCRTCNACRTA			
GyRT-1	RMCVDYR (F)/ Metaviridae/ Ty3-gypsy	5' MRNATGTGYGTNGAYTAYMG	55-56	420	Kubis *et al.* 1998; Friesen *et al.* 2001
GyRT-3	LSGYHQI (F)/ Metaviridae/ Ty3-gypsy	5' YKNWSNGGNTAYCAYCARAT		300	
GyRT-4	YAKLSKC (R)/ Metaviridae/ Ty3-gypsy	5' RCAYTTNSWNARYTTNGCR		-	
BEL-1MF	[E/D/K/S] [E/D/N]/LINE/	5'-RVNRANTTYCGNCCNATHAG	42 (*Taraxacum* 48)	410	Kubis *et al.* 1998
BEL-2MR	RQGDPLS/LINE	5'-GACARRGGRTCCCCCTGNCK			
LTR6150	BARE-1 ⬅	CTGGTTCGGCCCATGTCTATGTATC CACACATGGTA	40 (+LTR-6149 =48)	-	Kalendar *et al.* 1999
LTR6149	BARE-1 ➡	CTCGCTCGCCCACTACATCAACCGC GTTTATT	40 (+NIKITA =50)	-	Kalendar *et al.* 1999
Nikita	Nikita ➡	CGCATTTGTTCAAGCCTAAACC	45 (+LTR-6150 =50)	-	Leigh *et al.* 2003
Sukkula	Sukkula ➡	GATAGGGTCGCATCTTGGGCGTGA C	50 (+NIKITA=47)	-	Manninen *et al.* 2000
Ty1	Wl, W3, W7, W8 ⬅	CCYTGNAYYAANGCNGT	48 (+Ty-2=50)	-	Teo *et al.* 2005
TY2	W1, W3, W7, W8 ➡	TRGTARAGRAGNTGRAT	48	-	Teo *et al.* 2005
3' LTR	BARE-1 ➡	TGTTTCCCATGCGACGTTCCCCAACA	48	-	Teo *et al.* 2005
5' LTR1	BARE-1 ⬅	TTGCCTCTAGGGCATATTTCCAACA	42 (+NIKITA=49)	-	Teo *et al.* 2005

Y=C+T; **R**=A+G; **M**=A+C; **K**=G+T; **S**=G+C; **W**=A+T; **H**=A+T+C; **D**=G+A+T; **B**=G+T+C; **V**=G+A+C; **N**=A+G+C+T

3.3.4. Dot blot and Southern hybridization

A genomic library was constructed to identify repetitive DNA sequences from partial digests of genomic DNA from *Taraxacum sp.* and *Hieracium northroense*. The digestions were performed with *Hae*III, *Hin*dIII, *Bam*HI, *Sau*3A I*, Dra*I and *Eco*RI restriction enzymes (New England BioLabs) in the presence of appropriate buffers following manufacturer's instructions (Table 3.2).

For more details about cloning the fragment DNA to pGEM-T Easy Vector (promega), preparing cloning library, transformation of cloning library to positive charged membrane, and finally southern hybridizations of clones with labelled genomic DNA of *Hieracium* and *Taraxacum*, see materials and methods chapter (Chapter 2) section 2.2.7 and 2.29.

Table 3.2. Shows different restriction enzymes with their recognition sequence site, optimal buffer and optimal incubation temperature.

	Enzyme name	Recognition sequence site	Company	Optimal buffer	Buffer company	Optimal incubation Temperature (˚C)	Inactivation temperature (˚C)
1	BamHI	5'...G↓GATCC ...3' 3'...CCTAG↑G ...5'	NEB	2, 3, 4 + BSA	NEB	37	No
2	EcoRI	5'...G↓AATTC ...3' 3'...CTTAA↑G ...5'	NEB	1	NEB	37	65
3	HaeIII	5'...GG↓CC ...3' 3'...CC↑GG ...5'	NEB	4	NEB	37	80
4	HindIII	5'...A↓AGCTT ...3' 3'...TTCGA↑A ...5'	NEB	2	NEB	37	80
5	Sau3AI	5'...↓GATC ...3' 3'...CTAG↑ ...5'	NEB	1+BSA	NEB	37	65
6	DraI	5'...T T T↓A A A...3' 3'...A A A↑T T T...5'	Promega	B, 4	Promega NEB	37	65

3.3.5. Analyzing of DNA sequencing and phylogenetic constructions

Sequencing was performed commercially with one direction for primer of chloroplast, nuclear, M13, and degenerated retrotransposon primers.

Geneious version 7.1.9 and later (Kearse *et al.* 2012) program was used to perform preliminary phylogenetic analyses. Then phylogenetic analyses were

performed through maximum likelihood (ML) using Mega6 program (Tamura *et al.* 2013). For ML analysis, a nucleotide substitution model was selected.

3.3.6. Analysis of IRAP diversity

Fingerprints were scored to prepare binary matrices (Kalendar and Schulman 2006) of presence (1) or absence (0) of clear and distinguishable fragments of particular mobility IRAP primer and accession assuming that each band represents a single locus. A dendrograms constructed using the UPGMA method (Saitou and Nei 1987) showing relationships among *Taraxacum* accessions with 1000 bootstrap replicates using PowerMarker. The consensus bootstrap tree was generated using Geneious version 7.1.9.

3.4. Results

3.4.1. Relationship among *Taraxacum* and Hieracium agamospecies by using chloroplast and nuclear PCR marker

Amplification of nuclear ITS, chloroplast coding *mat*K, *rbc*L and noncoding *trn*H-*psb*A sequences gave a single band in each accession (18 asexual agamospecies of *Hieracium* and *Taraxacum* and one sexual species of each genus). The chloroplast genome of *Taraxacum* (Salih *et al.* 2017) was used as a reference for the intergenic spacer region between *trn*H and *psb*A (381-403 bp in *Taraxacum*, 358-397 bp in *Hieracium*). PCR amplification of plastid noncoding intergenic spacer between *trn*H and *psb*A and ITS showed poor resolution and the plastid coding *mat*K and *rbc*L region showed a lack of polymorphism between all studied agamospecies sequences in both genera so these were not analysed further. As reference for the fragments ITS1-5.8S-ITS2 (710-756 bp in *Taraxacum*, 500-820 bp in *Hieracium*), the whole 45S rDNA region (7,720 bp long; 18S-ITS1-5.8S-ITS2-26S) was assembled from whole genome sequences of three *Taraxacum* genomes (see Salih *et al.* 2017), showing three transition/transversions (one in the 5.8S rRNA and two in 26S rRNA) and SNPs (T in S3 genome against C in

A978 and O978 genome, and C, G in S3 and O978 genome against G, A in A978 genome) (submit seq to GenBank accession numbers XY123456).

After alignment (Geneious optimized by hand), phylogenetic trees were constructed using maximum likelihood in MEGA6. K2+G (Kimura 2-parameter and +Gamma distribution) using nucleotide substitution models, for ITS sequences for both genera; for intergenic spacers of *tran*H-*psb*A, T92 (Tamura 3-parameter) and T92 +G were the best DNA model for *Hieracium* and *Taraxacum* agamospecies respectively. The outgroup species *Lactuca sativa* (AJ633337 nuclear ITS, NC007578 *trn*H-*psb*A) and *Cichorium intybus* (AJ633451 nuclear ITS, HQ596644 *trn*H-*psb*A) were well resolved from the target genera (Figure 3.1, 2). Within the two genera, bootstrap values were low; most species and accessions were not resolved.

3.4.2. Isolation, amplification and identification of repetitive DNA

Genomic DNA digests (10 µg) from *Taraxacum* and *Hieracium* agamospecies with a total of six restriction enzymes showed few clear bands representing restriction satellites except for weak bands in HaeIII and DraI digests (Figure 3.3). After cloning the bands, the resultant colonies were hybridised with labelled total genomic DNA of *Hieracium northroense* and *Taraxacum.* From 100 colonies from *H. northroense* and 50 colonies from *T.* sp., 23 showed strong hybridization, and these clones were sequenced. None of the sequences nor gels showed evidence for tandem repeated DNA (ladders of obvious multi-mers), and there were no notable homologies with the NCBI database.

Major families of retrotransposons were amplifies and isolated using degenerated proimers designed to amplify the conserved domain of the reverse transcriptase (RT) gene (Table 3.1); sequence comparison confirmed most products were retroelement-related. Along with the strong band from an internal retroelement domain of 270bp (Figure 3.4 a), genomic DNA from *Hieracium* agamospecies gave multiple larger bands and retroelement fragments of 270bp and 420bp were cloned and sequenced from *H. hethlandiae* (H9) and *H. lissolepium* (H10).

Figure 3.1: Phylogenetic trees derived from maximum likelihood analysis of alignments of DNA sequences of 19 different *Hieracium* (a) and *Taraxacum* (b) agamospecies of an ITS nuclear region. Numbers above node are bootstrap support values.

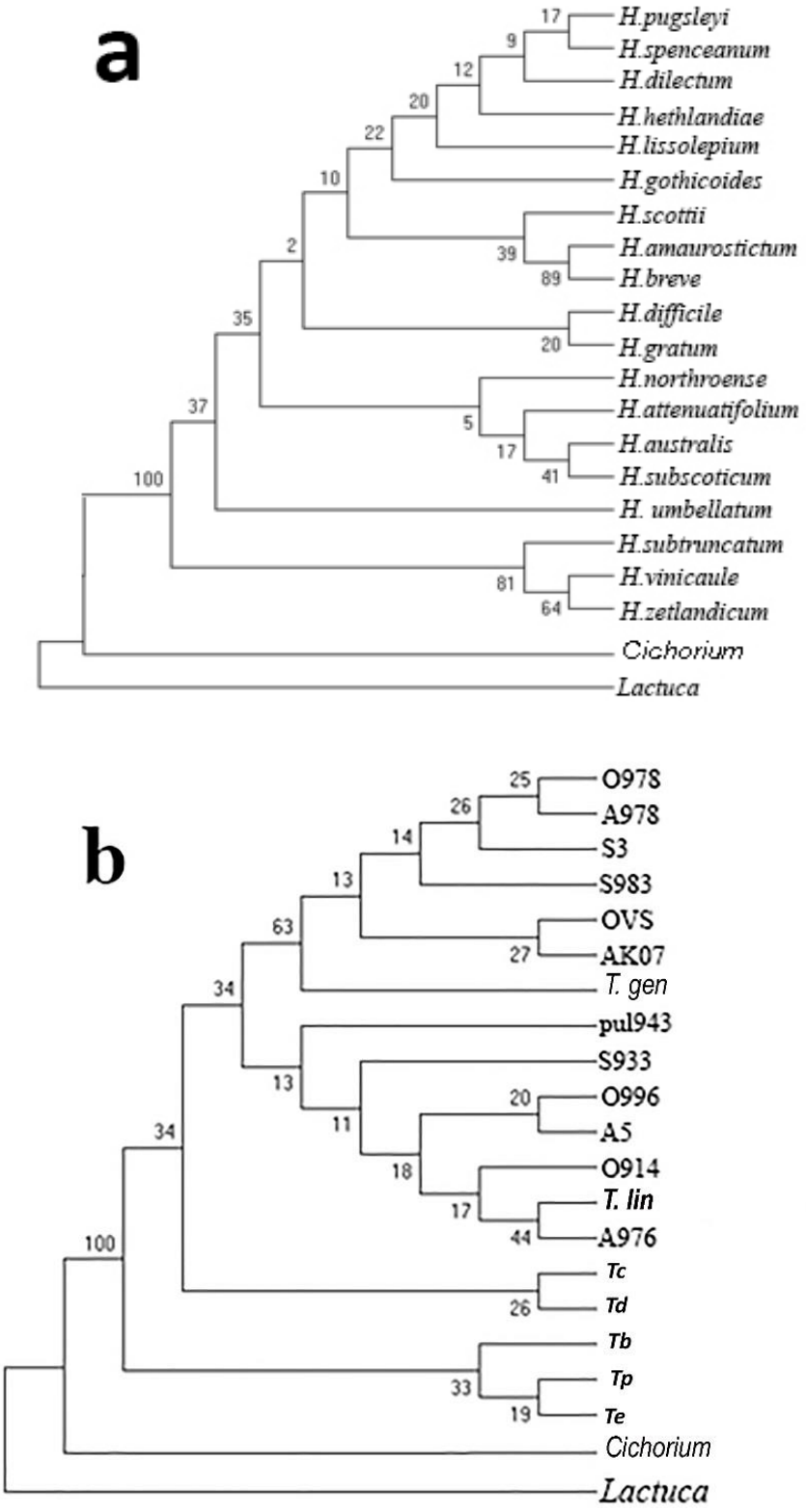
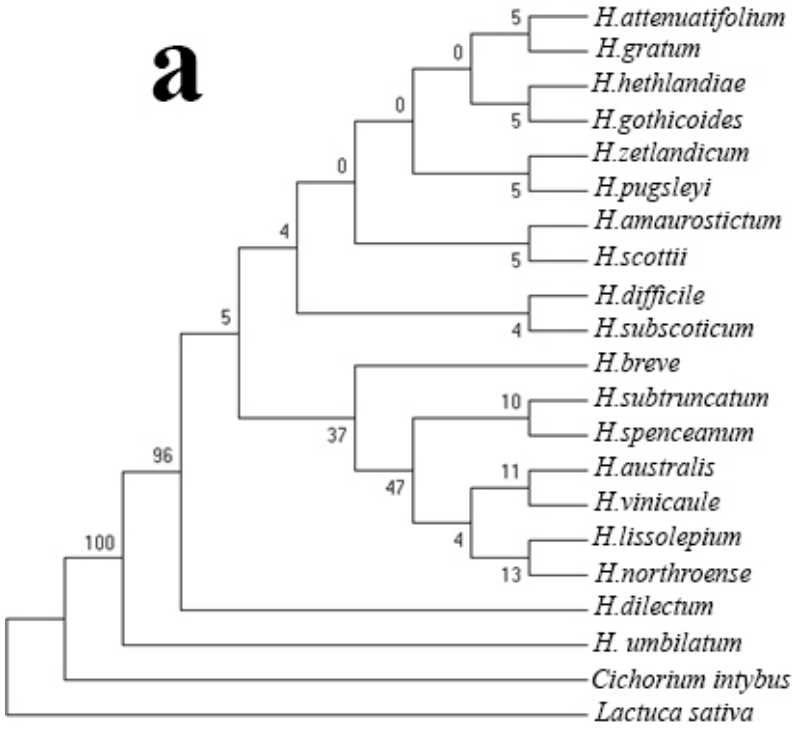
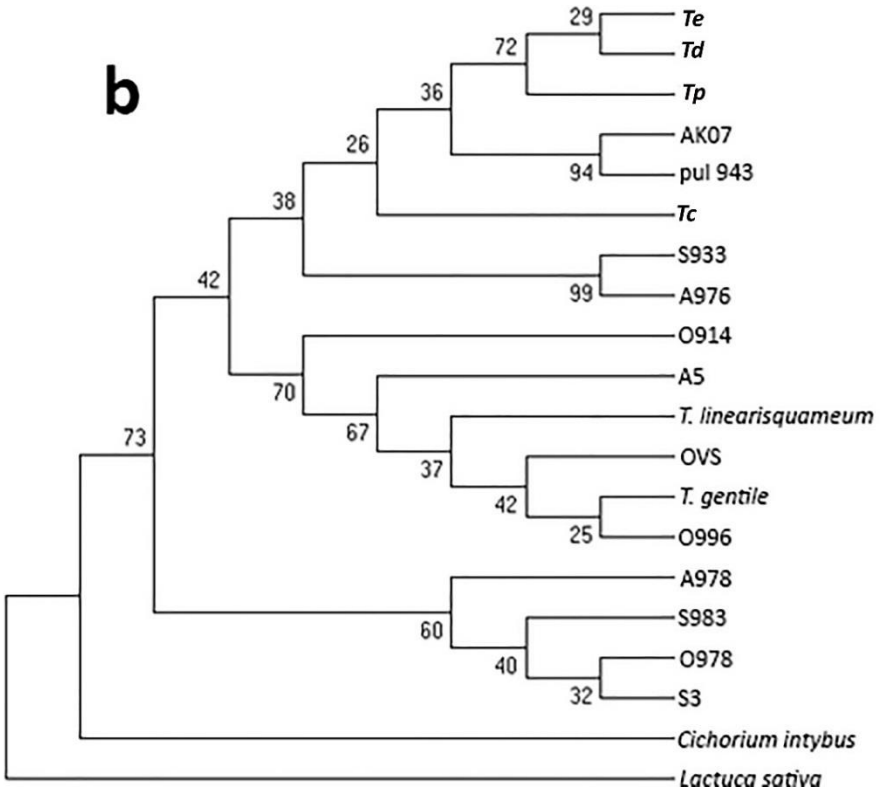


Figure 3.2: Phylogenetic trees derived from maximum likelihood analysis of alignments of DNA sequences of 19 different *Hieracium* (a) and *Taraxacum* (b) agamospecies of the chloroplast inter genic spacer region between *trn*H and *psb*A. Numbers above node are bootstrap support values.





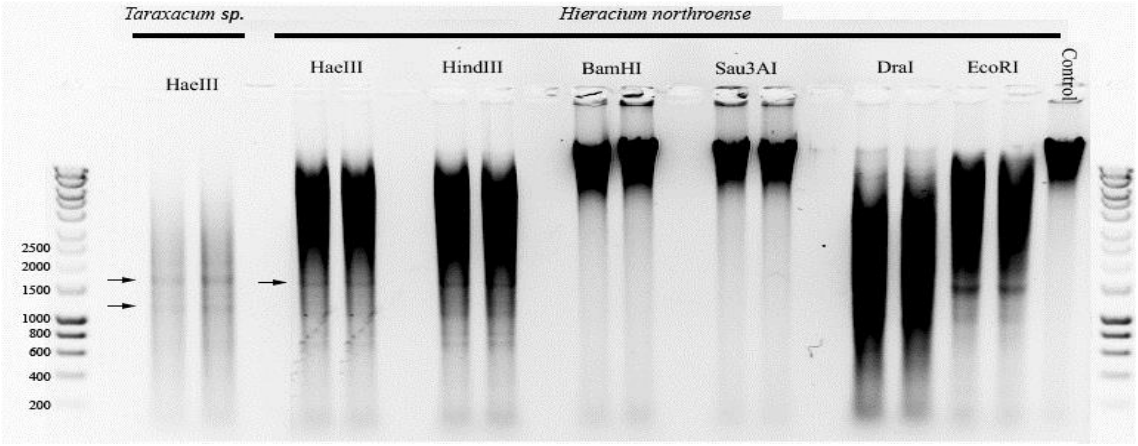


Figure 3.3: Ethidium bromide stained gel of restriction enzyme digestions of genomic DNA from *Taraxacum sp*. and *Hieracium northroense*. Tracks from left to right: Bioline Hyperladder I; HaeIII on *Taraxacum spp.*; (HaeIII; HindIII; BamHI; Sau3AI; DraI; EcoRI) on *Hieracium northroense*; control DNA without enzyme; Bioline Hyperladder I. Black arrows represent the separated clear band in the gel.

Reverse transcriptase (RT) gene fragments were amplified from LINE retrotransposons in *Hieracium* agamospecies DNAs. The sequence of the major 600bp fragment (Figure 3.4b) showed homology to a retrotransposon protein. Isolation of Ty3-gypsy-like retrotransposons from *Hieracium* agamospecies by amplification of genomic DNA with primer pairs GyRT1 and GyRT4 generated the expected fragment of about 420 bp (Figure 3.4 c). GyRT3 and GyRT4 amplified a Ty3-gypsy-like RT gene, giving the expected band size at 300bp and bands at 400, 600, 800bp (Figure 3.4d); most sequences showed homology to *Helianthus* Ty3-gypsy-like RT genes, although some had no similarity in GenBank.

Several clones were chosen randomly to sequence with M13 primer from different degenerated primer amplifications. For examination of the evolution and phylogenetic relations between all groups of retroelements, an unrooted phylogenetic analysis was carried out Maximum Likelihood methods conducted in MEGA6. General Time Reversible (GTR) along with Gamma distribution (+G) were considered as models with the lowest BIC scores (Bayesian Information Criterion) to describe the substitution pattern the best, and were used for the trees. The tree was build up by using 24 Ty1-copia-like, 37 Ty3-gypsy-like and 12 LINEs from *Hieracium* agamospecies. The

retrotransposons from the related elements clustered, supporting the monophyletic origin of the copia, LINE, and most gypsy clades (Figure 3.5), but there were no notable species-specific clades.

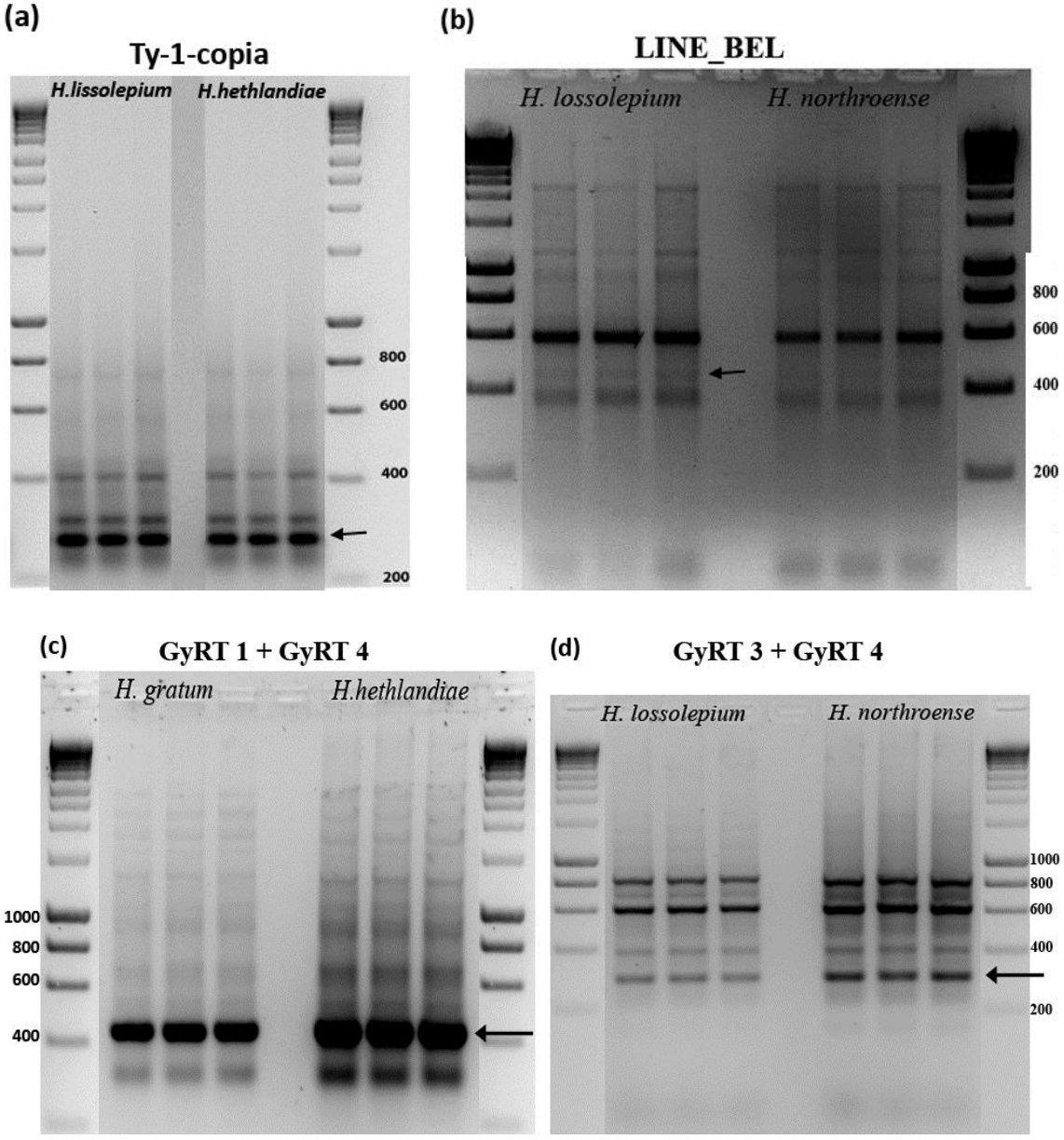


Figure 3.4: Detecting of (a) Ty1-copia, (b) LINE, (c) GyRT 1+4, and (d) GyRT 3+4 retrotransposons by amplification of reverse transcriptase-coding domain by PCR. Total genomic DNA from Hieracium agamospecies was subjected to PCR-amplification and products were separated on 1.5% agarose gel. An arrowhead indicates DNA fragments of expected band size.

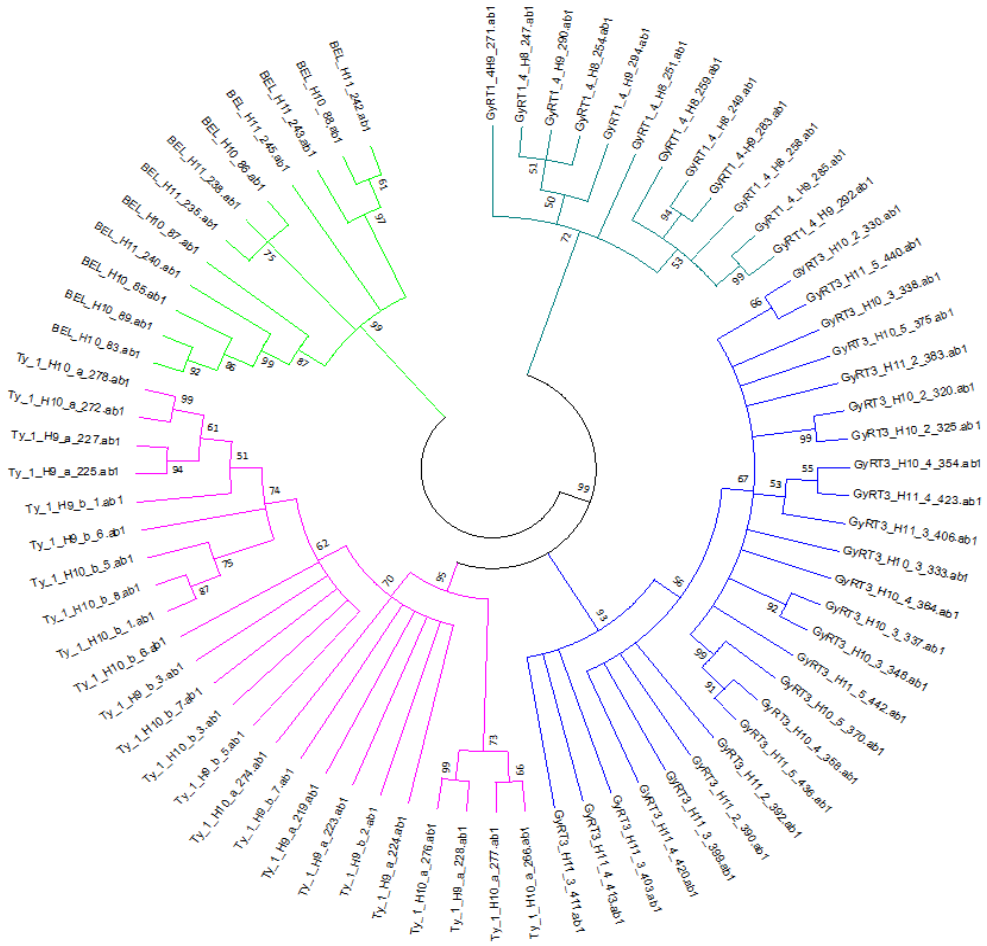
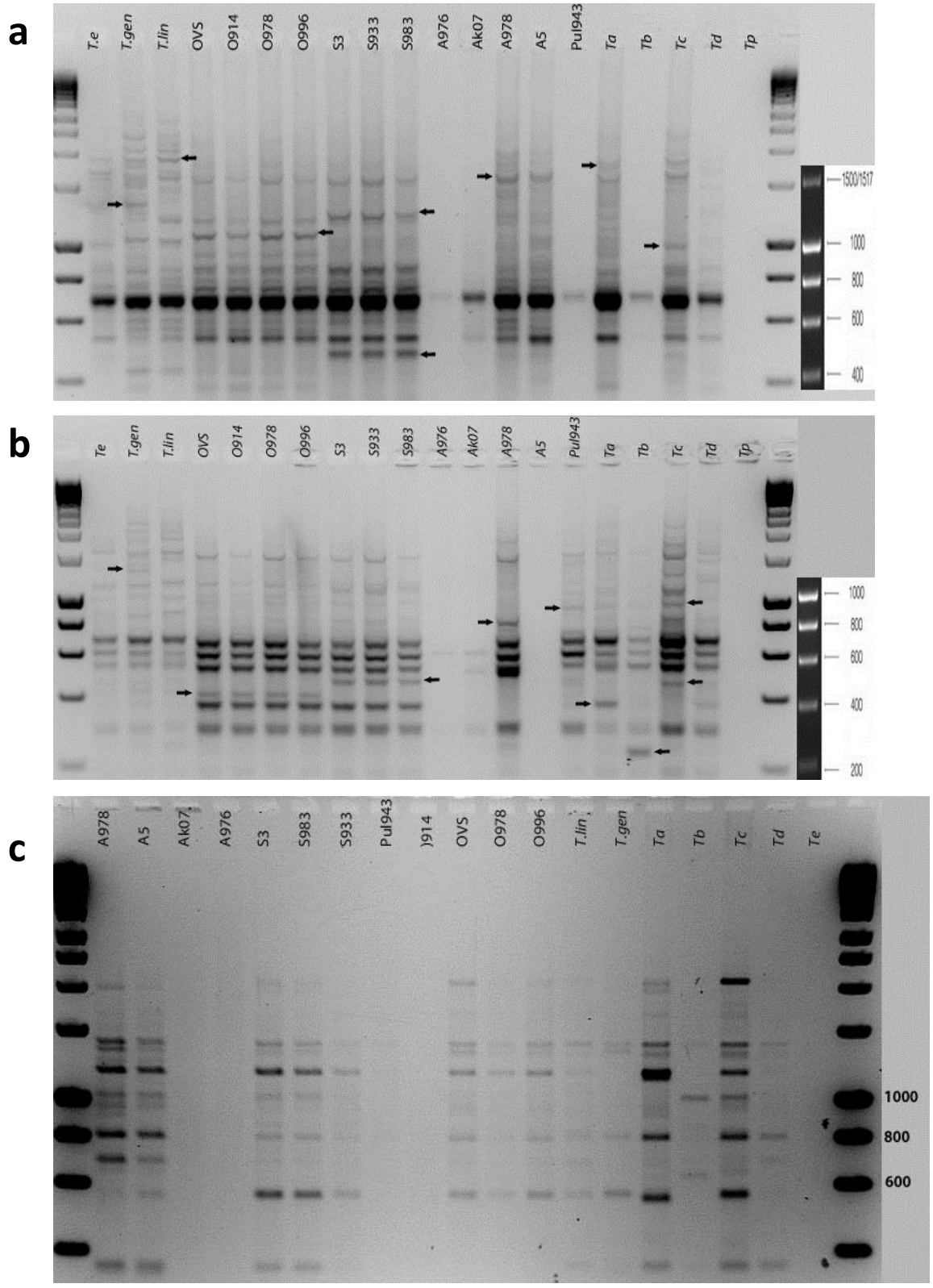


Figure 3.5: Unrooted bootstrap tree represent phylogenetic analysis of reverse transcriptase sequences of retrotransposons. The dendogram representing the relative similarities between the sequences was made with Maximum Likelihood. The number on the branches indicate the supporting bootstrap values.

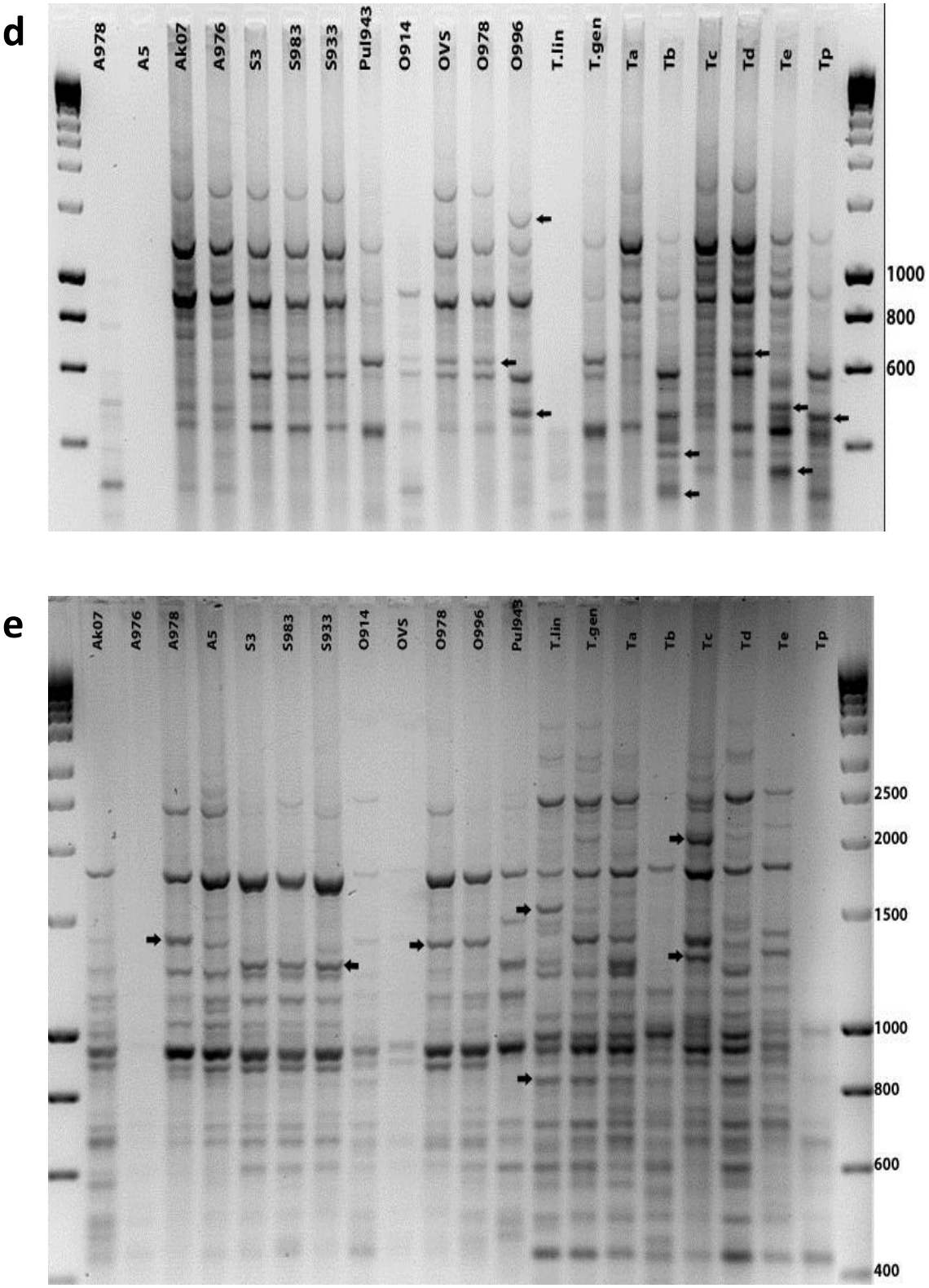
3.4.3. IRAP amplification and diversity within *Taraxacum* microspecies

IRAP markers were used to find whether transposon-based markers were able to classify variability of *Taraxacum* agamospecies. With 20 microspecies of *Taraxacum* and 14 primer combinations (Table 3.1), banding profiles yielded distinct and polymorphic fingerprints with bands ranging from 150 bp (SUKULA+NIKITA) up to 5,000 bp (SUKULA), (Figure 3.6, 7). The binary IRAP fragment scores were used for the reconstruction of distances between agamospecies by UPGMA, giving eight well-supported clusters (bootstrap value >90%), with major taxa (agamospecies) collected in Europe grouping clearly; UK and other species did not group with the European accessions (Figure 3.8).

Figure 3.6: IRAP amplification from 20 *Taraxacum* accessions. (a) Nikita, **(b)** NIKITA and 5'LTR, **(c)** LTR 6149, **(d)** LTR 6150, **(e)** SUKULA, **(f)** Combination between LTR6149 and LTR 6150, **(g)** combination between SUKULA and NIKITA. Arrow heads indicate the polymorphic bands.



← Figure 3.6 continue…



← Figure 3.6 continue ….

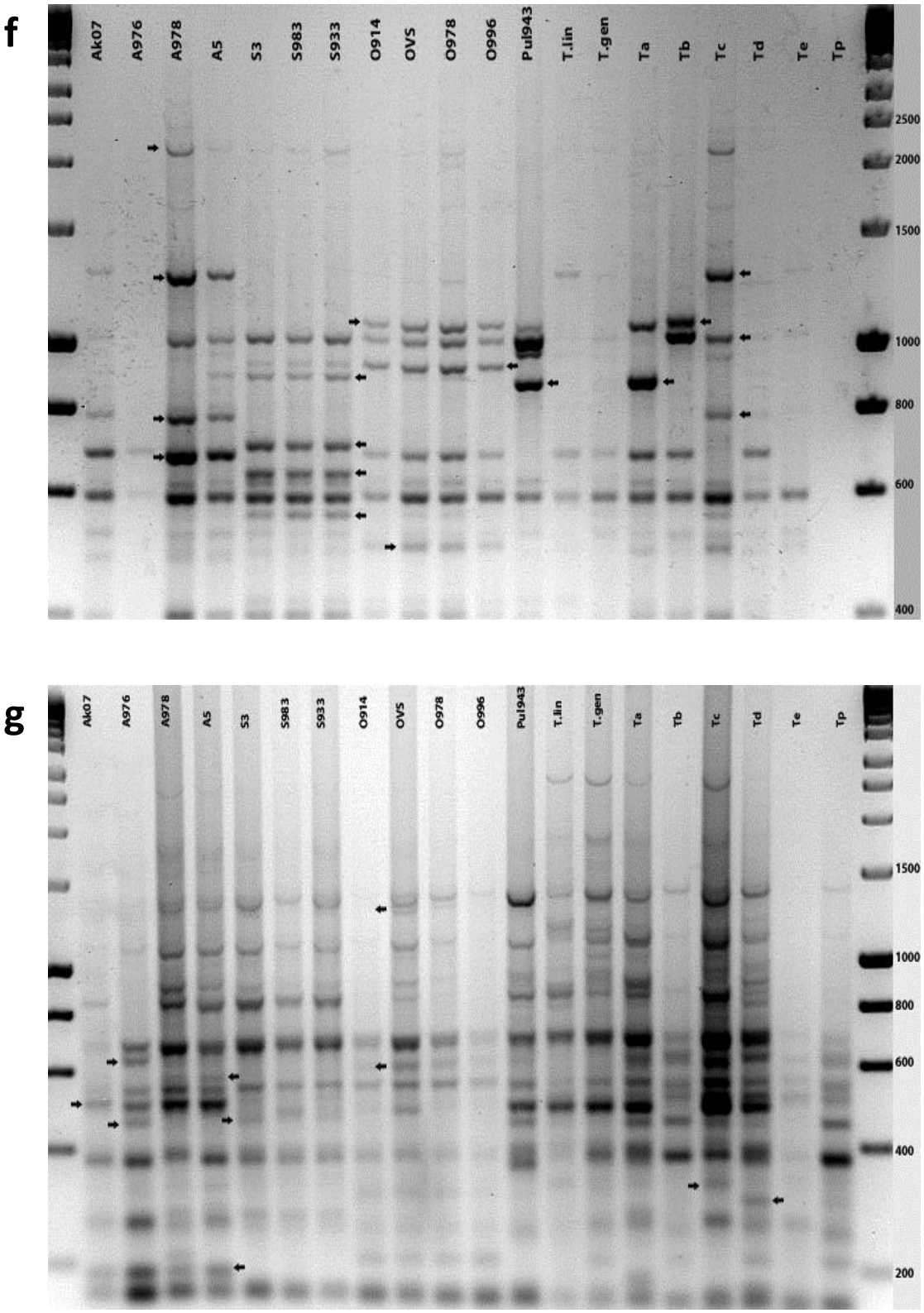


Figure 3.7. IRAP pattern from multiple *Taraxacum* accessions show no variation was evident.

The primer used were **(a)** 3'LTR , **(b)** RTY-1, **(c)** combination of RTY-1 and RTY-2.

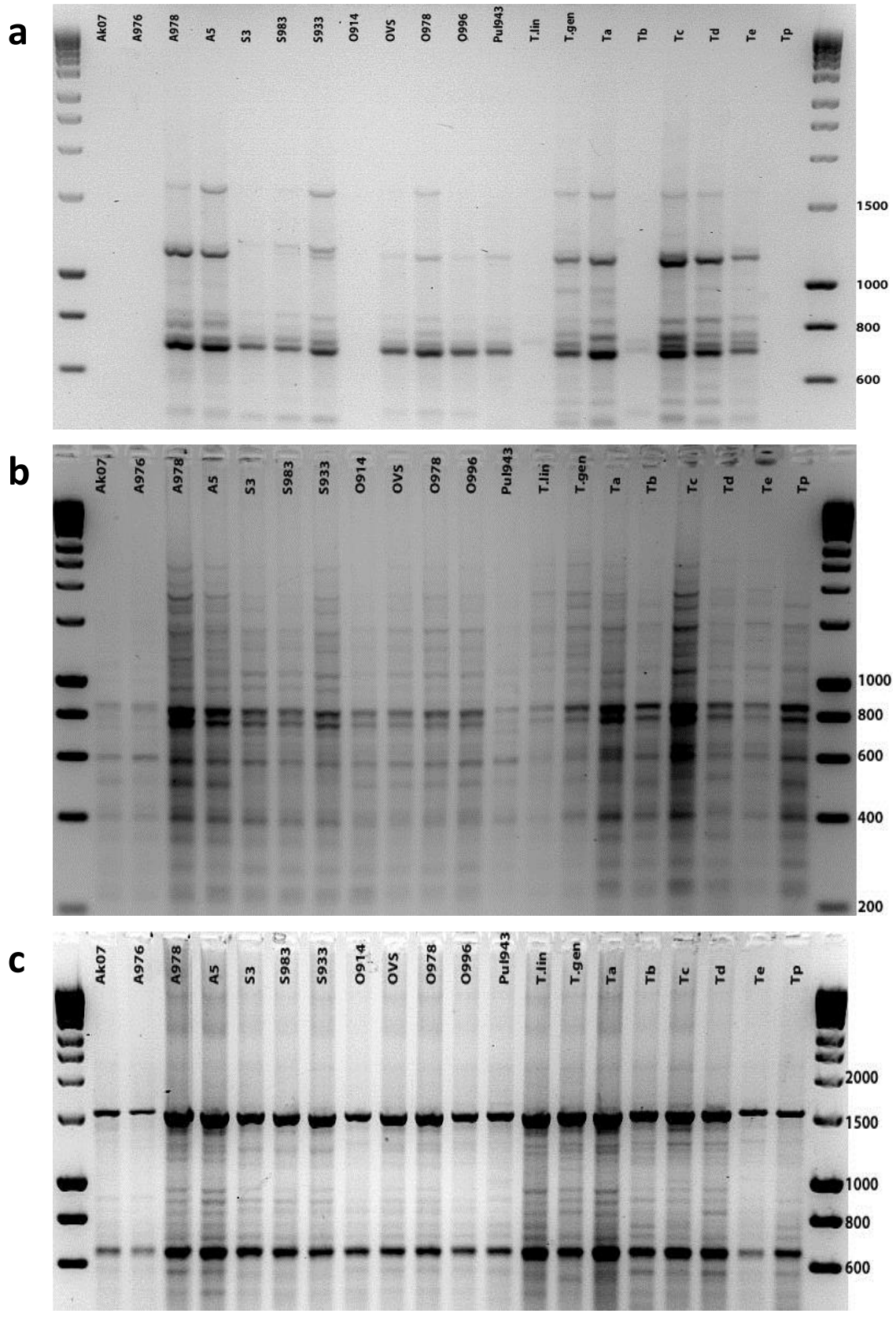
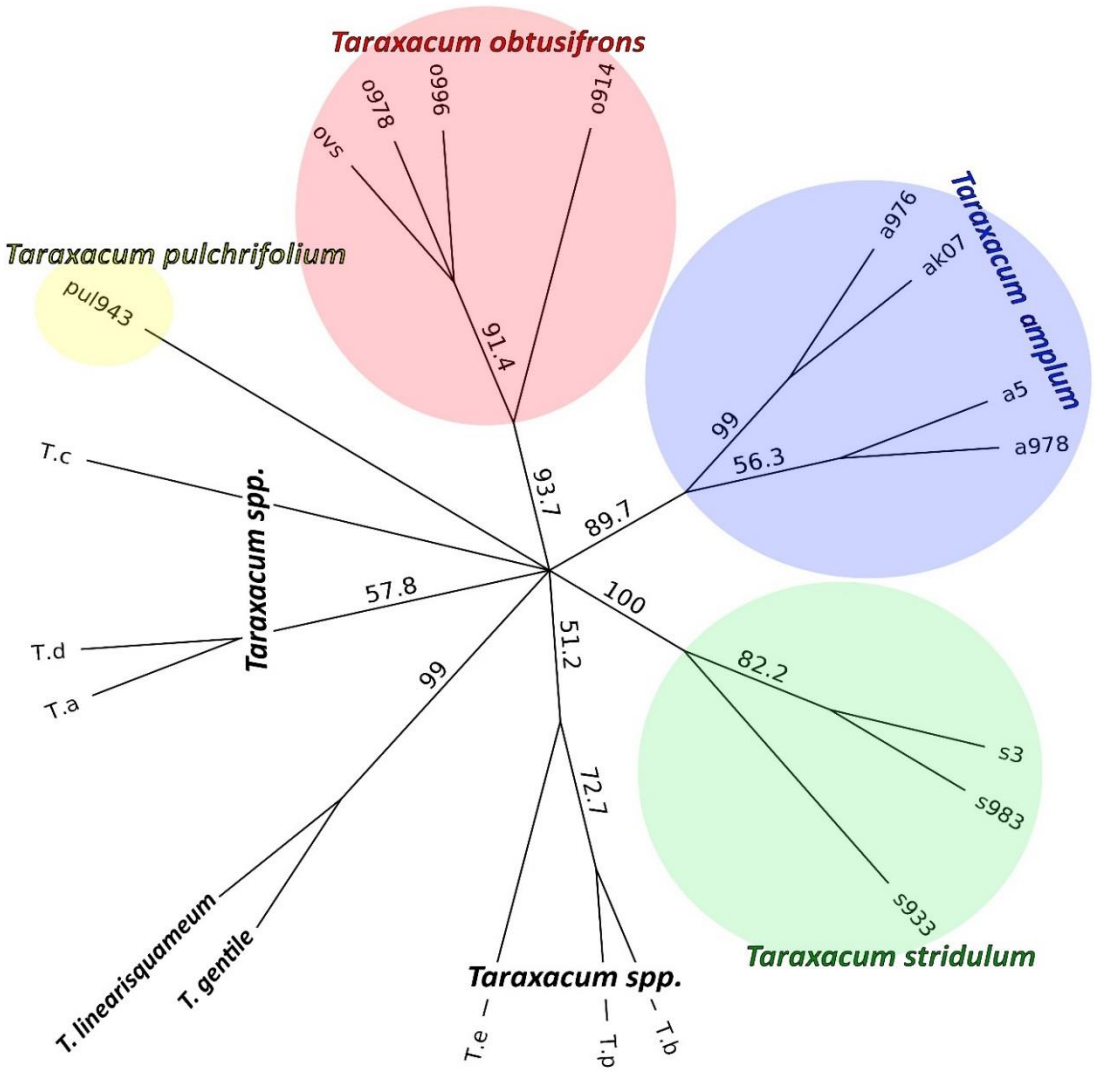


Figure 3.8: Consensus UPGMA dendrogram generated using IRAP data for 20 *Taraxacum* agamospecies. The bootstrap consensus (% shown at notes) is inferred from 1000 replicates computed by PowerMarker software (percentage support shown at nodes). Accession labels as Table 2.1 in materials and methods chapter.



3.4.4. Chromosomes and *in situ* hybridization

The chromosomes of *Hieraciumn vinicaule* (Figure 3.9 A–C) and *Hieracium northroense* (Figure 3.9 L - 2n=3x=27), 5.2-8.6 µm long, are all similar sub-metacentrics, with no conspicuous morphological features other than the secondary constriction at the NORs.

The DNA of the genomic clones HH40 of HaeIII fragment and HD14 of DraI fragment both clones from *H. northroense*, and TH10 from *Taraxacum spp.* (HaeIII) were sequenced well. These clones were used for *in situ* hybridization to chromosome preparations. HH40 on *H. northroense* (2n=3x=27) dispersed over most of the chromosomes with some exceptions of centromeric gaps and absent signals on some arms can be seen (Figure 3.9 A). Each of HD14 and TH10 probes showed widespread hybridization on *Hieracium* chromosomes (Figure 3.9 B, C).

The chromosomes of *Taraxacum* are all sub-metacentric and between 3.2 and 5.1 µm long at metaphase (Figure 3.9 D-I). In the triploids (2n=3x=24), no individual chromosome types could be unequivocally identified apart from the NOR chromosome although some were larger and there was some variation in centromere index. Phase contrast microscopy (not shown) and DAPI staining did not reveal any conspicuous domains of heterochromatin (e.g. at centromeres or telomeres) in either metaphases or interphases.

DNA from some polymorphic bands from IRAP analysis of *Taraxacum* was used for *in situ* hybridization to find the abundance and distribution of the PCR products. Most were abundant and widespread on many chromosomes (Figure 3.9 D-I) but each showed unique hybridization patterns. Typically, probes hybridized to broad (Figure 3.9 D) or narrower (Figure 3.9 E, F) bands or smaller dots at centromeres of almost all chromosomes (Figure 3.9 G, I) with signals absent on some chromosomes (Figure 9.3 E). The probe in Figure 3.9 H showed more widespread labelling all over the chromosomes.

Genomic *in situ* hybridization (GISH) with labelled DNA of *H. umbellatum* (2n=2x=18) gives differential staining of the chromosomes, including dispersed signal on three whole chromosome triplets, and whole arms of two triplets, one very weakly labelled triplet, and stronger centromeric bands on three smaller chromosomes (Figure 3.9 J). *Hieracium* chromosomes had two sites of 5S rDNA and one site of 45S rDNA per haploid genome. 45S rDNA localized at the end of the short arm of chromosome with a sub-terminal 5S rDNA site; the other 5S site was also sub-terminal (Figure 3.9 K, L).

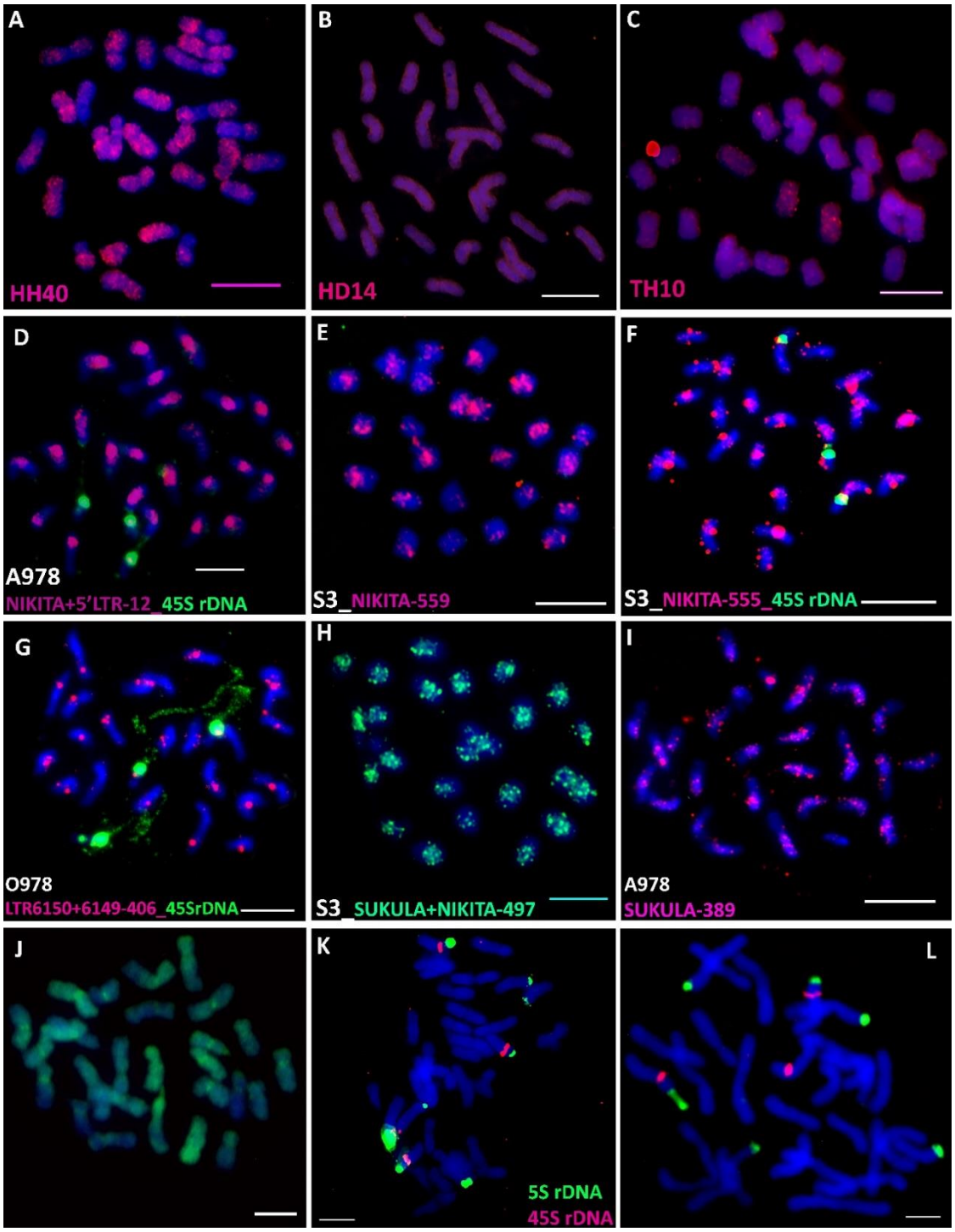


Figure 3.9: Mitotic metaphase spreads of *Taraxacum* (2n = 24) and *Hieracium* agamospecies after FISH and GISH. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. (A, B, C) root-tip metaphase chromosomes of *H. northroense* (2n=3x=27) hybridized with probes from DNA of different colonies of dot blot hybridizations from *Hieracium* and *Taraxacum;* (D-I) *Taraxacum* chromosomes hybridized with probes labelled from extracted polymorphic bands from IRAP of different primer combinations, primer name and metaphase chromosome genomes written beside the picture; (J) GISH from *H. umbellatum* genomic DNA labelled with Digoxigenin-11-DUTP hybridized on *H. northroense* (2n=3x=27); (K) root-tip metaphase chromosomes of *H. dilectum* (2n=4x=36); (L) root-tip metaphase chromosomes of *H. vinicaule* (2n=3x=27).

3.5. Discussion

Although agamospecies or microspecies are recognizable and reasonably well-defined taxonomically within both the *Taraxacum* and *Hieracium* genera, the phylogenies generated using either a nuclear ITS marker nor the chloroplast trnH and psbA markers (Figure 3.1) showed minimal resolution, and sequence data would not enable confident species identification. Chloroplast markers and nuclear used by others have also shown only limited phylogenetic signal in these genera, with some rare alleles and other alleles being shared among distantly related taxa (even between sexual and apomictic accessions, Wittzell *et al.* 1999; Van der Hulst *et al.* 2000; Mes *et al.* 2000; 2002; Kirschner *et al.* 2003, Majeský *et al.* 2012; 2015). Kirschner *et al.* (2003) showed that sequence data from commonly used barcode primers was not congruent with morphological analysis, perhaps a consequence of reticulate evolution. We have suggested that whole chloroplast genome sequences may be valuable to define unequivocally the maternal lineages within hybridizing and apomictic genera such as Taraxacum (Salih *et al.* 2017).

The current study presents the 45S rRNA of three *Taraxacum* agamospecies: while fragments of ITS may not be informative for phylogeny, the data show O978 and S3 genome are more similar than the diverged A978, so longer DNA stretches may be more amenable to analysis, although the effort for a population-level analysis would be considerable.

The results of this chapter show that there are many repetitive elements within the *Taraxacum* and *Hieracium* genera resulted in the variation within and between agamospecies. Retroelements and their derivatives are an abundant component of the genomes and all major families could be isolated using degenerate primers (Figure 3.4), as in other species (Flavell *et al.* 1992; Voytas *et al.* 1992; Hirochika and Hirochika 1993; Matsuoka and Tsunewaki 1996, 1999; Friesen *et al.* 2001; Kubis *et al.* 2003; Hill *et al.* 2005). Genomic DNA digestions and visual examination of the chromosomes did not suggest presence of major tandemly repetitive DNA families apart from the 45S and 5S rDNA (Figure 3.3 and 3.9 a, b, c).

In contrast to the low-copy markers, DNA markers based on multiple sequence sites have proved able to discriminate *Taraxacum* species: Van der Hulst et al. (2000) and Majesky *et al.* (2012) used AFLP markers in the latter case along with multi-locus SSR polymorphisms. IRAP markers have been used to study multiple polymorphic retrotransposon insertion sites to measure genomic diversity and relationships (Nair *et al.* 2005; Saeidi *et al.* 2008; Kalender *et al.* 2011; Alsayied *et al.* 2015). Vukich *et al.* (2009) used IRAPs to analyse intraspecific variability based on retrotransposon sequences among wild and cultivated accessions in *Helianthus*, reporting that the high polymorphism of IRAP bands indicates activity of retrotransposons after speciation. Here, we found many polymorphic IRAP bands in *Taraxacum* agamospecies, suggesting that retrotransposons are active in *Taraxacum*. The tree discriminated agamospecies and undefined members of the *T. officinale* aggregate (Figure 3.8), showing that retroelement-based markers are valuable for the study of diversity and for discriminating and grouping the different accessions of *Taraxacum*. Notably, the major sub-species collected in Europe were grouped exactly in agreement with the AFLP-based classification of Majeský *et al.* (2012).

IRAP can be amplified with a single primer matching either the 5´or 3´ end of the LTR but oriented away from the LTR itself, or with two outward-facing primers (Figure 3.10). The two primers may be from the same retrotransposon element family or may be from different families. IRAP was analysed on 20 microspecies of *Taraxacum*. The banding profiles yielded a considerable number of distinct and polymorphic fingerprints with bands ranging from 150 bp (SUKULA+NIKITA) up to 5,000 bp (SUKULA), (Figure 3.5).

Hieracium umbellatum is relatively distantly related sexual species to *H. northroense*; a closely related species would normally show more uniform GISH signal along chromosomes. The chromosomes grouping into three are consistent with *H. northroense* being autotriploid (Figure 3.9 J). The IRAP probes show no ‘genome-specificity’; primers give each though a characteristic location (Figure 3.9 D-I).

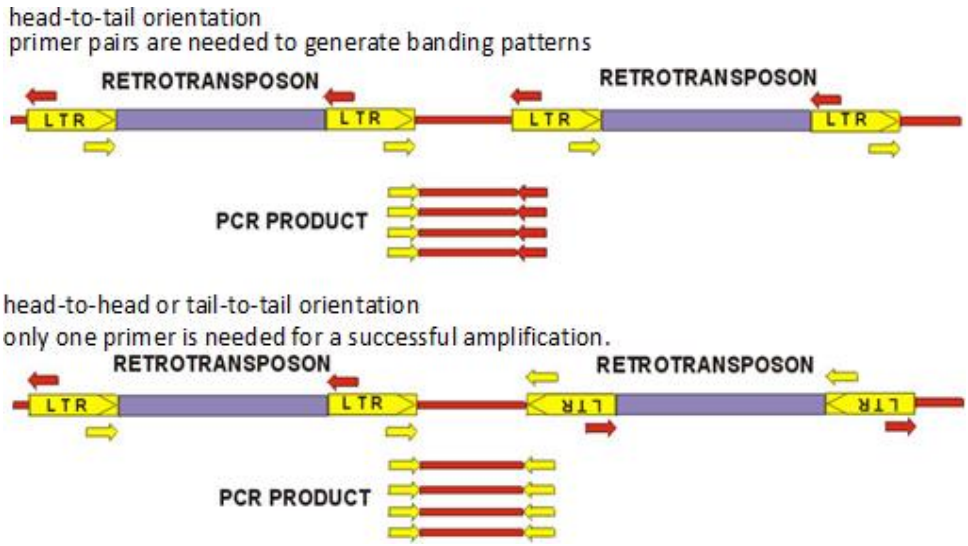


Figure 3.10. Schematic representation of IRAP markers. Arrows indicate direction of amplification and priming sites within LTR retrotransposons. Yellow boxes indicate LTR motifs, Purple boxes represent internal domains, and Red lines show intervening genomic DNA. Colour arrows designate primers.

The data presented here add to our knowledge of genome structure of the *Hieracium* and *Taraxacum*, allowing construction of molecular karyotypes of them and integrating information about major classes of repetitive DNA sequences with the morphology of chromosomes. The rDNA sequences show chromosome-specific distribution patterns and allow identification of individual chromosomes.

The results in this chapter shown that the taxonomic relationships of the microspecies we were using is appropriate to examine evolutionary differences; the data show presence of different repeat types and some variability between accessions. In the next two chapters, to identify the full complement of retroelements, tandem repeat arrays and other types of abundant repetitive sequence, we aimed to use next generation sequencing (NGS) to survey the whole genome of three closely related agamospecies of *Taraxacum officinale* agg. including (A978, O978, S3) which will investigate their repetitive DNA contents involving the diversity between them in the next two chapters.

CHAPTER 4

Complete Chloroplast Genomes from Apomictic *Taraxacum* (Asteraceae): Identity and Variation between three Microspecies

Published in PLoS ONE

Abstract

Chloroplast DNA sequences show substantial variation between higher plant species, and less variation within species, so are typically excellent markers to investigate evolutionary, population and genetic relationships and phylogenies. We sequenced the plastomes of *Taraxacum obtusifrons* Markl. (O978); *T. stridulum* Trávniček ined. (S3); and *T. amplum* Markl. (A978), three apomictic triploid (2n=3x=24) dandelions from the *T. officinale* agg. We aimed to characterize the variation in plastomes, define relationships and correlations with the apomictic microspecies status, and refine placement of the microspecies in the evolutionary or phylogenetic context of the Asteraceae. The chloroplast genomes of accessions O978 and S3 were identical and 151,322 bp long (where the nuclear genes are known to show variation), while A978 was 151,349 bp long. All three genomes contained 135 unique genes, with an additional copy of the *trn*F-GGA gene in the LSC region and 20 duplicated genes in the

IR region, along with short repeats, the typical major Inverted Repeats (IR1 and IR2, 24,431bp long), and Large and Small Single Copy regions (LSC 83,889bp and SSC 18,571bp in O978). Between the two *Taraxacum* plastomes types, we identified 28 SNPs. The distribution of polymorphisms suggests some parts of the Taraxacum plastome are evolving at a slower rate. There was a hemi-nested inversion in the LSC region that is common to Asteraceae, and an SSC inversion from *ndh*F to *rps*15 found only in some Asteraceae lineages. A comparative repeat analysis showed variation between *Taraxacum* and the phylogenetically close genus *Lactuca*, with many more direct repeats of 40bp or more in *Lactuca* (1% larger plastome than *Taraxacum*). When individual genes and non-coding regions were used for Asteraceae phylogeny reconstruction, not all showed the same evolutionary scenario suggesting care is needed for interpretation of relationships if a limited number of markers are used. Studying genotypic diversity in plastomes is important to characterize the nature of evolutionary processes in nuclear and cytoplasmic genomes with the different selection pressures, population structures and breeding systems.

4.1. Introduction

The organization of chloroplast genomes (plastomes) has similarities at the structural and gene level across higher plants (Palmer and Thompson 1982; Jansen *et al.* 2005). The DNA sequences show characteristic variation depending on their taxonomic position, and sequence fragments are widely exploited in molecular taxonomy (Hollingsworth *et al.* 2009). The chloroplast (or, more generally, plastid) genome (plastome, ctDNA, cpDNA) shows maternal inheritance in most species (Birky 2001) and normally there is only one haplotype in a plant. Since there is no sexual recombination among plastomes [although horizontal transfer of whole chloroplasts (Stegemann *et al.* 2012), or chloroplast capture (Wang *et al.* 2013) may occur], chloroplast markers can give robust phylogenies and are then used to estimate divergence times between lineages (Moore *et al.* 2010). The sequencing of the first plastome in *Nicotiana tabacum* (Shinozaki *et al.* 1986) has been followed by some 626 chloroplast whole plastomes belonging to 133 different plant families (including 18

well-defined species from 16 genera in the Asteraceae) deposited in the NCBI Organelle Genome Resources database by early 2016 (Jansen *et al.* 2007; Moore *et al.* 2007; Parks *et al.* 2009).

Typical angiosperm plastome sizes range from 135 to 160 kb (although much reduced in hemi-parasitic plants). The plastome has a conserved quadripartite structure composed of two copies (ca. 25 kb) of an Inverted Repeat (IR) which divides the remainder of the plastome into one Large and one Small Single Copy region (LSC and SSC) (Palmer and Thompson 1982; Jansen *et al.* 2005). One monophyletic clade within the legumes (including the tribes Cicereae, Hedysareae, Trifolieae and Fabeae and some other genera; see Wojciechowski (2006) and all conifers (Raubeson and Jansen 1992) have smaller plastomes in which one copy of the inverted repeat is missing, defining evolutionary lineages. Using whole plastid sequences from two orchid species, Luo *et al.* (2014) demonstrated that chloroplast structure, gene order and content are similar but differ with expansions and contractions at the inverted repeat-small single-copy junction and *ndh* genes.

PCR-amplified sequences within plastomes are used extensively for species identification and reconstruction of phylogeny at around the species level. Several regions are consistently the most variable across angiosperm lineages and some are widely used for barcoding approaches for purposes such as species discovery, floristic surveys, identification of plants, or identification of composition of natural products (e.g. Bruni *et al.* 2010; Bruni *et al. 2015*; Hollingsworth *et al.* 2016), following amplification and sequencing: *ndh*F-*rpl*32, *rpl*32-*trn*L-UAG, *ndh*C-*trn*V, 5'*rps*16-*trn*Q, *psb*E-*pet*L, *trn*T-*psb*D, *pet*A-*psb*J, and *rpl*16 (e.g. Dong *et al.* 2012). However, there is no universal 'best' region. The average number of regions applied to inter specific studies is about 2.5, which may be too little to access the full discriminating power of this plastome (Shaw *et al.* 2014). It is important to have multiple complete plastome for species across a family as references both to characterize any major structural changes, which would be difficult to identify from fragments, and to aid design of conserved PCR primers to exploit polymorphic regions in larger samples within and between taxa.

What are the limits on use of chloroplast sequences for addressing taxonomic questions? The answer depends on the rates of evolution and nature of variation found at different regions of the plastome. Shaw *et al.* (2014) commented on the use of plastome sequences at increasingly low taxonomic levels: the genes most commonly analysed after amplification by PCR may be appropriate for delineation of species but may not represent the most variable regions of the chloroplast. In date palm, chloroplast haplotypes may correlate with populations (Zehdi-Azouzi *et al.* 2015), although founder effects may be strong in such species. Särkinen and George (2013) used full plastome sequences of *Solanum* chloroplasts to identify the most variable plastid markers, concluding that different chloroplast regions are appropriate for study of evolution at different taxonomic levels from family downwards.

In the Asteraceae, Wang *et al.* (2015) have analysed 81 genes from chloroplasts of 70 different species, showing the family is monophyletic and branching is consistent with tribal relationships as understood on the basis on morphology. The Asteraceae family includes an inversion in the plastome relative to other eudicots (Kim *et al*. 2005). The boundaries of a 22.8 kb inversion define a split within the family, and a second 3.3 kb inversion is nested within the larger inversion. Generally, one of the end points of the smaller inversion is upstream of the gene *trn*E, and the other end point is located between the gene *trn*C and *rpo*B. The two inversions are similar among members of the Asteraceae lineage suggesting that the second inversion event occurred within a short evolutionary time after the first event. Estimates of divergence times based on *ndh*F and *rbc*L gene sequences suggest that two inversions originated during the late Eocene (38–42 MYA), soon after the Asteraceae originated in the mid Eocene (42–47 MYA - Kim *et al*. 2005).

The genus *Taraxacum* (Cichorieae, Asteraceae) is known for its complex reticular evolution including polyploidy events, hybridization and apomixis (Asker and Jerling 1992) that makes it difficult to reconstruct a reliable phylogeny. Repeated hybridization between sexual (diploids or rarely tetraploids) and apomictic (triploids and higher ploidies) taxa, rapid colonization of wide areas by apomicts after the Last Glacial Maximum (LGM), low levels of morphological differentiation and remaining

ancestral sequence polymorphisms have been of interest and a challenge to botanists for more than a century [e.g. Nägeli, having seen the results of Mendel (1866), suggested that Mendel should investigate the apomictic *Hieracium* species, see (Nogler 1984; Mogie and Ford 1988; Richards 1973; King and Schaal 1990; Kirschner *et al*. 2015)]. Investigation of genotypic diversity in pure apomictic and mixed sexual-apomictic populations showed variation arises from both mutation (accumulation of somatic mutations/allele divergence) and recombination (gene flow between sexual-apomictic individuals), (Van der Hulst *et al.* 2000; Mes *et al.* 2002; Majeský *et al.* 2012; 2015). Utilization of common chloroplast markers from coding and non-coding regions showed at best weak differentiation within the genus but helped to distinguish evolutionary old and primitive from evolutionary younger or more advanced groups of haplotypes (Wittzell 1999, Kirschner *et al*. 2003). Nevertheless, observed haplotypes were not species specific, some being rare while others were frequent and shared among different and not related taxa, even between sexual and apomictic plants [e.g. (Wittzell 1999; Majeský *et al.* 2012; 2015)]. Mes *et al.* (2000) showed a high level of homoplasy in several non-coding plastome regions.

Here we aimed to sequence whole chloroplast genomes (plastomes) of three morphologically well-defined apomictic microspecies or agamospecies from the *Taraxacum officinale* aggregate (dandelions), namely *T. obtusifrons, T. stridulum* and *T. amplum*. Our goals were to characterize the nature and scale of differentiation between plastomes in three related apomictic taxa and see if there were features of plastome variation that may be a consequence of apomixis. We then aimed to find the evolutionary relationships between the plastomes in the microspecies, and place them phylogenetically in the genus *Taraxacum*, the tribe Cichorieae and the Asteraceae. The results also aimed to identify appropriate regions for use as markers in future studies comparing mutation and inheritance of the nuclear genome in the apomicts with the maternally inherited plastome.

4.2. Materials and methods

4.2.1. Plant material and DNA sequencing

Three agamospecies (2n=3x=24) of *Taraxacum officinale* agg. [section *Taraxacum* (formerly *Ruderalia*), Asteraceae], *T. obtusifrons* Markl. (O978); *T. stridulum* Trávniček ined. (S3); and *T. amplum* Markl. (A978) were germinated and planted in pots. The seeds came from the agamospermous progeny of maternal plants genotyped by nuclear markers by Majeský *et al.* (2012) and ploidy was measured by chromosome counts and flow cytometry (Majeský *et al*. 2012). Geographical records of origin and voucher specimens are deposited in the Herbarium of the Department of Botany, Palacký University, Olomouc, Czech Republic (herbarium abbreviation: OL). Nuclear markers confirm the genotypes used for sequencing; plants were karyotyped showing 2n=3x=24 chromosomes, and voucher specimens of the sequenced plants have been deposited in the University of Leicester, UK, herbarium (LTR). Total DNA including nuclear, mitochondrial and plastome DNA was extracted from fresh green young leaves using standard cetyl-trimethyl-ammonium bromide (CTAB) methods (Doyle and Doyle 1987) to obtain high quality DNA.

DNA was sequenced commercially (Interdisciplinary Centre for Biotechnology Research, University of Florida, USA); accession S3 was sequenced with Illumina Miseq 2x300bp paired end reads while accessions O978 and A978 were sequenced using Illumina Hiseq500 2x150bp reads. About 59,258,642 paired-end reads were obtained for S3 (22 Gb), and 58,713,854 and 69,056,774 paired-end reads (12 Gb) were obtained for A978 and O978 respectively.

4.2.2. Sequence assembly

Assembly and analysis of the plastomes were performed on Ubuntu Linux 13.10, with Geneious version 7.1.4 and later (Kearse *et al.* 2012), (available from http://www.geneious.com/). Using paired end reads from S3, *de novo* assembly generated one large contig of >150,000 bp (420,584 reads) which was largely homologous to the *Lactuca sativa* var. *salinas* (DQ_383816; Asteraceae), (Timme *et al.*

2007) plastome which was then used to generate a consensus reference sequence. For A978 and O978, and for final assembly of the S3 plastome, all raw reads were mapped to the S3 reference (five iterations). The initial assembly showed some areas of double-coverage of repeated regions, and minimal coverage at the four junctions between IRs and the SSC/LSC regions; repeated assembly to short regions corrected these, until uniform coverage with no assembly gaps, high similarity of all assembled reads to the consensus, and minimal unmatched paired reads, was achieved. Plastome bases were numbered so the first base pair after IR2, immediately before the *trn*H gene, became base number 1.

4.2.3. Plastome annotation

Coding sequences and directions were identified in the *Taraxacum* plastome and genes; rRNA and tRNAs were annotated with the Geneious annotation function and DOGMA (Dual Organellar Genome Annotator (Wyman *et al.* 2004), http://bugmaster.jgi-psf.org/dogma/) with reference to published plastomes. In particular, the *Taraxacum* annotation was optimized by comparison with *Lactuca* (DQ_383816) to identify gene and exon boundaries, and tRNA genes were further confirmed with the online tRNAscan-SE 1.21 search server (Lowe and Eddy 1997). A circular plastome map was drawn using the online program GenomeVX (Conant and Wolfe 2008).

4.2.4. Short repeat motifs

REPuter (Kurtz *et al.* 2001) was used to identify and locate DNA repeats including direct (forward), inverted (palindrome) repeats, reverse, and complementary sequences more than 20 bp long (90% identity; Hamming distance 2). TandemRepeatFinder (Benson 1999) was used to find tandem repeats.

4.2.5. Comparison of chloroplast features and phylogenetic analyses

To see the extent of difference between *Taraxacum* and 21 Asteraceae accessions with full plastome sequences, GC content, genome size, gene content and nature of

LSC/SSC/IR were compared. Further, we compared the plastid sequences among 18 species and 16 genera in Asteraceae aligning the entire chloroplast (downloaded from GenBank) and the three *Taraxacum* plastomes. Based on primary alignment, regions with the highest sequence divergence were visualised in mVISTA program (Frazer *et al*. 2004) in Shuffle-LAGAN mode with default parameters to reveal their sequence variation. The alignments were visually checked and edited manually. Based on the comparison of plastome sequences, the regions with highest sequence polymorphism levels were chosen for further phylogenetic analyses. The aim of the phylogenetic analyses was to examine the congruence of the phylogenetic trees with respect to placement of the three *Taraxacum* microspecies within the subsampled Asteraceae family (with the whole plastome sequences available) and with respect to used plastome region for phylogeny reconstruction.

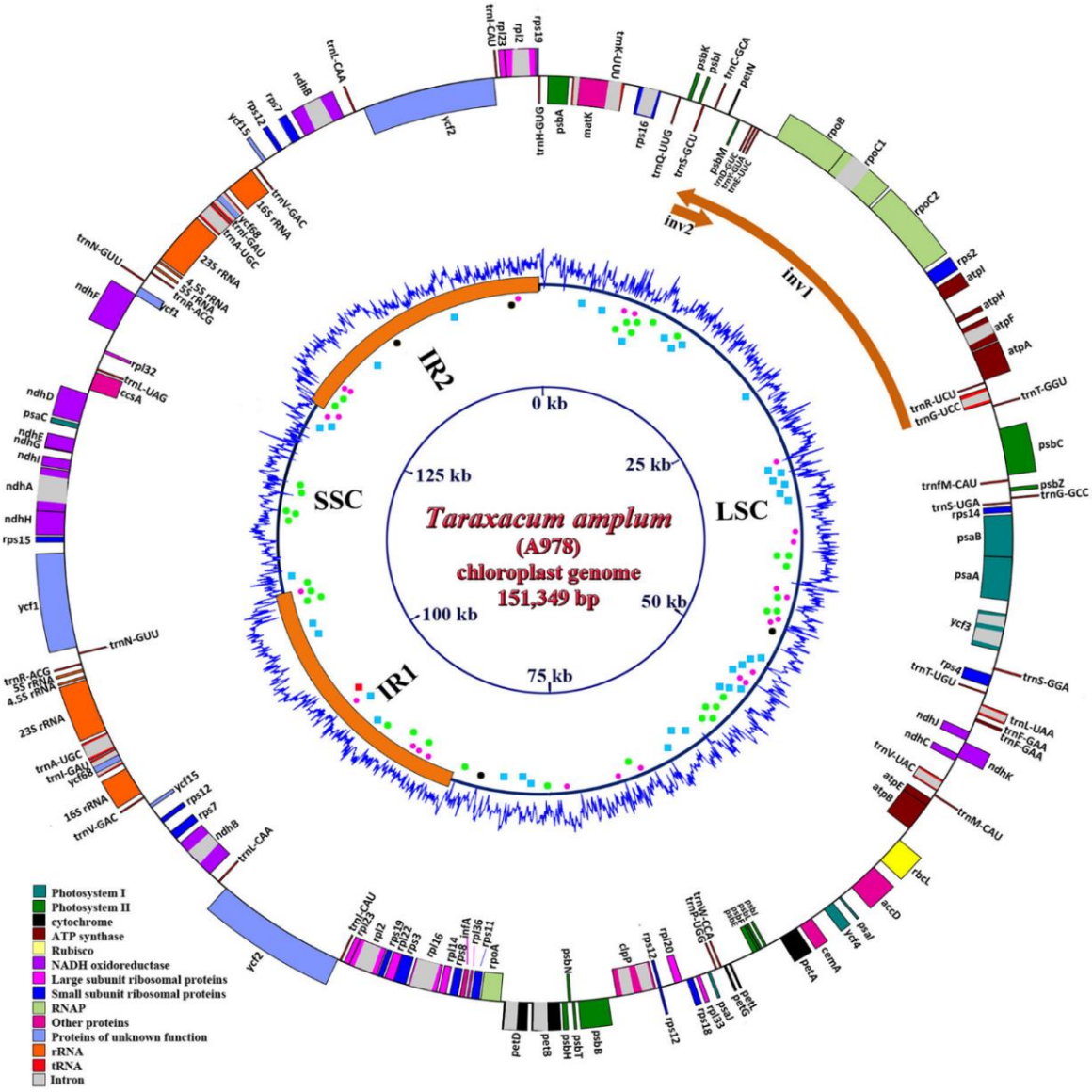
Maximum Likelihood fits of 24 different nucleotide substitution models for 22 accessions using the whole chloroplast genome plus 40 genic and inter-genic regions were calculated, and evolutionary analyses were conducted in MEGA6 (Tamura *et al.* 2013).

Phylogenetic analysis was conducted using the maximum likelihood (ML) method based on the best-fitted model of evolution as outlined in Supplementary Table 4.1. The bootstrap consensus tree was inferred from 1000 replicates (Felsenstein 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories). All three codon positions were included. Analyses were conducted in MEGA6 (Tamura *et al.* 2013). Trees were built for the entire plastome, 24 non-coding intergenic regions, 11 coding regions (including one intron), as well as separate analyses for the LSC, SSC and IR regions, tRNA and rRNA, genes in order to evaluate intragenomic variation in rates of molecular evolution, using *Nicotiana tabacum* (Solanaceae) as the outgroup.

4.3. Results

4.3.1. Structure of *Taraxacum* chloroplasts

Circular plastomes were assembled from the whole genome sequence data (average plastid coverage >2000 fold for each accession). The chloroplasts of accessions O978 and S3 were identical and 151,322 bp long, while A978 was 151,349 bp long. Figure 4.1 shows the circular map for the A978 accession, with genes, short repeats, the major Inverted Repeats (IR1 and IR2; 24,431 bp; see Figure 4.2), and LSC/SSC regions (LSC 83,889bp and SSC 18,571bp in O978 and S3). GC content (blue graph) was higher than average in the 7kb of the Inverted Repeat regions adjacent to the SSC.



← Figure 4.1. Map of the plastome of *Taraxacum amplum* (A978). Genes are shown inside or outside the circle to indicate clockwise or counterclockwise transcription direction respectively. The Inverted Repeat (IR, 24,431bp) is indicated by a thicker line for IR1 and IR2. GC content is show in the inner blue graph. Small Single Copy (SSC) and long single copy (LSC) regions are indicated, and the inverted regions (Inv1 and Inv2) within LSC relative to other species are shown as orange arcs. Short tandem repeats (microsatellites and minisatellites) are indicated by blue dots, palindromes by red dots, forward repeats by green dots and reverse repeats by black dots.

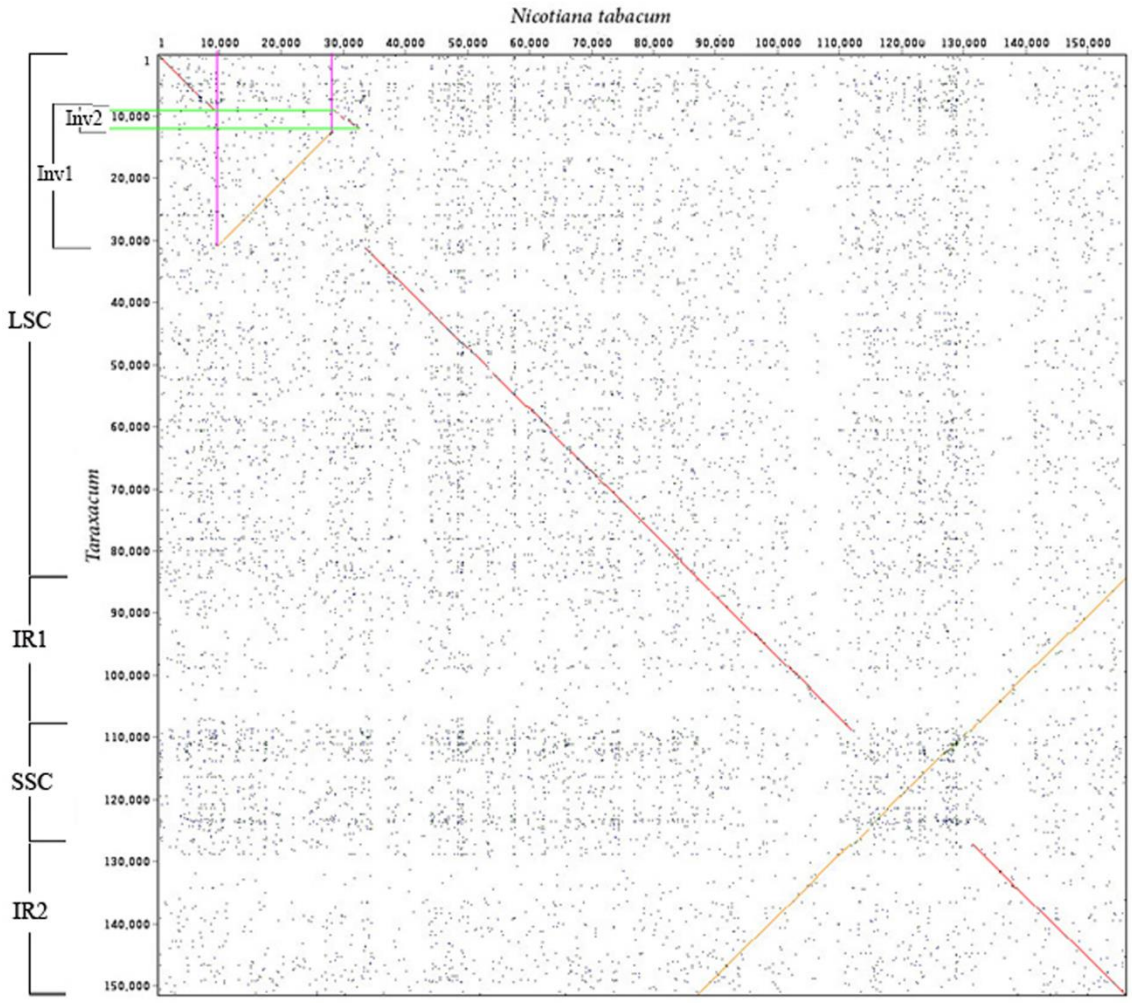


Figure 4.2. Dot-plot sequence comparison of *Taraxacum* and *Nicotiana* chloroplast sequences, showing the Inverted Repeats (IR1 and IR2), hemi-nested inversions between the two plastomes (Inv1 and Inv2) and inversion of the SSC.

4.3.2. Chloroplast genome polymorphism between *Taraxacum* microspecies

Between the two *Taraxacum* plastomes, there were 28 SNPs (9 transversions and 19 transitions; Chi-square=15.1; p=0.0001), occurring in all regions of the plastome (13 in LSC, 13 in SSC and 2 in IRs; Table 4.1 and Figure 4.1). Two SNPs in LSC genes (*rpo*C1 and *acc*D) were non-synonymous changes with the other 9 SNPs in genes being synonymous. There were 16 indels between 1 and 24 bp long, all but one occurring within the LSC region (the LSC representing 55% of the plastome; p<0.001; Table 4.1). A unique 22bp insertion, the duplicated 11bp motif TGTAGACATAA in an intron of the *trn*L-UAA gene, was present in accession A978 (Supplementary Figure 4.1). Overall, non-coding regions show a higher sequence divergence than coding regions in *Taraxacum* (Table 4.1). In the sequence alignment, the highest divergence was seen in regions including the intergenic spacer of *trn*H-*psb*A*, trn*K-*rps*16*, rps*16-*trn*Q*, trn*S-*trn*C, *trn*C-*pet*N*, rpo*C2-*rps*2*, psb*Z-*trn*G*, trn*G-*trnf*M*, ycf*3-*trn*S*, trn*T-*trn*L*, trn*F-*ndh*J*, trn*M-*atp*E*, pet*B-*pet*D*, trn*N-*ycf*1, *ycf*1-*rps*15*, ndh*D-*ccs*A*, rpl*32-*ndh*F*, psbI-trnS, ndhF-ycf1* and *ndhI-ndhG*.

Table 4.1. Transition/transversion and insertion/deletion events between *Taraxacum* microspecies S3/O978 and A978; where indel occurs in a gene, the gene name is indicated; other indels are intergenic.

#	Type	Position	Location	Nucleotide position	S3/O978	A978
1	SNP	LSC/*trn*K-*rps*16	IGS*	4907	T	C
2	SNP	LSC/*rps*16-*trn*Q	IGS	6402	A	G
3	SNP	LSC/*trn*S-*trn*C	IGS	8856	A	G
4	SNP	LSC/*trn*F-ndhJ	IGS	47823	G	A
5	SNP	LSC/*ndh*C-*trn*V	IGS	50219	T	C
6	SNP	LSC/*psb*B	gene	72455	A	G
7	SNP	LSC/*rpl*22	gene	83275	T	C
8	SNP	IR-1/*ycf*2-*trn*L	IGS	93366	G	A
9	SNP	SSC/*ycf*1	gene	109145	A	G
10	SNP	SSC/*ycf*1	gene	111110	G	A
11	SNP	SSC/*ycf*1	gene	112536	T	C
12	SNP	SSC/*ycf*1-*rps*15	IGS	113190	A	G

← Table 1 continue …

#	Type	Position	Location	Nucleotide position	S3/O978	A978
13	SNP	SSC/*ndh*D	gene	120836	A	G
14	SNP	SSC/*ndh*D-*ccs*A	IGS	121274	A	G
15	SNP	SSC/*ndh*D-*ccs*A	IGS	121275	G	A
16	SNP	SSC/*rpl*32-*ndh*F	IGS	124080	T	C
17	SNP	IR-2/*trn*L-*ycf*2	IGS	141978	C	T
18	SNP	LSC/*trn*H-*psb*A	IGS	222	A	C
19	SNP	LSC/*trn*S-*trn*C	IGS	8715	T	G
20	SNP	LSC/*rpo*C1	gene	18257	A	C
21	SNP	LSC/*rpo*C2	gene	20210	A	C
22	SNP	LSC/*acc*D	gene	57577	T	A
23	SNP	LSC/*psa*L-*ycf*4	IGS	59566	A	C
24	SNP	SSC/*psb*B	gene	72515	C	A
25	SNP	SSC/*ycf*1	IGS	112416	C	G
26	SNP	SSC/*ndh*F	gene	126757	A	C
27	SNP	SSC/*ndh*D-*ccs*A	IGS	121278	A	G
28	SNP	SSC/*ndh*D-*ccs*A	IGS	121279	G	A
29	InDel	LSC/*trn*H-*psb*A	IGS	167	-	AAATC
30	InDel	LSC/*rps*16 intron	gene	5417	C	-
31	InDel	LSC/*trn*C-petN	IGS	9481	T	-
32	InDel	LSC/*rpoc*1-intron	gene	16831	GGAAACTTGAGTAAGGAGTAGATC	-
33	InDel	LSC/*rpoc*2-*rps*2	IGS	23086	T	-
34	InDel	LSC/*psb*Z-*trn*G	IGS	35508	-	A
35	InDel	LSC/*trn*G-*trn*fM	IGS	35818	-	AGCCTTC
36	InDel	LSC/*ycf*3-trnS	IGS	43835	A	-
37	InDel	LSC/*ycf3*-trnS	IGS	44098	T	-
38	InDel	LSC/*trn*L_intron	gene	46911	-	TGTAGACATAA
39	InDel	LSC/*trn*M-*atp*E	IGS	52127	-	TTAAAT
40	InDel	LSC/*acc*D	gene	56925	-	GTCTTG
41	InDel	LSC/*ycf*4-*cem*A	IGS	60146	-	AGAAAT
42	InDel	LSC/*clp*P	gene	70273	-	T
43	InDel	LSC/*pet*B-*pet*D	IGS	76172	-	TTTATTTAACATAATATAGTTGA
44	InDel	SSC/*ndh*D-*ccs*A	IGS	121280	ATTTTTATTC	-

* IGS=intergenic spacer region

Gene content and arrangement were identical in all three sequenced *Taraxacum* plastomes. The plastome contains 135 unique genes, including a total of 81 protein-coding genes (plus 9 duplicated in IR), 4 rRNA (all duplicated in the IR) and 38 unique tRNA genes (one in the SSC region, 23 in the LSC region and 7 duplicated in the IR region) with two copies of the *trn*F-GGA gene in the LSC region and four rRNA genes in the IR region (Table 4.2; Figure 4.1). Within the IRs, there are 19 genes duplicated: all four rRNA, seven tRNA and eight protein-coding genes. Only the 5' end of the *ycf*1 genes (467 bp) and 3' end of *rps*19 (67 bp) are present in the IRs, and the gene *rps*12 is trans-spliced, with the 5' exon in the LSC and the remaining two exons in the IRs (Figure 4.1). There are 18 different intron-containing genes (of which six are tRNA coding genes). All intronic genes contain one intron, except two (*ycf*3*, clpP*) that contain two introns. The *trn*K-UUU gene had the largest intron (2,557 bp) with another gene, *mat*K, located in it (Table 4.3).

Table 4.2. Genes present in the *Taraxacum* plastomes.

Category	Gene name
Photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J, ycf3[a], ycf4
Photosystem II	*psb*A, B, C, D, E, F, H, I, J, K, L, M, N, T, Z
Cytochrome b6/f	*pet*A, B [b] , D [b] , G, L, N
ATP synthase	*atp*A, B , E, F[b] , H, I
Rubisco	*rbc*L
NADH Oxidoreductase	*ndh*A[b] , B[b,c] , C, D, E, F, G, H, I, J, K
Large subunit ribosomal proteins	*rpl*2[b,c] , 14, 16[b], 20, 22, 23[c] , 32, 33, 36
Small subunit ribosomal proteins	*rps*2, 3, 4, 7[c] , 8, 11, 12[b,c,d] , 14, 15, 16[b], 18, 19[c]
RNAP	*rpo*A, B, C1[b] , C2
Other proteins	*acc*D, *ccs*A, *cem*A, *clp*P[a] , *mat*K, *inf*A
Proteins of unknown function	*ycf*1, *ycf*2[c] , *ycf*15[c] , *ycf*68[c]
Ribosomal RNAs	rRNA23[c] , 16[c] , 5[c] , 4.5[c]
Transfer RNAs	*trn*A(UGC)[bc], *trn*C(GCA), *trn*D(GUC), *trn*E(UUC, *trn*F(GAA)[f], *trnf*M(CAU), *trn*G(GCC), *trn*G(UCC)[b], *trn*H(GUG), *trn*I(CAU)[c], *trn*I(GAU)[bc], *trn*K(UUU)[b], *trn*L(CAA)[c],*trn*L(UAA)[b], *trn*L(UAG), *trn*M(CAU), *trn*N(GUU)[c], *trn*P(UGG), *trn*Q(UUG), *trn*R(ACG)[c], *trn*R(UCU), trnS(GCU), *trn*S(GGA), *trn*S(UGA), *trn*T(GGU), *trn*T(UGU), *trn*V(GAC)[c],*trn*V(UAC)[b],*trn*W(CCA), *trn*Y(GUA)

[a] Gene containing two introns; [b] Gene containing a single intron; [c] Two gene copies in the IRs; [d] Gene divided into two independent transcription units; [f] Duplicated gene in LSC.

Sequences have been submitted to GenBank (GenBank accession number: KX499523, KX499524, KX499525), and the full raw reads from the three genotypes have been uploaded into SRA with BioSample accessions: SAMN05300515, SAMN05300516, SAMN05300517.

A total of 26233 codons in S3 and O978, and 26253 codons in A978 represent the coding regions of 90 protein-coding genes. Codon usage was biased towards A and T at the third codon position. Among the codons, serine (8.8% and 8.9% of O978, A978 respectively) and methionine (1.77% and 1.80 % of O978, A978 respectively) are the most and the least abundant amino acids (Supplementary Table 4.2).

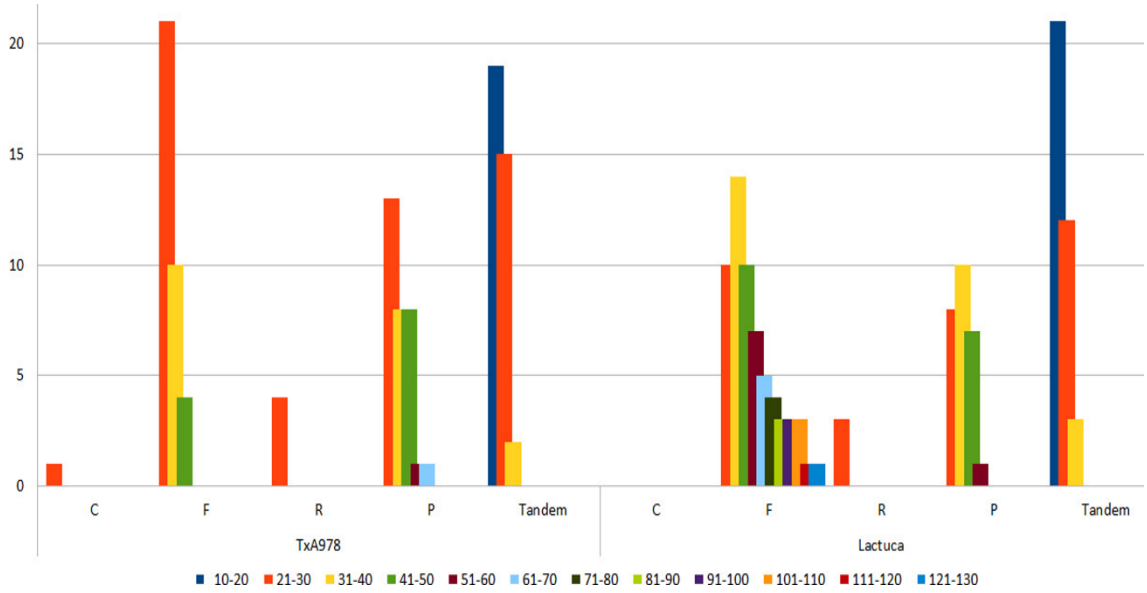
Table 4.3. Intron and exon sizes in genes in the *Taraxacum* plastome.

Genes	Regions	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
*atp*F	LSC	145	707	410	-	-
*ndh*A	SSC	553	1054	539	-	-
*ndh*B	IR	777	669	756	-	-
*pet*B	LSC	642	769	6	-	-
*pet*D	LSC	475	707	8	-	-
*rpl*2	IR	391	665	434	-	-
*rpo*C1	LSC	453	709	1638	-	-
*rps*12	LSC/IR	114	-	243	-	-
*rpl*16	LSC	408	1058	9		
*rps*16	LSC	40	860	227	-	-
*trn*A(UGC)	IR	38	814	35	-	-
*trn*G(UCC)	LSC	23	726	47	-	-
*trn*I(GAU)	IR	43	772	35	-	-
*trn*K(UUU)	LSC	37	2557	35	-	-
*trn*L(UAA)	LSC	37	440	50	-	-
*trn*V(UAC)	LSC	38	572	38	-	-
*ycf*3	LSC	124	690	230	740	153
*clp*P	LSC	71	623	291	812	229

Investigation of various types of repeats present in *Taraxacum* plastome showed the presence of five main types of repeats (complement, forward, reverse, palindromic and tandem) (Figure 4.3, Supplementary Table 4.3). The most abundant

were short repeats of sequence motifs with 21-30 nucleotides, except for tandem repeats, were the most abundant were motifs with only 10-20 nucleotides. Comparison with *Lactuca* (DQ_383816) showed difference in both types of present repeats and length of repeats (Figure 4.3).

Figure 4.3. Repetitive motif abundance in *Taraxacum* (only A978 shown since the three accessions were similar) and *Lactuca* plastomes. C=Complement repeats, P=Palindromic repeats, F=Forward repeats, R=Reverse repeats.



4.3.3. Comparison of chloroplast features between *Taraxacum* and 21 accessions of Asteraceae and phylogenetic analyses.

Comparison of chloroplasts between *Taraxacum* and other Asteraceae (Table 4.4) showed no dramatic difference in compared features (Figure 4.4, numerical data in Supplementary Table 4.4). The most prominent difference was observed in the number of genes with *Taraxacum*, together with *Helianthus annuus* (Supplementary Table 4.4), having the highest gene content (136 genes) from all of the compared species. Genome size, GC content and size of LSC did not vary considerably, while size of SSC was slightly bigger for two taxa (*Parthenium argentatum* and *Leontopodium leiolepis*) and of IR was lower for *Ageratina adenophora* and *Praxelis clematidea* (Figure 4.4).

Table 4.4. List of plastomes from GenBank used for comparison.

Sub-family	Tribe	Organism name	Ref seq.	Reference
Asteroideae	Heliantheae alliance	*Guizotia abyssinica*	NC_010601.1	(Dempewolf *et al.* 2010)
		Helianthus annuus	NC_007977.1	(Timme *et al.* 2007)
		Parthenium argentatum	NC_013553.1	(Kumar *et al.* 2009)
	Anthemideae	*Artemisia frigida*	NC_020607.1	(Liu *et al.* 2013)
		Artemisia montana	NC_025910.1	-
		Chrysanthemum indicum	NC_020320.1	(Liu *et al.* 2012)
		Chrysanthemum X Morifolium	NC_020092.1	(Liu *et al.* 2012)
	Astereae	*Aster spathulifolius*	NC_027434.1	(Choi and Park 2015)
	Senecioneae	*Jacobaea vulgaris*	NC_015543.1	(Doorduin *et al.* 2011)
	Gnaphalieae	*Leontopodium leiolepis*	NC_027835.1	(Lee *et al.* 2015)
	Eupatorieae	*Ageratina adenophora*	NC_015621.1	(Nie *et al.* 2012)
		Praxelis clematidea	NC_023833.1	(Zhang *et al.* 2014)
Cichorideae	Cichorieae	*Lactuca sativa*	NC_007578.1	(Kanamoto *et al.* 2004)
		Lactuca sativa var. salinas	DQ_383816_	(Timme *et al.* 2007)
		Taraxacum amplum (A978)	KX499525	This study
		Taraxacum obtusifrons (O978)	KX499524	This study
		Taraxacum stridulum (S3)	KX499523	This study
Carduoideae	Cynareae	*Centaurea diffusa*	NC_024286.1	(Turner and Grassa 2014)
		Cynara cardunculus	NC_027113.1	(Curci *et al.* 2015)
		Carthamus tinctorius L.		(Lu *et al.* 2015)
	Madieae	*Lathenia burkei*	Km360047	(Walker *et al.* 2014)
Solanaceae		*Nicotiana tabacum*	NC_001879	(Shinozaki *et al.* 1986)

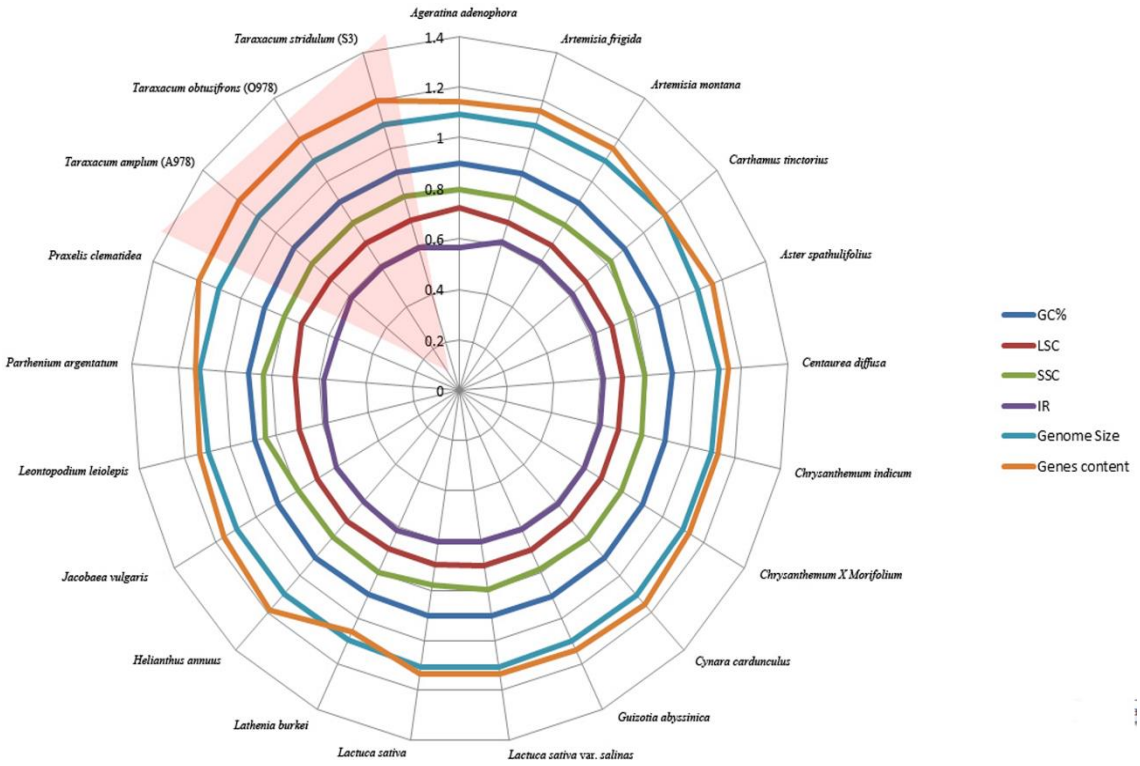


Figure 4.4. A radar-plot comparing features of the plastomes of 21 accessions of Asteraceae, showing, from inside to out, sizes of major plastome regions, GC content, genome size and number of different types of genes.

Based on comparison of sequences of whole plastomes, higher sequence divergence was present within non-coding regions. The most divergent coding regions between *Taraxacum* plastomes and the others 18 Asteraceae plastomes were *rpo*C1, *rpo*C2, *trnL*, *acc*D, *clp*P, *psb*B, *ndh*D, *ycf*1, *ndh*A, *rps*16 and *ndh*F (Supplementary Figure 4.2). Using the Maximum Likelihood method and nucleotide substitution models with minimum Bayesian information criterion (BIC) value for each tree from MEGA6 (Tamura *et al.* 2013), (Supplementary Table 4.1), 41 trees were produced. In all of them, the three *Taraxacum* microspecies appeared as a clade which usually (in 33 of the 41 trees) showed a well-supported sister group relationship to Lactuca. This is consistent with both genera belonging to subfamily Cichorioideae. Some DNA regions showed either a paraphyleteic (rRNA, tRNA, *trn*G-*trnf*M, *pet*A-*psb*J, *clp*P) or polyphyletic (*trn*H-*psb*A, *rpo*C2-rps2, *trn*S-*trn*C) Cichorioideae (Supplementary Figure

4.3), but most of these were relatively short sequences. Species in subfamily Carduoideae (belonging to the genera *Cynara*, *Centaurea* and *Carthamus*) were often sister to Cichorioideae (in 17 of the 41 trees), but there were several where other groups showed this relationship.

4.4. Discussion

Species in the Asteraceae family have contrasting evolutionary pressures from intense selection by people in agricultural and weedy species, with presumably relaxed selection in favourable niches, and there are some invasive species with genetic bottlenecks. The species also have various breeding systems including apomixis, sporophytic self-incompatibility, cleistogamy, wind and insect pollination and there is interest in the use of more apomictic crop species. With whole plastome sequences and comparisons between families, it will be valuable to identify the nature of evolutionary processes in nuclear and cytoplasmic genomes with the different selection pressures, population structures and breeding systems. Here, we provide brief discussion of main features of *Taraxacum* plastome gained form sequencing of whole chloroplasts in three apomictic accessions.

4.4.1. Chloroplast genome polymorphisms between *Taraxacum* microspecies and differentiation power of plastome sequences at low taxonomic level

The three apomictic accessions for which whole plastome sequences were generated in the present study belong to a group of common dandelions (generally called *T. officinale* aggregate). Sequenced individuals represent agamospermous progeny of maternal plants genotyped by nuclear markers by Majeský *et al.* (2012). This genotyping showed two defined groups (OSP and AMP) and supported the presence of nine tight genetic clusters among the nine studied apomictic accessions (for details see Majeský *et al.* 2012). The genotyping agreed with the morphologically-based division of the accessions into separate apomictic microspecies (a taxonomic rank for apomictic taxa based on morphology). Of the three apomictic microspecies sequenced in the

present study, two (O978, S3) belong to the OSP group and A978 belongs to the AMP group. Despite their clear and robu st nuclear differentiation, sequencing of the chloroplast *trn*L–*trn*F intergenic spacer showed they shared the cp*1a* haplotype: haplotype cp*1a* (haplotype 18a in Wittzell 1999) is the most common (derived) haplotype shared among wide spectrum of different sections (dandelion groups) in *Taraxacum* (Wittzell 1999; Majeský *et al.* 2012; 2015). This suggests haplotype cp*1a* might be derived from the most recent common ancestor of many derived *Taraxacum* sections.

Van der Hulst *et al.* (2003 their Figure 3), identified three *Taraxacum* chloroplast haplotypes in more than two plants (namely C1, C2 and C4), and found these were not restricted to single clades based on nuclear marker data (AFLPs (amplified fragment length polymorphisms) and microsatellites). They were neither monophyletic nor congruent with nuclear markers, thus negating the model that matrilineal markers would delimit nuclear marker data to matrilineal groups and thus detect clonal lineages. However, this study employed population-based sampling (randomly sampled individuals within a 'park lawn'). In such a habitat many different morphological clones (microspecies) coexist (see e.g. Ford 1985; Richards 1986) with different origin. In the case of apomicts, like *Taraxacum*, nuclear markers are able to delimit clonal lineages (Majeský *et al.* 2012, Kirschner *et al.* 2016) and the extent of a clonal lineage can be supported by matrilineal markers, although not unambiguously, (e.g. see Majeský *et al.* 2012). However, the markers used only consider a small fraction of the whole chloroplast and inevitably cannot discover all differences within particular chloroplast lineages. Whole plastome sequencing of well-defined samples measured all genetic variability among the three apomictic dandelions. The plastome sequences were identical in the two apomictic accessions O978 and S3, belonging to same morphological group OSP, and differed by 27bp in length, 28 SNPs and 16 indels from A978, belonging to different AMP group (Table 4.1).

What do these results show about the relationship between the apomictic microspecies where we sequenced the plastome, following the work of Majeský *et al.* (2012). Plastomes are evolving at a different, slower rate, compared to nuclear

markers, as noted by Wolfe *et al.* (1987). While nuclear markers showed genetic boundaries between the O and S sub-groups, whole chloroplast sequences did not. This may point to the young evolutionary age of the two microspecies (*T. obtusifrons* and *T. stridulum*): they have not accumulated any chloroplast mutations between each other and their most recent common ancestor. Morphologically, they are well-defined as separate morphological units (Majeský *et al.* 2012) with a low number of observed genotypes within investigated individuals from the O and S microspecies: two (*T. obtusifrons*) and four (*T. stridulum*) multilocus genotypes were detected by six nuclear SSRs (simple sequence repeats) among 21 and 23 genotyped individuals, while AFLPs showed only one AFLP-phenotype among 10 fingerprinted individuals of both microspecies. Apomictic reproduction cuts off a lineage from genetic recombination so an asexual lineage is expected to rapidly diverge as a result of accumulation of mutations and transposon activity that become the major generators of diversity and driver for genome evolution (Richards 1989; Heslop-Harrison *et al*. 1997).

4.4.2. Comparison of *Taraxacum* plastome with other genera

Sequence comparison of the plastome of *Taraxacum* with the reference *Nicotiana tabacum* (Shinozaki *et al*. 1986) revealed hemi-nested inversions in the LSC region (Inv1: 21,737 bp in S3/O978, and 21,711 bp in A978; inv2 of 2,543 bp in S3/O978 and 2,542 bp in A978; Figure 4.1 and 4.2). The nested inversion ended just upstream of the *trn*E-UUC gene with the large inversion. The other end-point of the inversion is located between the *trn*C-GCA and *rpo*B genes (Figure 4.5). The inversion in the LSC [Inv1 and Inv2; (Kim *et al*. 2005, Timme *et al*. 2007)] is conserved across all 21 Asteraceae chloroplast sequences. Liu *et al.* (2013) suggested that the LSC inversion region has undergone inversion followed by reinversion in Asteraceae, and that this could be a particularly active region for sequence rearrangements in the plastome: the existence of within-species variation in the presence of this major inversion supports the hypothesis that this region is a hotspot for inversion events (Figure 4.5).

Another large inversion between *N. tabacum* and *Taraxacum* (Figure 4.2 and 6) is present between base pair positions 108321 in S3, O978 (108,358 in A978) and 126891

in S3, O978 (126,919 in A978); it is flanked by inverted repeats and encompasses the entire SSC region (18,571 bp in S3, 18,561 bp in A978) (Figure 4.2, 4.6, 4.7). The SSC inversion from *ndh*F to *rps15* is present in all of the Asteraceae lineages involved in this study except *Artemisia frigida* (NC_020607) (Liu *et al*. 2013), *Artemisia montana* (NC_025910), *Carthamus tinctorius* (KP404628) (Lu *et al*. 2015), *Centaurea diffusa* (NC_024286) (Ahmed *et al*. 2012) and one reported *Lactuca sativa* (NC_007578, Kanamoto *et al*. 2004).

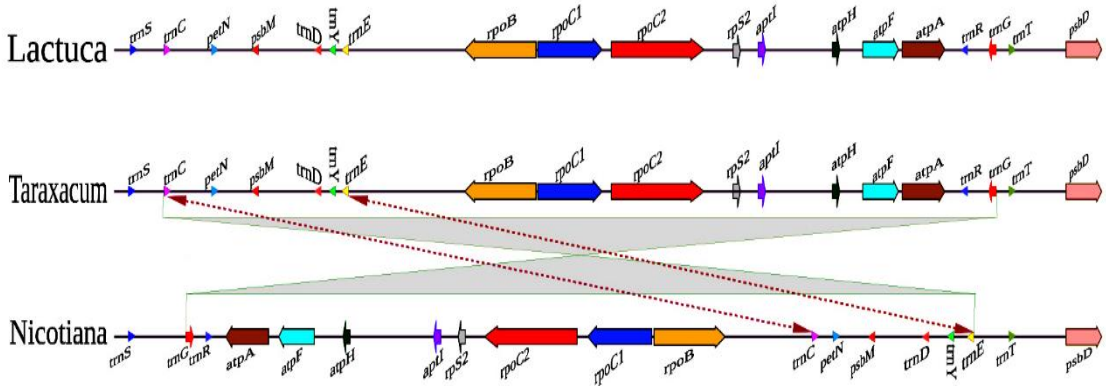


Figure 4.5. Comparative plastome maps. Endpoints of the large 22 kb inversion present in most Asteraceae and of a small inversion (3.3 kb in other Asteraceae).

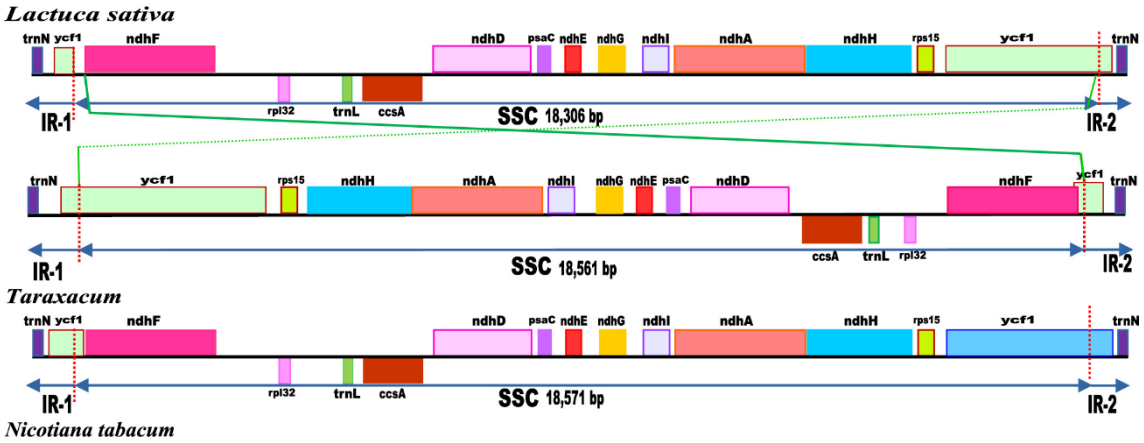


Figure 4.6. Comparative plastome maps. Gene order and inversion of the SSC region. Gene sequences were annotated and indicated along the black lines. Genes above the black lines indicate their transcription in reverse direction and genes below the black lines represent their transcription in forward direction.

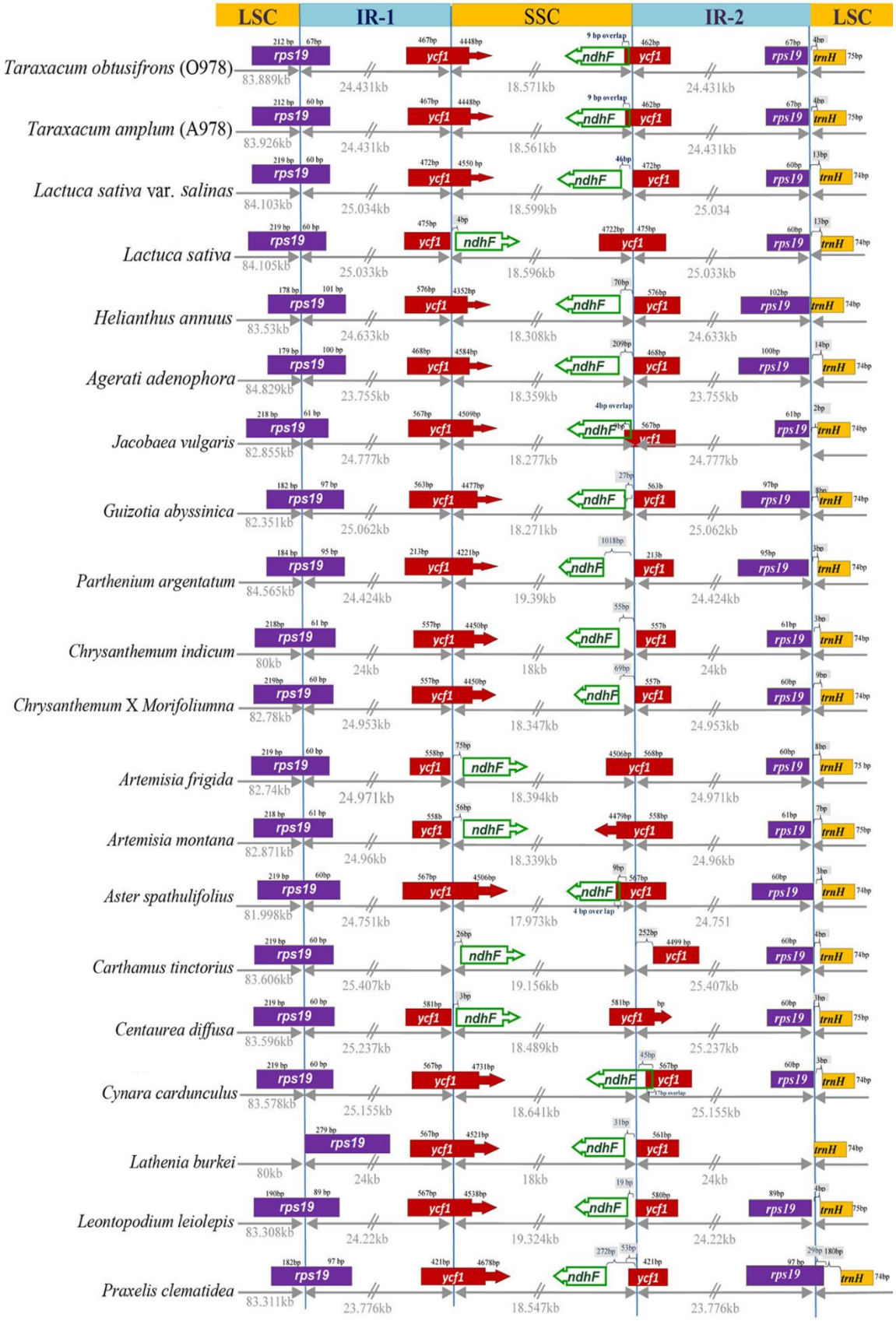


Figure 4.7. Comparative plastome maps. Border position of LSC, IR and SSC region among the 20 Asteraceae plastomes. Genes are indicated by colored boxes.

Comparison of features of the plastomes of 21 accessions of Asteraceae, showed overall similarity of chloroplasts across wider spectrum of different evolutionary lineages. There were even no dramatic differences among representatives of the three main subfamilies (Carduoidae, Cichorioidae, Asteroidae), what may stress overall high stability of chloroplast features at lower taxonomic level (Supplementary Table 4.4, Figure 4.4). The most remarkable difference was seen in the number of total tRNA and coding genes (Supplementary Table 4.4), with *Lasthenia burkei* being taxon with the lowest number of genes (119 - Total Gene N°/79 - N° Coding Genes/20 - N° tRNA) comparing with the three *Taraxacum* (136 – Total Gene N°/90 - N° Coding Genes/38 - N° tRNA). Holmquist (1992) considered that recombinogenic domains of chromosomes may be GC rich. Figure 4.1 shows that the GC content was lower in the SSC region flanked by IR1 and IR2, and higher in 7kb (of the 24kb) of the IR regions 1kb away from the SSC border, with an evident spike from low GC at end of both IRs; both ends of Inv1 had a low GC content. Thus, as found by Walker *et al.* (2014), high GC content was not associated with inversion breakpoints in the plastome.

The number of direct (forward), reverse, palindromic and tandemly repeated sequence motifs of various length classes in *Taraxacum*, compared with *Lactuca* (DQ_383816), can be seen in Figure 4.3 (see also Supplementary Table 4.3). The notable difference was the increased frequency of direct repeats more than 50bp long in *Lactuca*, where there were 27 compared to none in *Taraxacum* (37 compared to 4 repeats >40bp long). Liu *et al.* (2013) commented on variation in number and variety of repeats in the Asteraceae plastomes. Repeats have a role in plastome organization, but like Liu *et al.* (2013), we found no correlation between large repeats and rearrangement endpoints. Our comparative repeat analysis showed considerable variation between even *Taraxacum* and *Lactuca*, with many more direct repeats of 40bp or more in *Lactuca* (Figure 4.3; 1% larger plastome than *Taraxacum*). Relationships of repeats and mutation have been considered in chloroplast genomes (Ahmed *et al.* 2012), although in the related *Taraxacum* plastomes, SNPs and non-repeat indels showed little relationship with repeats.

4.4.3. Phylogenetic utility of chloroplast regions

Polymorphisms between the two *Taraxacum* plastomes and between *Taraxacum* and other Asteraceae included many chloroplast regions widely used for phylogenetic analysis. The presence of two *trn*F-GAA genes duplicated in the LSC is unusual and would make this region difficult to use for phylogeny and diversity studies (Supplementary Figure 4.4). Duplication of *trn*F-GAA gene was encountered already by Wittzell (1999), who, based on sequence variation of *trn*L-*trn*F region in number of different *Taraxacum* taxa, provide support for the informal division of dandelions on evolutionarily old and evolutionary younger/derived taxa. The presence of duplicated *trn*F gene is not specific only for *Taraxacum*, but is present also in other compared species of Asteraceae: namely in *Carthamus tinctorius*, *Guizotia abyssinica*, *Ageratina adenophora*, *Praxelis clematidea* and *Lasthenia burkei* (Supplementary Figure 4.4). Thus, duplication of the *trn*F-GAA gene probably occurred several (at least three or four) times independently in the three main Asteraceae subfamilies: Asteroideae, Cichorideae, and Carduoideae.

All three investigated apomictic *Taraxacum* microspecies represented separate clade sister to *Lactuca* in all phylogenetic analyses (Supplementary Figure 4.3). This was expected because *Taraxacum* and *Lactuca* belong to the same evolutionary lineage – Cichorioideae – within the Asteraceae family (no other species of Cichorioideae was included). This is also in accordance with the current knowledge of the relationships within the subfamily (Kilian *et al*. 2009). Although the close relationships of both genera, *Taraxacum* represent a distinct evolutionary lineage (Crepidinae) than *Lactuca* (Lactucinae) (Kilian *et al*. 2009) which according to Tremetsberger *et al.* (2013) have diverged during the Miocene, at least 16.2 MYA. Because of low level of sequence divergence between the investigated *Taraxacum* accessions and because these microspecies represent only a scant part of species known in the genus, it is not possible to draw some conclusions about their evolutionary relationships. In part of the phylograms accession A978 appeared to be basal to O978/S3, but other phylograms do not support this and the relations between the plastomes appeared as unresolved. Definitely, whole plastome sequences provide

far more discrimination power than individual markers, for phylogeny reconstruction. For deeper insight into the evolution of the *Taraxacum* genus, it will require wider sampling of more distinct taxa. Kirschner *et al.* (2003) used a parsimony analysis of morphological and chloroplast data (two intergenic spacers *psb*A-*trn*H + *trn*L-*trn*F) in *Taraxacum* to show an overall lack of congruence. They suggested the conflict was a consequence of reticulation affecting morphology (and presumably nuclear markers), a process unlikely for the chloroplast genomes. Intergenic spacer *psb*A-*trn*H belonged among the most divergent plastome regions (in the sense of sequence divergence between the two distinct plastomes A978 versus O978/S3) in our analyses (presence of one SNP and 5bp Indel; Table 4.1), but as noted above, no sequence variation was observed among the three investigated accessions for the *trn*L-*trn*F intergenic spacer.

Both the more conserved coding regions and variable non-coding regions of the chloroplast genome have proved useful for phylogenetic studies (Nie *et al*. 2012; Walker *et al*. 2014), with faster rates of evolution in noncoding regions; however the data here show care is needed in interpretation based on single regions as might be amplified by 'barcode' markers. Maybe some incongruences arise where mutations are reiterated (similarities are not identical by descent), although rare male chloroplast transmission (e.g. McCauley *et al*. 2007; Ellis *et al*. 2008) and recombination events cannot be ruled out.

It is important to select marker sequences which have a rate of evolution that is appropriate to the evolutionary distance of the accessions under analysis and the questions being addressed (Saeidi *et al*. 2006). Walker *et al.* (2014) have pointed out that rates of molecular evolution vary over the plastome, particularly in noncoding regions. Here, two of the plastomes, from accessions which are in well-defined clades based on morphology and nuclear DNA markers, were identical: without the full plastome sequence, there would always have been questions about whether the plastome markers we happened to use were appropriate. It was also evident that the most frequently used chloroplast markers (including *trn*L-*trn*F, and *mat*K) showed few polymorphisms between O/S and A *Taraxacum* and to position *Taraxacum* with respect to other *Taraxacum* microspecies, and of the species in the Cichorieae.

It would be interesting to determine the timings of evolutionary separation of the *Taraxacum* microspecies, and of the species in the Cichorieae. This would enable comparisons of evolutionary rates of sexual and apomictic species, and between nuclear and plastome sequences. Tremetsberger *et al.* (2013) used fossil-calibration based on pollen and a nuclear sequence to estimate divergence between species in the group, but the prehistoric and fossil record for the majority of the Asteraceae, including *Taraxacum,* is poor (Tremetsberger *et al.* 2013; Sterk *et al.* 1987; Richards 1973).

We expect whole genome sequencing (Walker *et al.* 2014) to be used increasingly for taxonomy and systematics, within-species biodiversity, population, phylogenetic and evolutionary projects. With the total cellular DNA used here, without enrichment for chloroplast sequences, 3.5 to 4% of reads mapped to the chloroplast [400 unreplicated plastomes per 1C (unreplicated haploid) nuclear genome], allowing robust assembly including the duplications and inversions. Even with automation, PCR amplification and sequencing of multiple regions of chloroplasts and nuclear plastomes is time-consuming and requires optimization, while whole plastome sequencing only requires DNA extraction and a service provider. Analysis and interpretation of whole-genome-sequencing results is, however not yet optimized nor routine.

In the current study, we sequence full chloroplast of three well characterized apomictic *Taraxacum* microspecies. We provide the full annotated plastome sequences for the genus, which can be used in diverse spectrum of further comparative analyses and provide reference plastome for primer design in taxonomic and phylogenetic studies of the genus. We also showed the low sequence divergence between the investigated apomictic taxa, what point to their recent origin (probably post-Pleistocenic). The sequenced plastome (A978) may represent the most common recent chloroplast type involved in origin of many evolutionarily younger *Taraxacum* taxa.

CHAPTER 5

Repetitive DNA in genomic sequences of *Taraxacum* (Asteraceae) and variation between microspecies

Abstract

Repetitive DNA including transposable elements (TEs) are largely and dynamically evolving parts of eukaryotic genomes, especially in plants. Their repetitive nature and their abundance in the genome makes them challenging to study using classical methods of molecular biology. Next generation sequencing and new computational tools have greatly facilitated the investigation of transposable element variation within species and among closely related species. An analysis of about 45Gb sequence (10x to 20× genome coverage) of three closely related *Taraxacum* microspecies, by implementing raw reads in RepeatExplorer and designing probes for *in situ* hybridization, has been performed. The analysis provided characterization of repetitive DNA, which makes up about 50-61% of the genome. The results showed that repeats in the *Taraxacum* microspecies are made of various types of Ty1-*copia* (13-16%) and Ty3-*gypsy* (10-14%) retroelements, while DNA transposons were found to be rare. Also, some other unknown repetitive DNA clusters were investigated. The results showed that asexual *Taraxacum* agamospecies may contain a lower diversity of

transposable elements than sexual taxa of *Helianthus*. The study shows differences between the three *Taraxacum* microspecies in both genomic proportions of repetitive DNA type and by analysing the repetitive DNA by *in situ* hybridizations. Specific 49bp tandem DNA repeats were characterized and by means of probes, these were located on the satellites of three chromosomes.

5.1. Introduction

Most angiosperm plant DNA sequences consist of many of different repetitive DNA families, classified as class I or II transposable elements (TEs), tandemly repeated DNA in one or several genomic locations, endogenous retroviruses, satellite DNA and simple repetitive DNA (Heslop-Harrison and Schmidt 1998; Biscotti *et al.* 2015). The repetitive DNAs are part of the wider pattern of genetic variation because of their differences in abundance among living organisms (McClintock 1984; Heslop-Harrison *et al.* 1997; Hilbrict *et al.* 2008). These differences make it difficult to analyse and study them but, with the continued development of next generation sequencing technology and the ability to sequence whole genomes, this problem is being overcome. Also, by means of *in situ* hybridization, repetitive DNA sequences can be localized on chromosomes and their characteristic loci visualised (Kubis *et al.* 1998; Heslop-Harrison and Schwarzacher 2011).

One of the largest and most diverse angiosperm families is the Asteraceae. It is believed to be a relatively young family, dating from the mid-Eocene and diversifying in the last 40 My (Jansen and Palmer 1987). The family includes many economically important crop plants, such as lettuce and sunflower, many horticultural plants including *Chrysanthemum*, and a number of invasive weeds with a worldwide distribution, including *Taraxacum* and *Hieracium*. The proportion of repetitive DNA in Asteraceae species genomes has been estimated as 66-71%, slightly higher than in other genomes that have been studied, e.g. *Oryza* 25–66%, *Vitis* 41.4%, *Sorghum* 61%, *Malus* Miller 67% and *Nelumbo* 50% (Rice genome 2005; Paterson *et al.* 2009; Cossu *et al.* 2013; Natali *et al.* 2013).

Taraxacum is distributed worldwide and belongs in subfamily Cichorioideae, tribe Cichorieae (syn. Lactuceae). *Taraxacum* species consist of sexual diploids (2n=16), and polyploids, almost all of which are apomictic (forming viable seed without fertilisation). Triploidy (2n=24) is the most common form of polyploidy (Richards 1973). Such a reproductive system results in the formation and spread of thousands of morphologically different clones.

In a sexually reproducing plant, the source of genetic variation is mostly from recombination and segregation during meiotic cell divisions. This is necessary for their continued evolution. In asexually reproducing plants, such as *Taraxacum* agamospecies, genetic recombination during the meiotic process is avoided so that segregation does not occur and progeny are produced that are genetically homogenous and identical to their mother. Genotypic variation, however, can be found among progeny produced asexually. This phenomenon has been of interest to many scientists and has raised many questions about the mechanism of origin of this variation in an apomictic plant (*Taraxacum* in particular). One of these questions centres on the contribution made by repetitive DNA (especially transposable elements) to the variation observed. So far, there have been many studies of variation in *Taraxacum,* especially in populations with obligate apomixis, including studies by Mogie and Richards 1983; Lyman and Ellstrand 1984; Ford and Richards 1985; Mogie 1985; van Oostrum *et al.* 1985; Hughes and Richards 1988; Majeský *et al.* 2012. The general conclusion is that the genotypic variation in *Taraxacum* microspecies comes from mutation.

So far, most studies of transposable element diversity and evolution have been made on sexually reproducing eukaryotic organisms. A few studies on different animal families, however, eg. Arkhipova and Meselson 2000; Zeyl *et al.* 1996; Goddard *et al.* 2001, have found that asexual taxa may contain less transposable element diversity than sexual taxa.

Many approaches have been used to analyse repetitive DNA sequences including identifing abundant restriction satellite DNA fragments, characterizing clone library sequences and probe with genomic DNA to find abundant clones then hybridization, using degenerated primers for amplification reverse transcriptase (RT)

gene (Chapter 3), and microdissecting regions of chromosomes with repeats/heterochromatin and clone. Alternatively, over the past decade, with the development of next generation sequencing, there have been many different computational programs used to identify and characterize repetitive DNA sequences in the eukaryotic genome, including RepeatExplorer, developed by Novak *et al.* (2010). It is one of the online computational programs and uses a graph-based clustering algorithm on raw read sequences to identify and analyse repetitive sequences in the genome, without the need for reference databases (Novak and Macas 2010). This program is designed to be effective in analysing repetitive DNA components of both plant (Macas *et al.* 2011; Novak *et al.* 2010) and animal (Pagán *et al.* 2012) genomes.

5.2. Aims

In this study, I employed bioinformatic analysis and the repeat clustering method for a comparative analysis of the genomes of three closely related microspecies of *Taraxacum officinale* agg. a total of ~45 Gb genomic DNA sequence data were generated to:

1) Identify repeat composition. I identified and quantified major groups of repetitive DNA family sequences, including transposable elements, in order to understand the nature and consequences of genetic variation between microspecies and the large-scale organization and evolution of an apomictic plant genome.
2) Estimate sequence diversity of repeats and diversity, distribution and abundance of transposable elements between the three related microspecies.
3) Generate bioinformatics resources for the development of repeat-based genome-specific markers.
4) Characterise the genomic organization of graph-based clusters by *in situ* hybridization and get insight into the organization of different cluster sequences on chromosomes. This builds towards identifying repeat families.
5) Analyse the nature of sequences in major clusters.

The role of transposable elements in generating variation in sexual versus agamospecies will be investigated by comparing the results with a sexually reproducing species in the Asteraceae family.

5.3. Materials and Methods

5.3.1. Plant materials and DNA isolations

Three closely related agamospecies (2n=3x=24) of *Taraxacum officinale* agg. [section *Taraxacum* (formerly *Ruderalia*), Asteraceae] were studied.

- *Taraxacum obtusifrons* Markl. (O978)
- *T. stridulum* Trávniček ined. (S3)
- *T. amplum* Markl. (A978)

Seeds were germinated and planted in pots and grown in a growth cabinet under suitable temperatures, daylight and humidity for *Taraxacum*. Genomic DNA was extracted from fresh leaves after removing the midribs, following the Cetyl TrimethylAmmonium Bromide (CTAB) procedure of Doyle and Doyle (1987), with minor modifications, to obtain high-quality DNA. Root tips collections were also made from these plants.

5.3.2. Illumina DNA sequencing

For whole genomic DNA sequencing and results of the sequencing, see Chapter 4 section 4.2.1.

5.3.3. Graph-based Clustering of *Taraxacum* sequences and data analysis

In order to identify repetitive DNA families, graph-based clustering (using RepeatExplorer; Novak *et al.* 2010) was performed on a random subset of *Taraxacum* genomic DNA.

First of all, the raw data of whole DNA sequences from three *Taraxacum* microspecies were processed to paired end sequence by Geneious version 7.1.4 and later (Kearse *et al.* 2012; available from http://www.geneious.com/). Then the paired-end whole genome raw sequences were exported as FASTA files. Clustering was not feasible on the full data set because of computational and size limitation requirements by the program, so the data were split into several subsets containing ca 1.8 Gb size

sequences each by the split command on Ubuntu Linux 13.10. Then, all split files from three *Taraxacum* genomes were subjected separately to bioinformatics analysis to clustering using the RepeatExplorer website (http://repeatexplorer.umbr.cas.cz/), within the Galaxy environment. A graph-based clustering approach was implemented on each of the three *Taraxacum* microspecies datasets in a default mode except checking the option of paired end.

5.3.3.1. Analysing RepeatExplorer outcome files

The main results presented by RepeatExplorer are in HTML format and contain a table listing all clusters, genome proportions of each cluster and similarity hits. Graph clustering resulted in thousands of clusters, however RepeatExplorer HTML files represented only those clusters >= 0.01% of genome proportion. Most of the cluster graphs, with cluster annotation and their build up unassembled reads from the archive files were investigated manually. Clusters that represented high proportions of genomic DNA, clusters with unique pattern, high similarity hits to specific repeats, unclassified clusters, “Low_complexity” and “Simple_repeats” clusters were among the most interested clusters that have been investigated here. Each cluster was consists of a number of contigs, the contigs chosen randomly and sometimes the longer contigs have been chosen.

5.3.3.2. Data analysis

Genome proportions of each repetitive sequence from the annotated clusters were calculated by taking their summation to investigate the total genome proportions of each repetitive DNA family.

Contig sequences were imported to the Geneious program to analyse, design primers, and search for any tandem repetitive DNA using dot plot and Tandem Repeats Finder (Benson 1999), by using the default parameters.

Microsatellite sequences were identified using Tandem Repeats Finder. The microsatellite of di-, tri-, tetra-, penta- and hexanucleotides which lie adjacent to each other was studied, with the minimum number of repeats = 3, excluding the mono-nucleotide repeats. Searching for microsatellite was conducted on some

"simple_repeat" clusters with different genome proportions from O978 genome RepeatExplorer output files. The repeated times of different microsatellite motifs were recorded, and then the results were compared.

5.3.3.3. GC% content analyses

The GC% content of different clusters annotated with different repetitive DNA families were investigated and compared. For this analysis, the raw sequences from archive files of RepeatExplorer results belonged to the chosen clusters were used. The analysis comprises both dinucleotide and trinucleotide frequencies within all the raw reads belonging to one cluster, using online program "genomatix" (http://www.genomatix.de/cgi-bin/tools/tools.pl?s=2aad241fefdd14b228c6272d22271c27;TASK=statistics) to create sequence statistics. A different program was used to generate randomized DNA sequences according to the "genomatix" mononucleotide percentage (Random DNA sequence generator- http://users-birc.au.dk/biopv/php/fabox/random_sequence_generator.php). Then, the Excel program was used to analyse all the data and make a comparison between the results of the two programs, and to compare the chosen clusters.

5.3.4. Primer design and PCR amplification

Primer pairs were designed from contigs related to the clusters in output HTML files for the three *Taraxacum* microspecies. Only those primers generating robust patterns were retained. The list of primer designed from RepeatExplorer HTML output files with their annealing temperature and expected band size are listed in Table 5.1.

Table 5.1. Primer pairs and there sequences from chosen clusters of three *Taraxacum* microspecies.

Genome	Cluster Number	Oligo name	Sequence : (5' to 3')	Tm	Expected product size	Labelled with
A978	1	TaRECL1_238F	TTCCGGCTAGACCTCCTTCC	66.5	1065	Bio
A978	1	TaRECL1_1295R	AGTTCCAAACGAGCTTGCTG	64.4		
A978	1	TaRECL1_440F	TTGTGACAGGTCCTGGACAC	63.9	857	Dig
A978	1	TaRECL1_1295R	AGTTCCAAACGAGCTTGCTG	64.4		
A978	1	TaRECL1_440F	TTGTGACAGGTCCTGGACAC	63.9	635	Bio
A978	1	TaRECL1_1090R	CTTCACCAGTCGCGTGTTTG	66.8		
A978	9	TaRECL9_874F	GGTCCACGTGTATTCCCTCG	66.5	987	Dig
A978	9	TaRECL9_1860R	AAGAGGCAGCACACCAGTAC	62		
A978	3	TaRECL3_64F	TACCAATCTGTCACACCCCC	64.7	643	Dig
A978	3	TaRECL3_706R	ATGGGGCGTTATCACACTCC	65.9		
A978	36	TaRECL36_72F	GGACCTCGTAGTCAGCATCG	64.8	577	Bio
A978	36	TaRECL36_648R	CGCTTACCTCTCCGTCAACC	65.9		
A978	16	TaRECL16_41F	TCAAACCCGACATCAAAACGC	69	314	Bio
A978	16	TaRECL16_354R	AACTCCGTGATGGTGAGACC	64.1		
A978	2	TaRECL2_201F	TCTGCCCCTGTGTTAATCGA	65.5	150	Bio
A978	2	TaRECL2_350R	AATGGGCATAACTTCCAAACGA	66.1		
A978	5	TaRECL5_18F	CTAGCCACACTAGTCAGCCG	62.6	193	Dig
A978	5	TaRECL5_210R	TGGTCGCCGGAAAACACATA	68.4		
A978	19	TaRECL19_7F	GCGGTTCTCAGAGATGAAGCT	64.8	128	Bio
A978	19	TaRECL19_134R	ACCGGGATTTCACCAACAGT	65.4		
A978	4	TaRECL4_15F	GTTGGGTGAGGGTGAGTGAG	64.7	168	Dig
A978	4	TaRECL4_181R	TCTCCTCACTCCCTCACTCG	64.7		
A978	7	TaRECL7_50F	AGAAGCTACCATGCCCATGC	66.3	160	Bio
A978	7	TaRECL7_209R	TGGATGCATGCAAGGAAGGA	69		
A978	40	TaRECL40_82F	CCCTGCATTCCATCATCAAGA	67.1	150	Dig
A978	40	TaRECL40_197R	TGACATTAAGACATGGTTAAACATCT	61.4		
A978	17	TaRECL17_6F	ACCAAATGCTTCCTACTCCTTCT	63	464	Bio
A978	17	TaRECL17_469R	TCTCCGGTTTACAAAAGCTCA	63.5		
A978	65	TaRECL65_18F	TCACTCACCACTCTCACTCTCT	60.7	562	Dig
A978	65	TaRECL65_579R	ACATTTCCATGAACTATCAACCAACA	66		
A978	78	TaRECL78_366F	ACAAAGGCGAAACAGAACAACA	65.6	223	Bio
A978	78	TaRECL78_588R	GTATCAATTGCCAAACCCGCA	68.6		
A978	127	TaRECL127_7F	TCCAAGGGGGAGTGTTGAAA	66.6	149	Bio
A978	127	TaRECL127_155R	TGAGAATAGTGTATCAGTACATCGT	58.3		
A978	128	TaRECL128_150F	CTGCTCTTGCTGCTTCCTCT	63.9	254	Bio
A978	128	TaRECL128_403R	GGTCAAGCCGGGTTAAACCT	66		
A978	154	TaRECL154_15F	TTTGGAGGGCATTTGTCGGT	68.6	188	Bio
A978	154	TaRECL154_202R	ACGGATAAGATTGCAGGTTCT	61.5		
A978	175	TaRECL175_14F	AGGGGGTGTGTGAGTAAGGA	63.4	123	Dig
A978	175	TaRECL175_137R	TTTATATGGTGCGTCGCCGT	67.2		
A978	186	TaRECL186_41F	GAGGAGTGAAGGTGGTGACG	64.8	334	Dig
A978	186	TaRECL186_374R	ACCCGCCAACACATGTTACT	64.2		
A978	206	TaRECL206_1007F	TATTTTGGTGCTCCCGGGTC	67.9	447	Dig
A978	206	TaRECL_1453R	CGCAACTGGAAGTGGTTGTG	66.7		
A978	235	TaRECL235_64F	TCTATCCCGATGAGCCGTCT	65.9	161	Bio
A978	235	TaRECL235_224R	TGGGCTGAGCACCTTTTTCA	68		
A978	236	TaRECL236_36F	GCTGCCTCCCACATCTACTC	63.8	233	Bio
A978	236	TaRECL236_268R	AGCTTGGCCCTTCTCTCCTA	64.2		
A978	250	TaRECL250_21F	TCCAGCTACCAGTTACCAGC	61.9	158	Dig
A978	250	TaRECL250_178R	GTGGCAGGTTCGAGACATCA	66.4		
A978	289	TaRECL289_19F	CACAAGGGCTGCTATCACCA	66.2	245	Dig
A978	289	TaRECL289_263R	GTTTGTCTAACCCGGGTCGT	64.9		

Table 5.1. continue ….

Genome	Cluster Number	Oligo name	Sequence : (5' to 3')	Tm	Expected product size	Labelled with
A978	311	TaRECL311_148F	CGACCGGTTGTATAGGCGAA	66.8	235	Bio
A978	311	TaRECL311_782R	GTGTACGACTCTTGTGCCCA	63.9		
A978	306	TaRECL306_64F	AGGGGAGGAAGTCAATCGGA	67	372	Bio
A978	306	TaRECL306_467R	AAGGGGGTGTTTGGTTGGTT	66.3		
A978	313	TaRECL313_91F	TTATCGGTGAGCGAAGTCCA	65.6	177	Bio
A978	313	TaRECL313_267R	CGAAGGAGCAAGAAAGGAGT	62.4		
A978	11	TaRECL 11_265F	ATCGGGGCGTTACAACTCTC	65.2	309	Bio
A978	11	TaRECL 11_573R	CTACGCGATGGTGTGAGTGT	63.9		
A978	20	TaRECL 20_45F	AGCGGTGGTCTTTGCACTAA	64.6	347	Dig
A978	20	TaRECL 20_347R	TCGCCTAGCACATGTTCTGG	66.3		
A978	64	TaRECL 64_2215F	GATTATGTTGCGGTGCGGAC	67.5	1372	Dig
A978	64	TaRECL 64_3568R	CACAATTGCGGTGGCTTGAA	69.2		
A978	46	TaRECL 46_115F	ATGCCCCAGCTCGAATTGAA	68.8	362	Bio
A978	46	TaRECL 46_476R	ATCGTTGACAGACTGCGGTT	64.7		
A978	116	TaRECL 116 _2F	GCCAGGAAACAAGGAGCCTT	66.6	155	Dig
A978	116	TaRECL 116_156R	ACGTCCACCGGTAACTAGGA	63.3		
A978	83	TaRECL 83_261F	GGTACCAAAGAAGTCGGGCA	66.1	632	Bio
A978	83	TaRECL 83_892R	CAAGGAAGGAACGGATGCCT	67.4		
A978	135	TaRECL 135_39F	TTTGCTGGTGTTTTGAGCCG	68.2	299	Bio
A978	135	TaRECL 135_337R	GGTGACATGGACTGAAGCCA	66.3		
A978	83	TaRECL 83_292F	AAGAATGGTGGACGGACGAC	66.2	239	Dig
A978	83	TaRECL 83_530R	ATTCGCAAAAAGTCCGCACC	68.1		
A978	129	TaRECL 129_77F	TGCACTCTCGTTTACACTCTAGT	59.7	412	Dig
A978	129	TaRECL 129_488R	GTGACTCGAGTGCAAAACGT	63		
A978	171	TaRECL 171_68F	TACTCCCTCCGTCCCAAATT	64.4	215	Dig
A978	171	TaRECL 171_282R	CGTTTTTGTCCATCATTGACCG	68.2		
O978	1	Tx_O978_CL1_1F	CAAATTCCCAACACGACG	62.7	86	Bio
O978	1	Tx_O978_CL1_86R	CAGACTGCATCCGATATTTATC	60		
O978	27	Tx_O978_CL27_156F	CTGTGTTATTTGATTCGGGC	61.9	317	Dig
O978	27	Tx_O978_CL27_472R	GATTAACTTGTCGCCGTTAG	58.9		
O978	118	Tx_O978_CL118_1301F	ATCGATCTATTATGCGCCTC	60.6	920	Bio
O978	118	Tx_O978_CL118_2220R	TGGTGTATAGTTCATCAGCG	58.8		
O978	38	Tx_O978_CL38_104F	TTTGAGAACAGAGACCGAAG	59.6	559	Bio
O978	38	Tx_O978_CL38_662R	ATGAACCATCAGACACTCAC	57.4		
O978	49	Tx_O978_CL49_42F	TGAGGAGGATCTAGTGGAG	56.7	259	Bio
O978	49	Tx_O978_CL49300R	GGTGTAGCGAAGAGGATATG	58.8		
O978	68	Tx_O978_CL68_81F	AGCCGTGAAAAAGAATCAAC	60.7	105	Dig
O978	68	Tx_O978_CL68_185R	TCACGTCTTTCTTTAACAGC	57.2		
O978	94	Tx_O978_CL94_101F	GAGTGTAAAACGAGTGTACG	54.8	257	Dig
O978	94	Tx_O978_CL94_357R	ATGCACTCTCGTTTACACTC	56.9		
O978	89	Tx_O978_CL89)220F	TAAATCATGGCCGGTCATAG	62.2	273	Dig
O978	89	Tx_O978_CL89492R	GCATTTGTGCTGGTATGATC	60.6		
O978	96	Tx_O978_CL96_9F	CGGGGGATTTGTGTAAATTG	63.7	131	Dig
O978	96	Tx_O978_CL96_139R	CGATAAGCTCCATTTGCATC	62.1		
O978	108	Tx_O978_CL108_2221F	AGGGATCAATTTCTAGTCGC	59.7	910	Dig
O978	108	Tx_O978_CL108_3130R	AATGGATTGCTACCTATGCC	60.4		
O978	164	Tx_O978_CL164_20F	TAAAAGCTTGATCATGTGCG	61.4	230	Bio
O978	164	Tx_O978_CL164_249R	TCAAAGTACAAGGTATGGGC	59.1		
O978	281	Tx_O978_CL281_153F	TGTGACTGGCTATTAAAGGG	59.2	319	Bio
O978	281	Tx_O978_CL281_471R	TGTGACTGGCTATTAAAGGG	59.2		
O978	172	Tx_O978_CL172_71F	ACCAAGGACAGAAGAGAACTCC	62.6	80	Bio
O978	172	Tx_O978_CL172_150R	CTTGAACTTCACTCGGCAGC	65.1		

Table 5.1. continue…

Genome	Cluster Number	Oligo name	Sequence : (5' to 3')	Tm	Expected product size	Labelled with
O978	214	Tx_O978_CL214_4F	AATCCAGCCATCACTATAGC	58.2	328	Dig
O978	214	Tx_O978_CL214_331R	AGCTTAGAATCCAATGTTGC	58.4		
O978	223	Tx_O978_CL223_347F	ATGAAGCCTTACAGAAAACC	57.4	368	Bio
O978	223	Tx_O978_CL223_714R	CTTTTTGGCTCTTTGGTTGG	63.4		
O978	225	Tx_O978_CL225_555F	CTTCTTCTGAAACTCCAAGC	58.2	994	Dig
O978	225	Tx_O978_CL225_1548R	ATATGTTACAAATGCGGTCG	59.9		
O978	224	Tx_O978_CL224_1422F	ACGAAAGATGGATTAGAGGC	59.7	1159	Bio
O978	224	Tx_O978_CL224_2580R	GTTTGGTCCATGTGAAATCC	61.7		
O978	259	Tx_O978_CL259_103F	AAGTTATCCAGAGTGTTTGG	55.7	104	Bio
O978	259	Tx_O978_CL206R	GTAGACATTTAAATTGGGACGG	61		
O978	75	Tx_O978_CL75_68F	CACCCGAAATACTATCCCTG	60.8	340	Dig
O978	75	Tx_O978_CL75_407R	TGTTGAAGCTCTGTTTTTGC	60.6		
O978	132	Tx_O978_CL132_214F	ATATGACGTGATGATGTGGG	60.7	500	Bio
O978	132	Tx_O978_CL132_713R	ACTGTGTATCGCTCGTAATC	56.8		
O978	93	Tx_O978_CL93_1240F	GAAACAAAAACGGCAATTGG	64	632	Dig
O978	93	Tx_O978_CL93_1858R	CTGGGTGGTCTATTTATGGG	60.6		
O978	38	Tx_O978_CL38_102F	TAACTAGTACAAAAGGTGGC	53.6	425	Dig
O978	38	Tx_O978_CL38_526R	GATTATCCCGACCCTTATGC	62		
S3	2	TRECL2F82	ATGTGCGGGTAAGAATCGGG	67.9	486	Dig
S3	2	TRECLL2R568	TCTTGACTCCCCCGGTCTAG	65.4		
S3	8	TsRECL8_182F	GTTTTGGCCACTTTAAAGGG	62.1	573	Bio
S3	8	TsRECL8_754R	AGTTCCTGTAGCTAAAACCG	57.6		
S3	213	TsRECL213_486F	CTACCATTTTTGACGCTTCC	61.1	621	Dig
S3	213	TsRECL213_1106R	CATCTCCATTGTCACTTTGC	60.8		
S3	148	TsRECL148_2F	GAAAACGAATCGACATCTCG	62.2	599	Bio
S3	148	TsRECL148_600R	AAGAGTAAGGATTGAAGCCC	58.8		
S3	130	TsRECL130_321F	GCGAACCAAAAACATTTTCC	63.1	608	Dig
S3	130	TsRECL130_928R	TTGTTCAATGCTGTTGTTCC	61.2		
S3	33	TsRECL33_598F	ACGATTAAGAAAGGTGGACC	59.5	1006	Dig
S3	33	TsRECL33_1603R	ATCCTAGTACTACTCTCGCC	54.2		
S3	96	TsRECL96_251F	TCCCTTCAACACAACATAGG	60	971	Bio
S3	96	TsRECL96_1221R	TCTTTTGTCTCATGGTAGCC	59.3		
S3	95	TsRECL95_2041F	AGGAGGGATCCTAGAATTCC	59.9	1118	Dig
S3	95	TsRECL95_3158R	CCCCATTTGAATGATTCACG	65		
S3	102	TsRECL102_924F	TACGATTTTCCATGCTTTGC	62.5	1921	Bio
S3	102	TsRECL102_2844R	ATGTCTTTCTATGGGATCGC	60.5		
S3	100	TsRECL100_13F	ACGAGATTAAAGTGTGACTGG	58.3	190	Bio
S3	100	TsRECL100_202R	TTTGTCCAACCTGTAGATCG	60.2		
S3	105	TsRECL105_52F	CTACACTCTTTCCCTACACG	56.4	720	Dig
S3	105	TsRECL105_1299F	ACAGTTCGAACGATCTATGG	59.2		
S3	105	TsRECL105_771R	ACCAAAGAGAATGATGACCC	60.4	1609	Bio
S3	105	TsRECL105_2967R	TGAATTGAGCTCGTATAGGC	59.5		
S3	80	RETCL80fiftymerFluor	TTAATGTGTTGATTGTGACATT AAGACATGGCTAAAACCCTATAAGATCT	76.2	-	Flc (green)

5.4. Results

5.4.1. Repetitive DNA in *Taraxacum* microspecies from RepeatExplorer

A subset of raw sequence reads of the three *Taraxacum* microspecies was used for graph-based sequence cluster analysis using RepeatExplorer. The genomic sequences represent 21x coverage for the S3 haploid genome and 10.4x, 12.2x for A978 and O978 respectively. More information about genome sequences and reads information is given in Table 5.2a.

Overall, results were consistent among the three microspecies (Figure 5.1), and RepeatExplorer results showed that the proportion of repetitive DNA in three *Taraxacum* microspecies was very similar. Each had about 300 clusters of sequences representing more than 0.01% of the genome, representing about 55% of all reads. Rather, the 150bp and 300bp read lengths altered numbers of sequences put into single clusters, and the analysis was sensitive to cutoff value for connection of clusters. Thus the 300bp reads of the S3 genome reported a greater number of larger clusters than in A978 and O978 (Figure 5.1) and single-copy/unclustered sequences in S3 genome were smaller than A978 and O978 genome, 0.4% in S3, 23% in A978 and 30% in O978 genome as might be expected because the longer reads include low-copy DNA domains which become clustered with adjacent repetitive DNA (Figure 5.1). Because of sensitivity to graph clustering parameters, cluster fragmentation may give multiple linear graphs. For example, RT and INT domains of a retroelements may be in a single cluster or separated into two clusters.

Additional analyses were conducted to understand the RepeatExplorer outcomes and investigate the results in three genomes and their correlation with sequence read lengths. I took 150 bp reads (first half and middle section) of the 300bp S3 reads for RepeatExplorer from the results provided from this analysis. There were no significant differences between S3 300 bp and first and middle section of S3 (150 bp) (Figure 5.2).

Table 5.2. (a) Cytogenetics, genomic and sequencing features of the *Taraxacum* microspecies. (b) RepeatExplorer results out-come.

(a) **Species**	***T. stridulum* (S3)**	***T. amplum* (A978)**	***T. obtusifrons* (O978)**
Chromosome number	2n=3x=24	2n=3x=24	2n=3x=24
1CX value (pg)	0.87	0.87	0.87
Haploid genome size (kb)	850,860	850,860	850,860
Whole genome sequenced read number	59,258,642	58,713854	69,056,774
Whole genome sequence (GC%)	37.9%	37.1%	37.5%
Sequenced read length (bp)	300	150	150
Illumina coverage	20.9x	10.4x	12.2x
Chloroplast number of reads present in the whole genome	4,452,141 (7.5%)	2,108,482 (3.6%)	2,739,214 (3.97%)
Size of the genome used to upload to RepeatExplorer (Gb)	1.791	1.791	1.791
Number (%) of chloroplast reads including the upload raw-read file	377,184 (7.5%)	355,491 (4.04%)	318,409 (3.6%)

(b)	**N0. of uploaded reads**	**No. of used reads in clustering**	**% of No. of used reads in clustering**	**Cluster number (genome proportion >= 0.01)**	**total number of similarity hits**
O978	8,844,500	1,827,812	20.7	318	368,380,374
A978	8,800,467	3,396,038	38.6	287	403456478
S3	5,000,000	3,505,488	93.2	291	687438086
S3_First	5,000,000	4030760	81	251	424,021,979
S3_middle	5,000,000	4861492	97	344	582,670,370
S3_chunks	5,000,000	5000,000	100	535	7,648,973

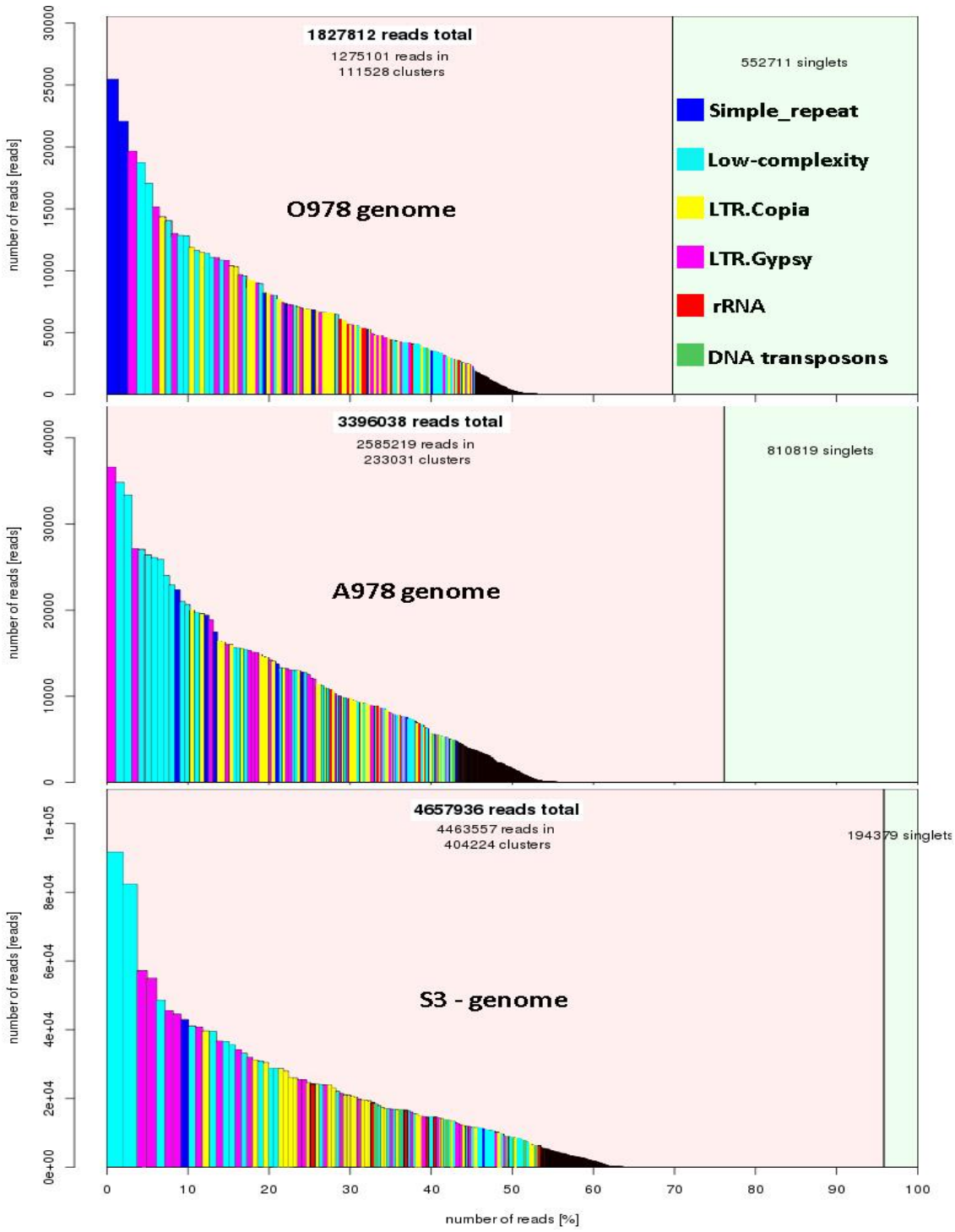


Figure 5.1. Distribution of clusters based on three *Taraxacum* genomes by size and class of repetitive element histograms, shows the result of clustered using RepeatExplorer. (A) *T. stridulum* (S3); (B) *T. amplum* (A978); (C) *T. obtusifrons* (O978). Each bar in the histograms shows the individual size (height) of each cluster and the size relative to the sampled genome (width). The Y-axis shows both the percentage of the reads and number of reads in the clusters and the X-axis shows their cumulative content. Moreover, single-copy and unclustered sequences are reflected to the right of the vertical bar. Bars are coloured according to the type of repeat present in the cluster, as determined by the similarity search.

Table 5.3: Repetitive DNA composition of the three *Taraxacum* microspecies.

Repeat type	Super family	Family	A978	O978	S3	S3-First 150bp	S3 Middle 150bp	S3 Chunk 150bp
Class I (retrotransposons)								
LTR	**gypsy**	Ogre/Tat	1.302	1.060	1.541	1.710	1.541	0
		Chromovirus	6.283	6.635	10.505	7.092	7.082	0.801
		Athila	1.096	0.938	1.834	1.273	1.302	0
		Unclassified gypsy	1.922	1.65	1.084	4.917	3.622	0.687
		Total gypsy	**10.603**	**10.283**	**14.964**	**14.992**	**13.547**	**1.488**
	copia	Tork	1.042	0.891	1.276	1.088	1.002	0
		TAR	0.360	0.253	0.412	0.590	0.451	0
		Maximus/SIRE	7.319	8.290	8.720	8.324	8.020	0.198
		Ivana/Oryco	0.145	0.094	0.485	0.166	0.195	0.058
		Bianca	1.637	1.732	1.541	1.511	1.520	0.341
		Angela	1.902	1.372	1.733	1.445	1.194	0.114
		AleII	0.149	0.224	0.251	1.057	0.425	0.029
		AleI-Retrofit	0.408	0	0.341	0.347	0.499	0.026
		Unclassified copia	0.755	1.751	1.634	0.871	2.063	0.920
		Total copia	**13.717**	**14.607**	**16.393**	**15.399**	**15.369**	**1.686**
	Caulimovirus/PARA-RT		0.159	0.278	0.261	0.189	0.190	0.012
	Unclassified LTR		0.074	0.074	0	0.075	0.077	0
Non LTR	**LINE**	L1	0.046	0	0.211	0.129	0.095	0.141
Class II (DNA transposons)-Subclass 1								
TIR	**hAT**	AC	0.178	0.251	0.485	0.414	0.356	0.012
		Tip100	0.167	0.047	0.496	0.221	0.446	0.066
		Tag1	0.105	0.033	0.058	0.055	0.0146	0.018
	PIF_Harbinger		0.255	0.182	0.467	0.354	0.449	0
	CMC-EnSpm		0.401	0.336	1.058	1.179	1.061	0.119
	MULE-MuDR		0.062	0.067	0.605	0.463	0.208	0.066
	TcMar-Stowaway		0.086	0.013	0.156	0.167	0.197	0
	Unclassified DNA		0.015	0.027	0.018	0.036	0.040	0
Class II (DNA transposons)-Subclass 2								
Helitron			0.280	0.172	0.473	0.337	0.222	0.026
Total DNA transposons			**1.269**	**0.956**	**3.343**	**2.889**	**2.903**	**0.281**
Satellites			0.223	0.227	0.518	0.523	0.515	0.337
Unclassified (Low complexity)			21.628	17.164	22.513	20.715	23.218	1.687
Unclassified (Simple repeats)			3.785	4.685	1.431	2.617	3.348	0.251
Unclassified			0	0.468	0	0.844	0.298	2.564
45S rDNA			1.0	1.9	1.2	1.879	1.831	0.645
5S rDNA			0.2	0.2	0.5			
Total Repetitive DNA			**52.876**	**50.712**	**61.778**	**60.588**	**61.613**	**9.118**

Further analyses were investigated by taking multiple 'chunks' of the reads (semi-randomization) and uploaded to RepeatExplorer, which resulted in fewer repeat (19% of the repetitive DNA contents of the genome) and much less well identified clusters which there was dramatically decrease in each repetitive DNA portions, and increase in unclassified repeat, this could be because of RepeatExplorer collected the reads to build up the clusters but could not annotate them to their repetitive types because of very low sequence similarity hits ((Supplementary Figure 5.1, Table 5.3).

Nevertheless, uploading complete randomization of reads (shuffled raw reads) to RepeatExplorer resulted in reporting the program with an error as no sequences are clustered, this might be due to the absent of similarity hits between the sequenced reads (Supplementary Figure 5.2).

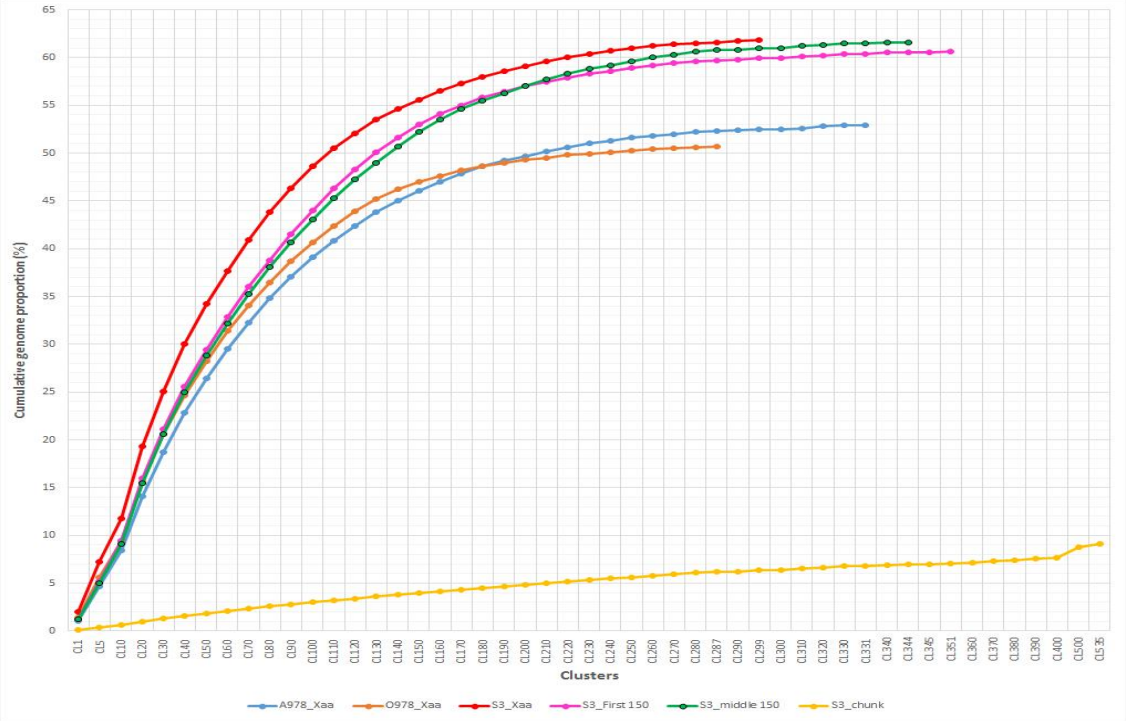


Figure 5.2. RepeatExplorer analysis of repeats output. The graph show differences in cumulative genome proportion.

The graph based clustering analysis resulted in considerable differences in the representation of the major families of repetitive elements in the genomes of the three microspecies. Figure 5.3 shows the ten largest clusters generated by RepeatExplorer, their identity and genome percentage, each representing between 1.8 and 0.9% of the DNA. Top 10 clusters from the three *Taraxacum* microspecies output files showed that about half the clusters, included retrotransposon protein domains but no notable repetitive DNA sequence features were identified in the remaining clusters (Figure 5.3). As shown, most of the top ten clusters blasted with "Low complexity" or "Simple repeat" and all three microspecies show that the most abundant clusters consist of low-complexity and simple-repeat DNA. However, the most striking difference is the O978 genome, which contrasts to the A978 and S3 genomes: Figure 5.3 shows that in O978 genomes the first two clusters are annotated as “simple_repeat” DNA, containing about 2.6 % of the reads used in the graph based clustering. In A978 and S3 genomes, however, the first two clusters represent low-complexity and contain about 2% and 5% respectively.

The “cut-off” feature presented in RepeatExplorer tends to connect different clusters into a new graph according to the similarity and number of mates shared between them. Through the connection of clusters, RepeatExplorer makes a new graph to show how related clusters are connecting, and only groups with more than two clusters are shown in the new graph (Supplementary Figure 5.3). So by this means, I could group the cluster results from RepeatExplorer outcomes into several groups to study the characteristics of the repetitive DNA and characterise the cluster shapes of each repetitive DNA family. The default cut-off value for the association of raw reads into a single cluster is 0.1. However, lower or higher cut-off values can be performed by RepeatExplorer. Analysis of the results from a cut-off value of 0.2 showed that all the clusters grouped have similar RepeatExplorer graph shapes with almost the same repetitive DNA annotations, suggesting that the sequences are indeed related. The results from the cut-off value of 0.1 showed that most of the low-complexity clusters were grouped together into two groups: a) the first group consisted of 23 clusters and showed that almost all of their patterns are linear-shaped (supplementary Figure 5.3 group 1); b) the second low-complexity group cluster graphs were of different shapes

with a lot off diffused line (supplementary Figure 5.3 group 8). c) the third groups were the LTR-Gypsy, which grouped into four groups with different lineages (domains) by grouping from 15 to 3 clusters, most of the clusters have the star shaped or sometimes linear with colour coded domains, (supplementary Figure 5.3 group 2, 7, 10, 13). d) another one of cut-off groups is LTR-Copia retroelements, grouped to six cut-off groups, each group represented from 9 to 3 clusters with different copia domain lineages for repetitive DNA (supplementary Figure 5.3 Group 3, 4, 6, 9, 11, 12). Additionally, the rDNA (45S rDNA) grouped in one group (supplementary Figure 5.3 Group 5.

5.4.2. Estimation ratio of LTR and Non LTR-Repetitive DNA composition in *Taraxacum* microspecies genomes

Clusters may be representative of a particular transposable element family and every contig is a possible consensus for a repetitive DNA family or subfamily or part of one. RepeatExplorer output suggests automated annotations for each cluster based on homology to known repetitive elements (e.g. rDNA sequences) or protein domains (e.g. transposon domains), also including low-complexity (often AT rich) and simple-sequence motifs. The graph (pattern) shapes, clustered sequences and often raw reads for each cluster were checked manually, and annotations used to group the sequence classes.

From the results, the genome proportions of each type of repetitive DNA in the three microspecies, including various groups of LTR retrotransposons, did not exceed a few percent of the genome, the genome proportions of each type of repetitive DNA and comparison among three *Taraxacum* accession genomes are shown in Table 5.3 and Figure 5.4.

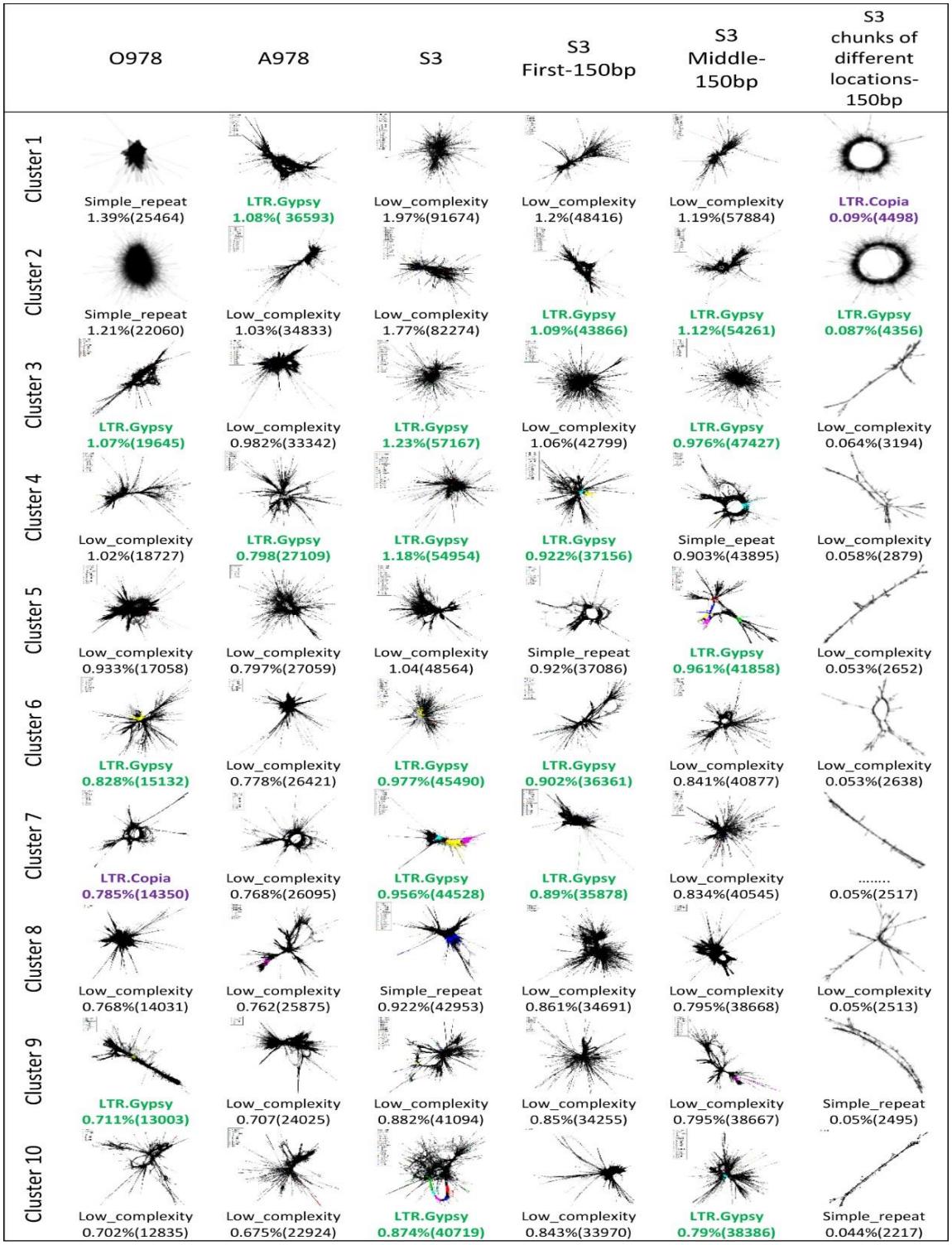


Figure 5.3: Graph based clustering analysis of repetitive elements in the three Taraxacum microspecies. Graph layouts of the ten top and largest clusters of repetitive elements detected in the graph based clustering analysis. Clusters are ordered by size, with largest at the top. Below each graph layout is the class of the repetitive element, the genome percentage of each cluster and number of paired reads belonging to it in parentheses. Coloured regions in some graphs represent conserved domains identified by RepeatExplorer.

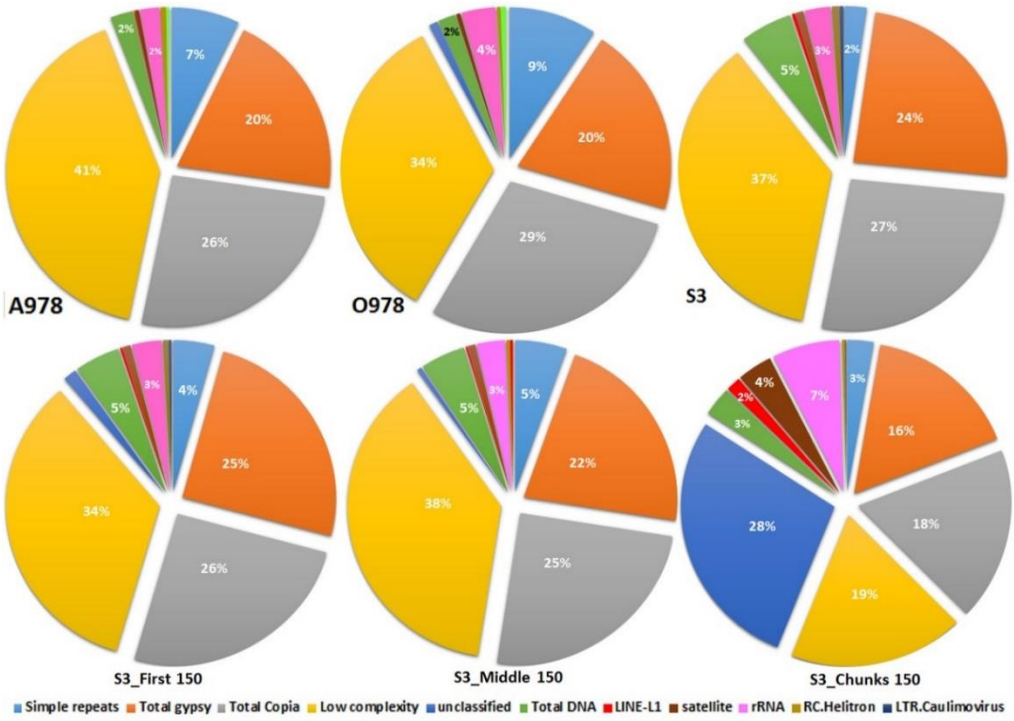


Figure 5.4A. Comparative analyses of the repetitive fraction of three *Taraxacum* genomes.

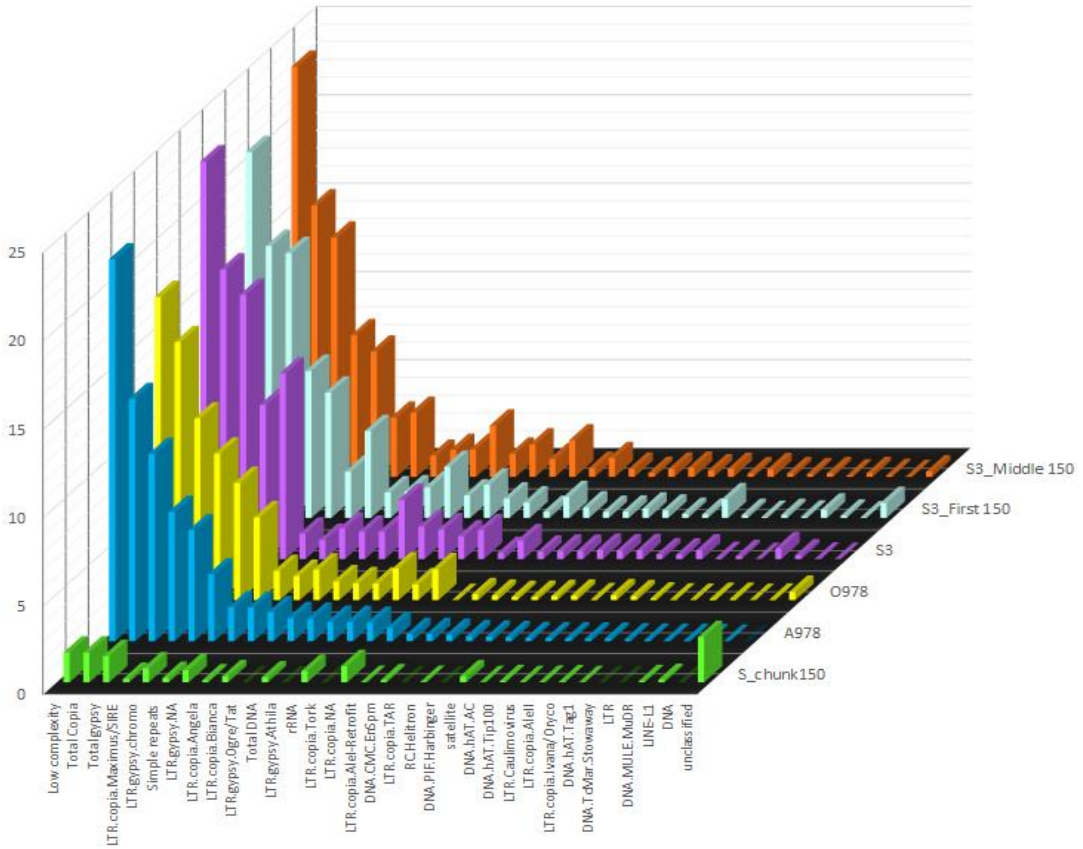


Figure 5.4b. Differences in repetitive DNA proportions in each genome analyzed.

The most abundant DNA sequences found in the three microspecies of *Taraxacum* genome were LTR-retrotransposons, forming 47-51% of the total repetitive genome. They were composed of 26-29% Ty1-copia, represented by eight distinct evolutionary lineages (Figure 5.4 A), of which Maximus/SIRE was the most abundant, representing 15% of total repetitive genome. The remaining Ty1/copia elements belonged to Tork, TAR, Ivana, Bianca, Angela, SleII and AleI-Retrofit, and represented about 13% of the total repetitive genome (Figure 5.4 B). Ty3-gypsy elements represented 20-24% of total repetitive genome (Figure 5.4 A), and belonged to three evolutionary lineages. The most abundant lineage is chromoviruses, comprising about 12-17% of the total repetitive genome. Two other lineages, Ogre/Tat and Athila, represented 5-6% of the total repetitive genome (Figure 5.4 B).

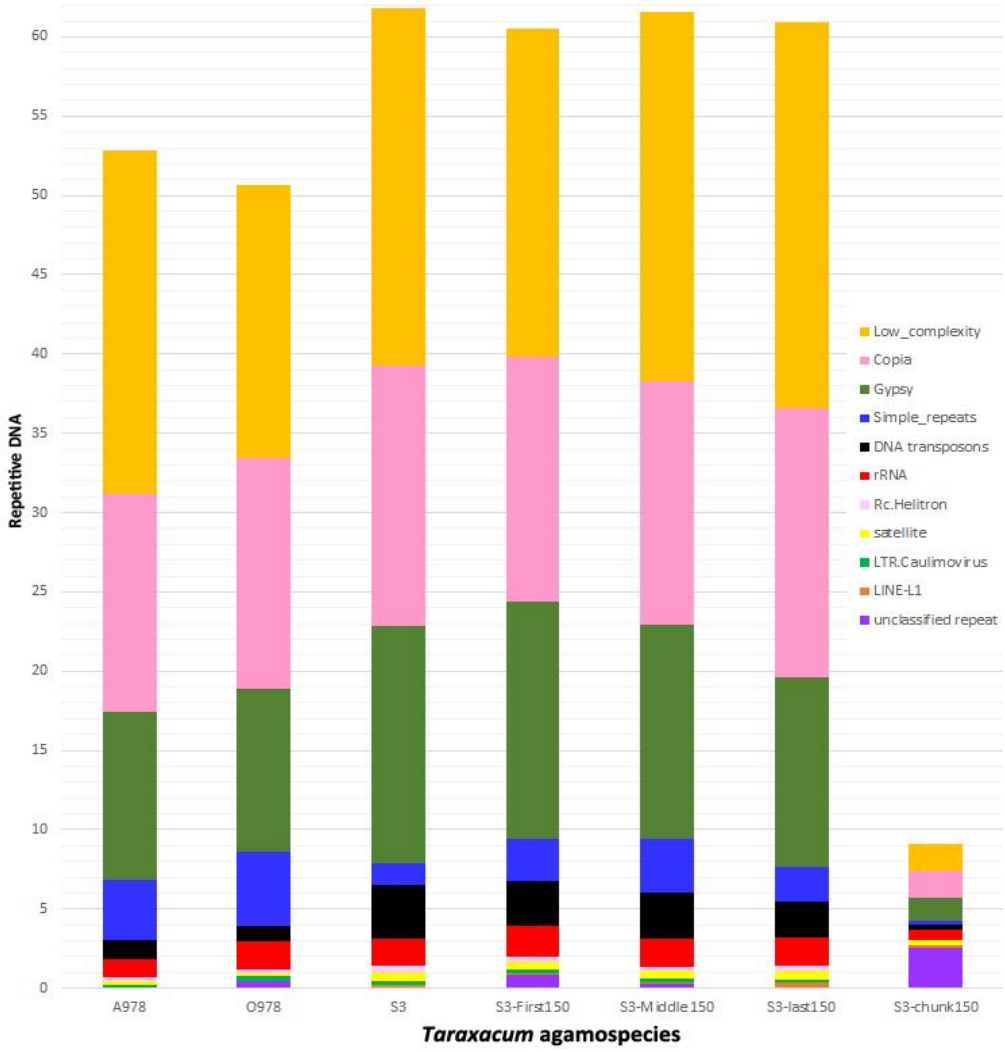


Figure 5.4 C. Shows the differences in repetitive DNA components in each genome analyzed.

Another type abundant DNA sequences clustered by graph based clustering annotated as 'low-complexity' and 'simple repeat' motifs, because of the AT-rich nature of the genome, they represent together about 39-48% of total repetitive genome.

Compared to LTR-retrotransposons, non-LTR retrotransposons and DNA transposons were found to be relatively rare (Figure 5.1, 4 and Table 5.3). LINE elements has not identified in the top 200 clusters in A978 and S3 genome, and it was not clustered in O978 genome, it was estimated to constitute about 0.1-0.2% of the *Taraxacum* genome and were almost undetectable (Figure 5.4). In addition, the top 50 clusters in the S3 genome and top 100 clusters of the A978 and O978 genomes were not annotated with DNA transposons. They were estimated to constitute about 2-5% of the *Taraxacum* genome (Table 5.3 and Figure 5.4).

Clusters containing 45S rDNA represented 2-3% of total repetitive genome. The 45S rDNA sequences consist of the rRNA locus (18S, ITS I and II, 5.8S, and 26S) surrounded by parts of the intergenic spacer (IGS). They separated into multiple clusters, each representing a separate region of the rDNA gene (Figure 5.5), i.e. the 45S rDNA is divided among four clusters in A978 and S3 and seven clusters in the O978 genome. In contrast, 5S rDNA was represented by one cluster in each HTML file for the three *Taraxacum* microspecies. In most of the cases the cluster annotated as a satellite. 5S rDNA represents up to 0.5% of the genome (Table 5.3).

The few clusters that annotated with "satellite" after manualy analysing the DNA sequences of these clusters they were blast with rDNA 5S and not satellite DNA sequences.

Telomeres are not usually clustered by RepeatExplorer in clusters that have a genome proportion of more than 0.01%, but it can found that RepeatExplorer output files contain reads within some clusters that are entirely the classic telomere repeat, i.e. CCCTAAA or in reverse direction TTTAGGG. The telomer-containing reads showed that telomere sequence motifs are repeated up to 8-9 times in the sequence. In several cases, the telomere sequence ends with poly A/G nucleotides.

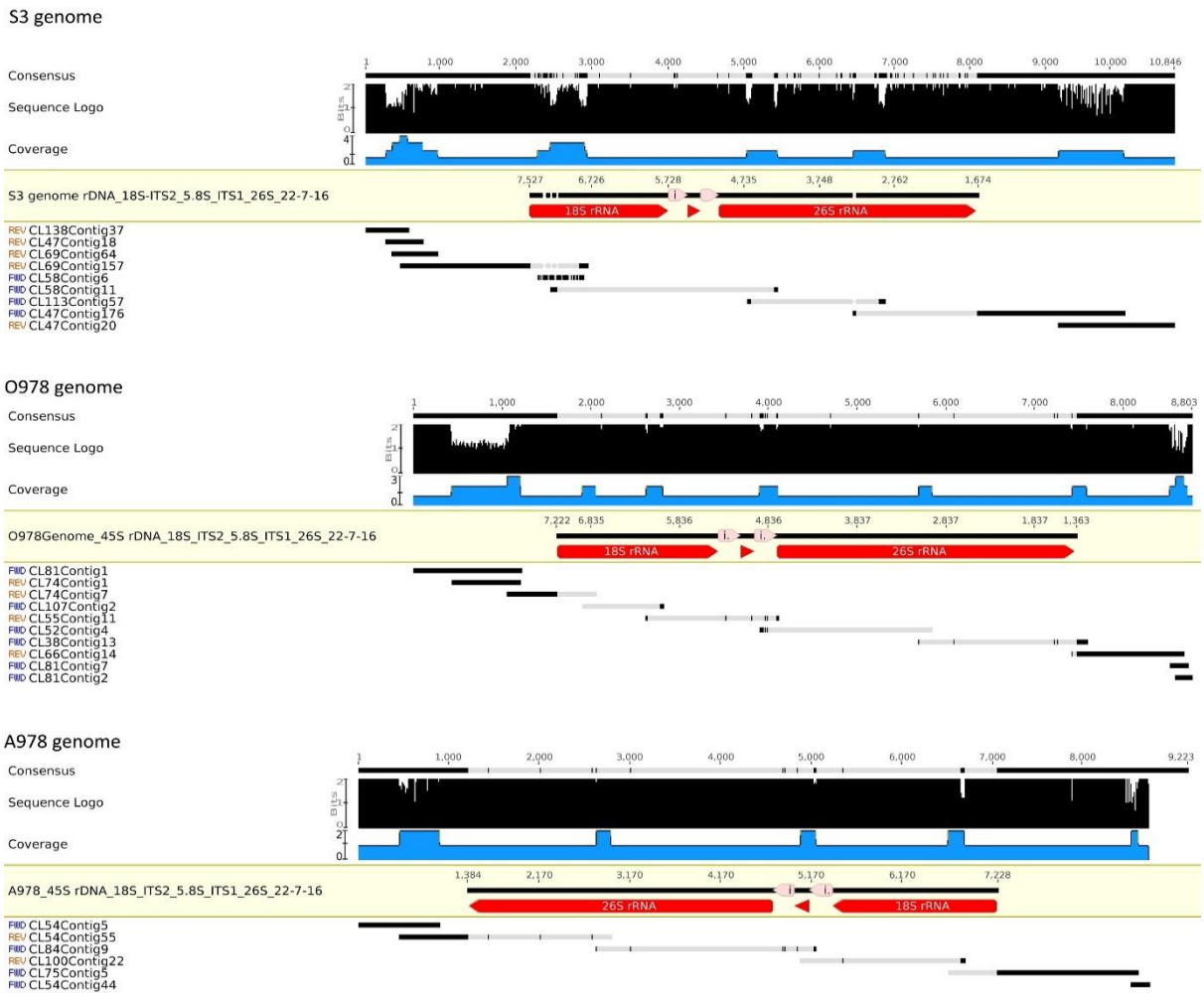


Figure 5.5. Alignment of the cluster contig sequences to complete sequences of the 45S rDNA.

5.4.3. Characteristic of “simple_repeat” and “low_complexity” clusters

Low-complexity and Simple-repeat together were the largest genomic proportion of analysed *Taraxacum* microspecies by RepeatExplorer. The CL1, CL2 in O978 and CL11, CL16 in the A978 genome output files, with almost same cluster graph, are among the largest clusters of the top 10 clusters with largest genomic proportions, they represent 2.6% in O978 genome and 1.227% in A978 genome with approximately 60% similarity hits to simple-repeat, however, the other simple_repeat clusters does not exceeds the 8% of similarity hits to simple_repeat. So further investigation took place to analyse what the characteristics of these clusters are. I looked at sequence reads of these

clusters in A978 and O978 genomes that annotated with simple repetitive DNA and realised these clusters mostly consist of poly-G with some poly-A sequences.

Further simple-repeat clusters were analysed with a different percentage of similarity hits to simple repetitive DNA, such as CL28 (2.18% hits) and CL34 (0.68% hits) and CL206 with highest similarity hits (7.95% hits) to simple repeat. In the O978 genome, all the contigs and their raw read sequences were checked manually. The tandem repeat finder *program (*Benson 1999*) was* used to analyse contigs sequences for identifications of the microsatellite repeats with repeat units ranging from 2 to 6 bp motifs. This analyse resulted in many different microsatellite repetitive DNA build up these clusters, including 2-mer repeat (AT), (GT), (GA), (CG); 3-mer (AAC) (AAG) (AAT), (AGG), (CAC), (CAT), (CAG), (ACT), (ACG), (GCC); 4-mer (AGAA), (AATA), (GATA), (ACCC, (ATAT); 5-mer (ACCAC), (CACAT), (CCACG); and 6-mer (AACAGC), (AGAGCC), (CAACAG), (GGCGGT). Also, all converting canonical and different directions were checked and discounting mononucleotides of CCC and AAA. Therefore, the results show that the most abundant microsatellite motifs are as dinucleotides AT and GA, and ACC for trinucleotides (Figure 5.6). The CL206 in O978 genomes, with genomic proportions of 0.025% to simple_repetitive DNA, showed that it mostly consists of repeated 2-mer sequences: (CA), (TG) and (TA) motifs and poly A/T sequences. Table 5.3 and Figure 5.3 show most of the repetitive proportions affected by shortening the reads length and they dramatically decreased but the ratio of unclassified repeats decreased.

Furthermore, analysis of the dinucleotide and trinucleotide frequencies within the raw reads (S3 genome as an example) gave a sample of these uncharacterized (low complexity, unclassified) clusters, these results have compared with random sequences with the same AT percentage. Results from the comaprisons showed the evidence that all the unannotated clusters in top 10 had non-random sequences, with a tendency for A rarely followed by nucleotide T in low-complexity and C rarely followed by G in simple-repeat in dinucleotide, and A to rarely followed by TA in low-complexity and C rarely followed by TA in simple-repeat in trinucleotides. Possibly also include ‘randomized’ sequence result in analysis.

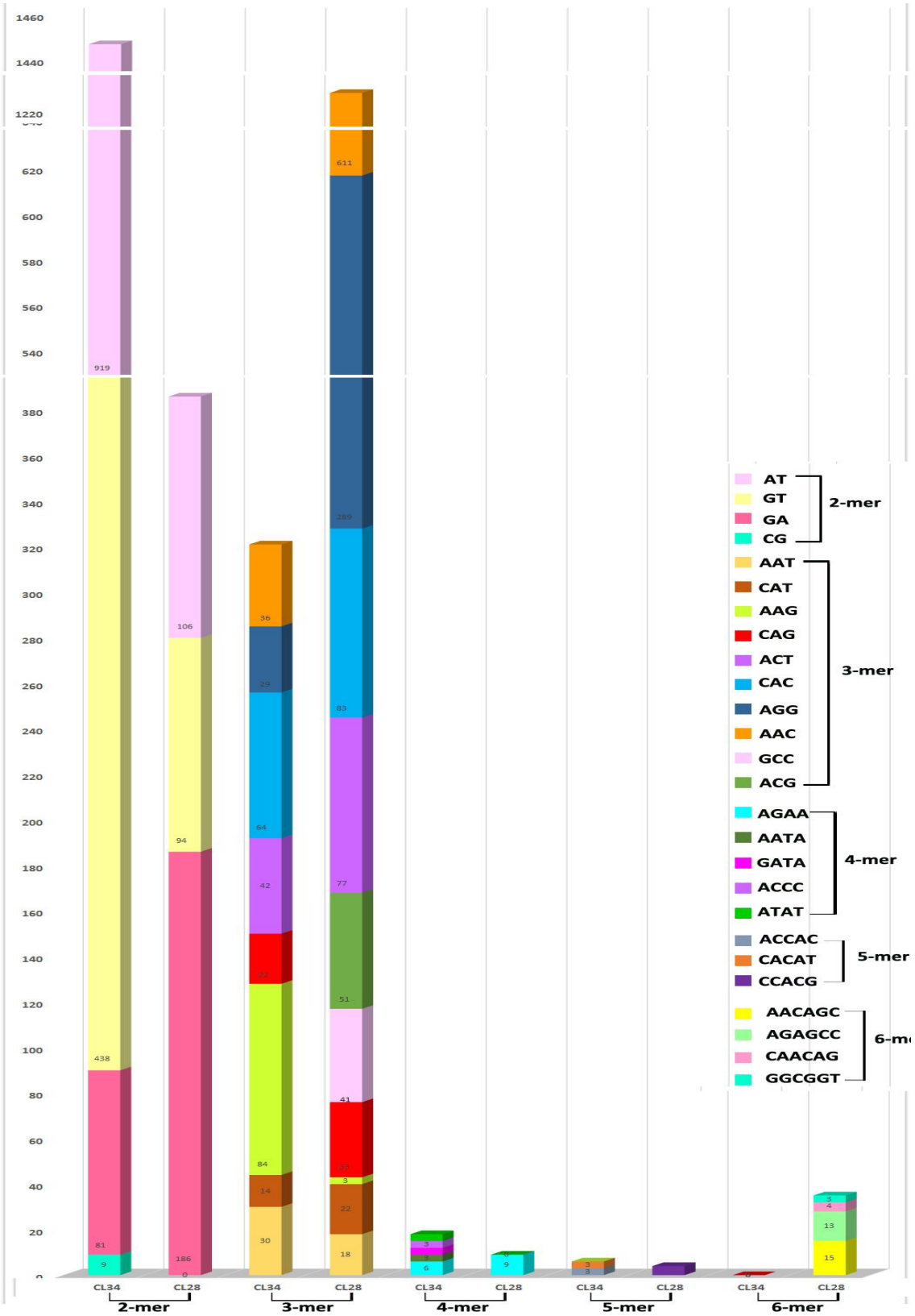


Figure 5.6. Frequency of di-, tri-, tetra-, penta- and hexaploid microsatellites in the simple-repeat clusters of CL28 and CL34 of *Taraxacum* (O978) genome outcomes from RepeatExplorer.

As a comparator to the low complexity sequences, the same analysis showed a different deviation from the random expectation for a gypsy and copia element (where C followed by T or G and C followed by TA were in excess), (Table 5.4 and analysis in supplementary Table 5.1). Further, the results showed that there are differences between AT-rich of the clusters and the whole genome sequences, as the GC% of the low-complexity and the simple-repeat cluster is less than 40%, and for the LTR-copia and LTR-gypsy clusters it is reached 42% and 45S rDNA has the highest GC content of 51% (Table 5.4).

5.4.4. Identification and characterization of repetitive DNA by cluster shapes

RepeatExplorer output files resulted in the identification of repetitive DNA by similarity hits of the cluster sequences against the pre-existing dataset in the RepeatExplorer database of known repeats, which made up the graph pattern according to the collection of strings of DNA sequence motifs for each cluster.

Supplementary Figure 5.4 shows the most usual shapes for some of the repetitive DNA annotated in current study. The Ty3-gypsy repetitive DNA shows graph shapes that are usually either star-shaped or linear, with some being diffuse-ray-shaped, and some (probably the younger gypsy repetitive DNA) appear circular. The graph patterns in clusters with a high percentage of similarity hits to LTR-gypsy, on the graphs the repetitive domains colour coded and in their order of Protease (PROT) - Reverse Transcriptase (RT) - RNaseH (RH) –Integrase (INT). Also, we can usually see circular or sometimes linear shapes with the repetitive domain colour codes for Ty1-Copia elements, with the retroelement coding domains in the order PROT - INT - RT - RH.

Likewise, when the HTML output files were checked manually, special shapes for other repetitive DNA were detected. This was especially true when I compared them in the three *Taraxacum* microspecies RepeatExplorer results. However, they had a very low percentage of similarity hits.

Table 5.4. GC% of different types of cluster annotations.

Repetitive type	condition	Tri-nucleotide	Di-nucleotide	GC %		
				MIN	MAX	Average
Low_complexity	rarely followed by	ATA	AT	30.0	43.6	36.4
	Often followed by	AAA	AA			
Simple_repeat	rarely followed by	CTA	CG	32.2	50.9	39.4
	Often followed by	AAA	CA			
LTR-gypsy	rarely folowed by	CTA	CT	39.9	45.8	41.8
	Often folowed by	AAA	AA			
LTR-copia	rarely folowed by	CTA	CG	35.8	42.5	41.5
	Often folowed by	AAG	CA			
DNA-transposons	rarely folowed by	ATA	AT	29.9	42.9	35.6
	Often folowed by	AAA	AA			
rRNA	rarely folowed by	CTA	CT	49.5	52.6	51.1
	Often folowed by	AAA	AA			

The multi-loop or arc shapes represented LTR.Caulimovirus; thin linear shapes were observed for DNA.CMC.EnSpm; linear or sometimes shortly branched or arc shapes represented RC.Heliton; and DNA.hAT were represented by linear shapes, some with very short branches. The low-complexity clusters mostly had a smooth, linear pattern graph; the 45S cluster pattern was mostly like a thick line or coil; but 5S rDNA was represented mostly by star-shapes, or sometimes by doughnut shapes. Simple-repeat DNA clusters appeared as condensed stars or ball shaped graphs.

5.4.5. Characterization of repetitive DNA by *in situ* hybridization

Identification and classification of the repetitive fraction of *Taraxacum* genome were investigated by labelling DNA fragments from chosen RepeatExplorer clusters and their sequence contigs randomly. This study focused on the highly abundant DNA clusters with high genome proportions, usually chosen from the top 10 clusters, low-complexity, some simple-repeat, some clusters with unique graph pattern, and clusters with high similarity hits to dispersed repetitive DNA. In total 73 primer pairs were designed, amplified and labelled for fluorescent *in situ* hybridization (FISH). The primer sequences were used to calculate their abundance and their copy number per genome

in the three studied microspecies by mapping them back as a reference to the whole genome sequences (Figure 5.7, see also numerical data in Supplementary Table 5.2). The probe was used along with 5S and 45S (18S-5.8S-25S) rDNA arrays. Their localization on the *Taraxacum* chromosomes was observed to characterise the physical genome organization. Each primer or labelled probe was hybridized on each of three *Taraxacum* microspecies metaphase chromosomes. From ten to twenty images were analysed from each probe and each accession. There was a large number of *in situ* pictures to analyse. To simplify analysis, all of the *in situ* pictures were organized into different groups according to a cut-off value of 0.1. Also, some *in situ* pictures grouped according to their similarity hits to known repetitive DNA.

The results showed that *Taraxacum* chromosomes possess six 5S rDNA sites, three with strong signals and three with weaker signals. Thus, 5S rDNA sites usually occurs at a frequency of two per haploid set of chromosomes. The 45S rDNA has three sites, occurring at a frequency of one per haploid set of chromosomes. The 45S rDNA localized at the secondary constriction near the end of the short arm of a chromosome which shows a satellite connected by strands in the *in situ* signal. It is characterised by a large sub-distal intercalary filiform region. This region is elastic and may vary in length as seen in the *in situ* pictures. None of the 5S and 45S ribosomal DNA sites shared chromosomes with each other.

Table 5.5 shows the classification of the *in situ* signals for different types of repetitive DNA sequence from different clusters. Each FISH figure (Figure 5.8-2, 5.9-2, 5.10-2, 5.11-2, 5.12-2) represents the RepeatExplorer information, including graph shapes, genome proportions, similarity hits to repetitive DNA families, and protein domains for chosen cluster sequences as probes for *in situ* hybridization.

According to the cut-off value of 0.1, the major group which contained a large number of clusters is group 1 (Supplementary Figure 5.3) and the majority of clusters are annotated with low-complexity. Results showed that the low-complexity probes mostly located in the centromeric region, either in a braod zone (Figure 5.8-1A, C, E, G) or tightly concentrated around the centromere (Figure 5.8-1D, I, J).

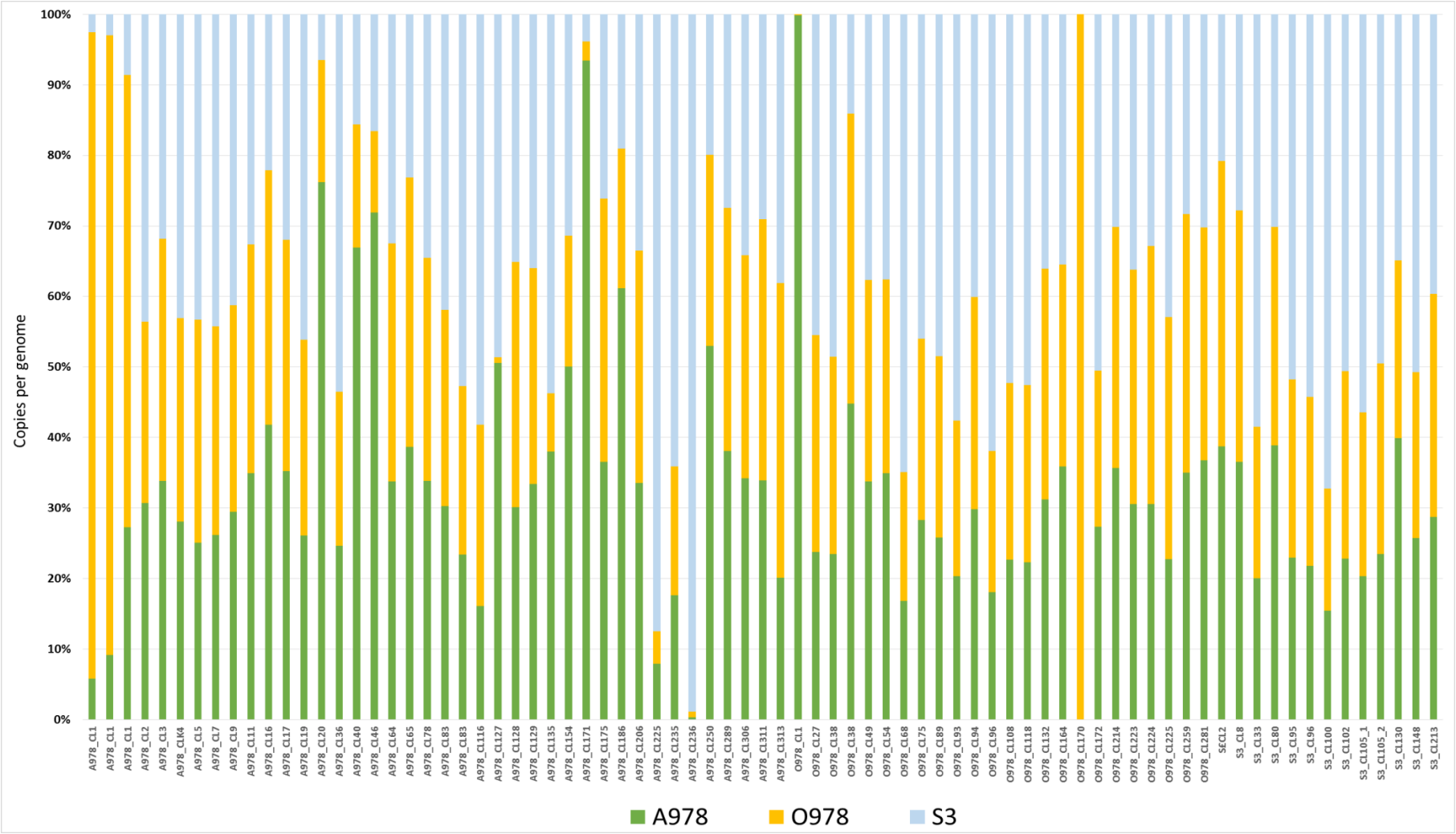


Figure 7. Copy number per genomes of primer region used in FISH of three *Taraxacum* genomes.

Table 5.5. Classification of FISH signals from different clusters outcomes from RepeatExplorer programs.

Repetitive localization on chromosomes	gypsy	copia	Simple_ repeat	Low_complexity	DNA. hAT.AC	DNA.CMC. EnSpm	Satellite	LTR. Caulimovirus	Helitron	DNA.TcMar. Stowaway	rRNA	No Similarity hits
Broad 'centromeric' region	CL1, CL135	CL46, CL64, CL95	CL1,	CL17, CL78, CL206, CL154, CL175, CL250, CL313, CL129, CL100, CL108, CL68, CL16, CL95		CL281, CL224						
centromeric except satellite chr. or some other chr.				CL18, CL235, CL93						CL259,		
Tightly concentrated to centromere.	CL9, CL83c4-a	CL36	CL38-1	CL128, CL186, CL105-1, CL8, CL27								
Centromere, gap on some chr.			CL164,	CL3,							CL96,	CL172,
sometimes whole arms and broader signals				Cl2, CL7, CL19,								
double dot signals on centromere, both intercalary and sub-telomeric				CL5								
dispersed signal on all chr.	CL1-F CL4	CL116, CL213	CL49, CL65,	CL2, CL102, CL105-2	CL130			CL94, CL225,				
Double dot on many chr. Not all chr.		CL80		CL40,								
dots over all chr. tended to centromere.	CL20			CL171		CL289, CL148						CL223,
Proximal distribution.				CL11, CL311, CL306, CL236, CL225K								
6 signals/5S rDNA	CL83c5-b						CL33, CL89,					
Associated with Satellite				CL132		CL96,			CL214,			

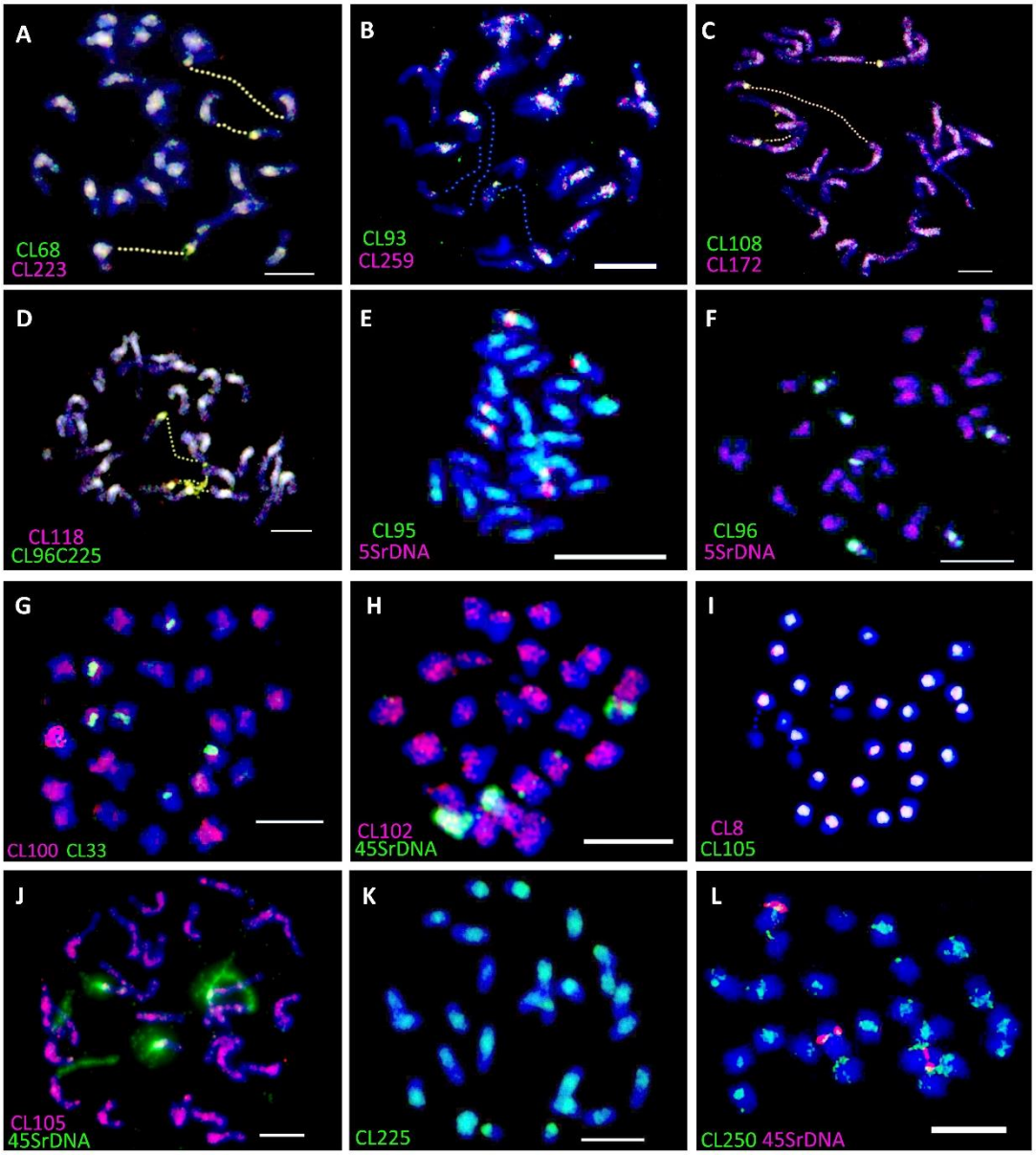


Figure 5.8-1. Genomic distribution of low-complexity clusters. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH patterns come from the clusters that written beside the picture and these clusters annotated with Low_complexity in RepeatExplorer outcomes. These clusters were grouped according to cut-off value 0.1.

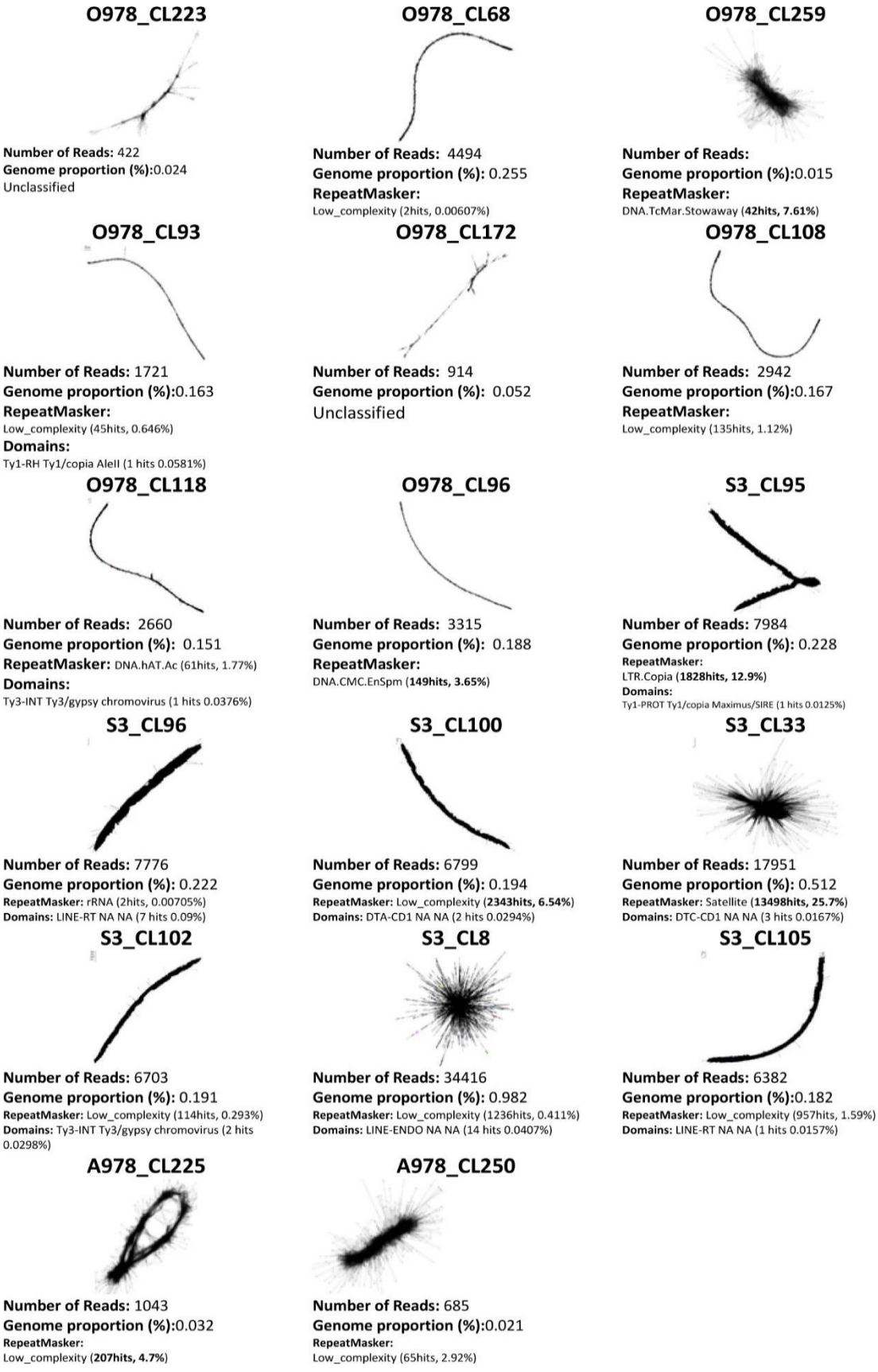


Figure 5.8-2: Pattern from RepeatExplorer outcomes that used to produce the probe for Figure 5.8-1 FISH.

There are some other low-complexity clusters that I studied but they are not grouped by the cut-off feature. They presented a very different distribution from the previous groups (Figure 5.8-1). The results shows that the low-complexity probes located mostly in the broad centromeric region (Figure 5.9-1C) or labelled tightly to the centromeric region (Figure 5.9-1D-red) and in another case the probe was distributed on all of the chromosomes and sometimes excluded from the broad centromeric region (Figure 5.9-1B, D-green, I, E-green arrows). In some cases double dot signals were obtained, intercalary and sub-telomeric (Figure 5.9-1A), and a broad proximal region on most of the chromosomes (Figure 5.9-1 F, G, H). The signals could include the gap in between (Figure 5.9-1 C-white arrows) or were absent on some chromosomes (Figure 5.9-1 E-white arrows), or the signal was absent on satellite chromosomes (Figure 5.9-1C, yellow arrows). In Figure 5.9-1J, the probe shows very low copy number in O978, and was not detected in either the A978 or S3 genomes.

The probe labelled from clusters of LTR-gypsy (Figure 5.10-1) show a variable signal location and distribution on metaphase chromosomes. In most cases, it gave a signal in a centromeric location but differed in the way it was distributed. Sometimes signals were distributed on all of the chromosomes (Figure 5.10-1 B-green, D-green, G) or had broad centromeric locations (Figure 5.10-1D-red, F, H, I), or sometimes had whole arms with broader, strong signals on some chromosomes but lighter on the others (Figure 5.10-1B-red), or were tightly located on the centromere of all chromosomes (Figure 5.10-1E). The signal is double dots (Figure 5.10-1F), located on centromeric region with including of gaps on some of the whole chromosomes (Figure 5.10-1A) absent on some chromosomes (Figure 5.10-1 B white arrows) located at centromere (Figure 5.10-1C). Figure 5.10-1 I shows the green signals from CL40. They are candidates for a tandem repeat at centromeres.

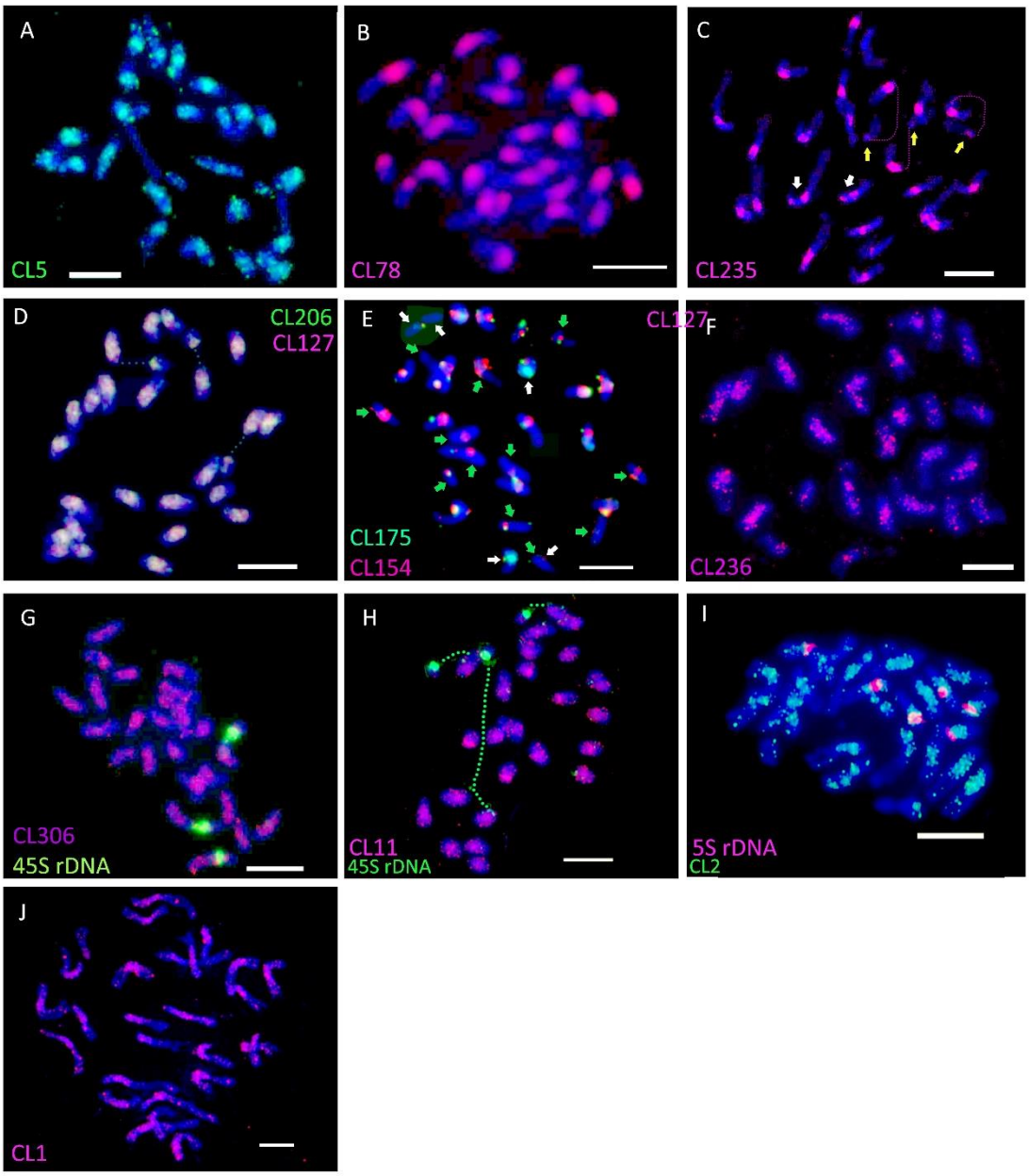
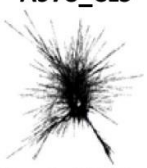


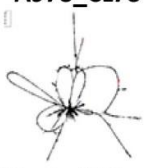
Figure 5.9-1. Genomic distribution other Low_complexity (A-I) and Simple_repeat, (J) clusters that not grouped by Cut-off feature. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. White arrows: gap position in the signal, green arrows: excluding signals on centromeric region, yellow arrows: absent of the signals on satellite chromosomes.

A978_CL5



Number of Reads: 27316
Genome proportion (%):0.834
RepeatMasker:
Low_complexity (447hits, 0.349%)
Domains:
Ty3-RT Ty3/gypsy chromovirus (2 hits 0.00732%)

A978_CL78

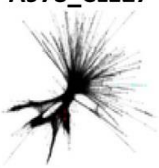


Number of Reads: 7619
Genome proportion (%): 0.233
RepeatMasker:
Low_complexity (347hits, 1.09%)
Domains:
DTM-CD1 NA NA (40 hits 0.525%)

A978_CL235

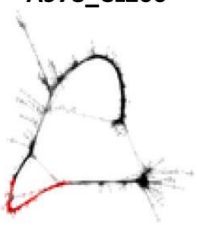
Number of Reads: 846
Genome proportion (%): 0.026
RepeatMasker:
Low_complexity (26hits, 1.47%)

A978_CL127



Number of Reads: 4754
Genome proportion (%):0.145
RepeatMasker:
Low_complexity (**836hits, 10.2%**)
Domains:
DTA-CD1 NA NA (17 hits 0.358%)

A978_CL206



Number of Reads: 1428
Genome proportion (%):0.044
RepeatMasker:
Low_complexity (**255hits, 5.79%**)
Domains:
DTH-CD1 NA NA (198 hits 13.9%)

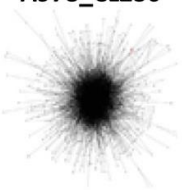
A978_CL154

Number of Reads: 3276
Genome proportion (%):0.1
RepeatMasker:
Low_complexity (**869hits, 7.28%**)

A978_CL175

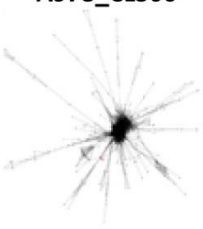
Number of Reads: 2157
Genome proportion (%): 0.066
RepeatMasker:
Low_complexity (3hits, 0.0272%)

A978_CL236



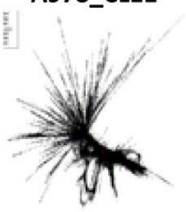
Number of Reads: 821
Genome proportion (%): 0.025
RepeatMasker:
Low_complexity (14hits, 0.387%)
Domains:
Ty1-INT Ty1/copia AleI/Retrofit (1 hits 0.122%)

A978_CL306



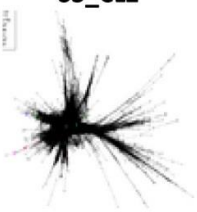
Number of Reads: 348
Genome proportion (%):0.011
RepeatMasker:
Low_complexity (6hits, 0.589%)
Domains:
Ty1-RH Ty1/copia Angela (1 hits 0.287%)

A978_CL11



Number of Reads: 21574
Genome proportion (%):
0.659
RepeatMasker:
Low_complexity (**2925hits, 3.55%**)
Domains:
Ty3-INT Ty3/gypsy chromovirus (5 hits 0.0232%)

S3_CL2

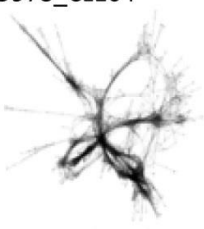


Number of Reads: 4282
Genome proportion (%):
1.12
RepeatMasker:
Low_complexity (513hits, 1.34%)
Domains:
LINE-ENDO NA NA (3 hits 0.0701%)

O978_CL1

Number of Reads: 24623
Genome proportion (%):1.4
RepeatMasker:
Simple_repeat (**24597hits, 50.8%**)

O978_CL164



Number of Reads: 1065
Genome proportion (%):
0.06
RepeatMasker:
Simple_repeat (4hits, 0.0993%)

Figure 5.9-2: Pattern from RepeatExplorer outcomes that used to produce the probe for Figure 5.9-1 FISH.

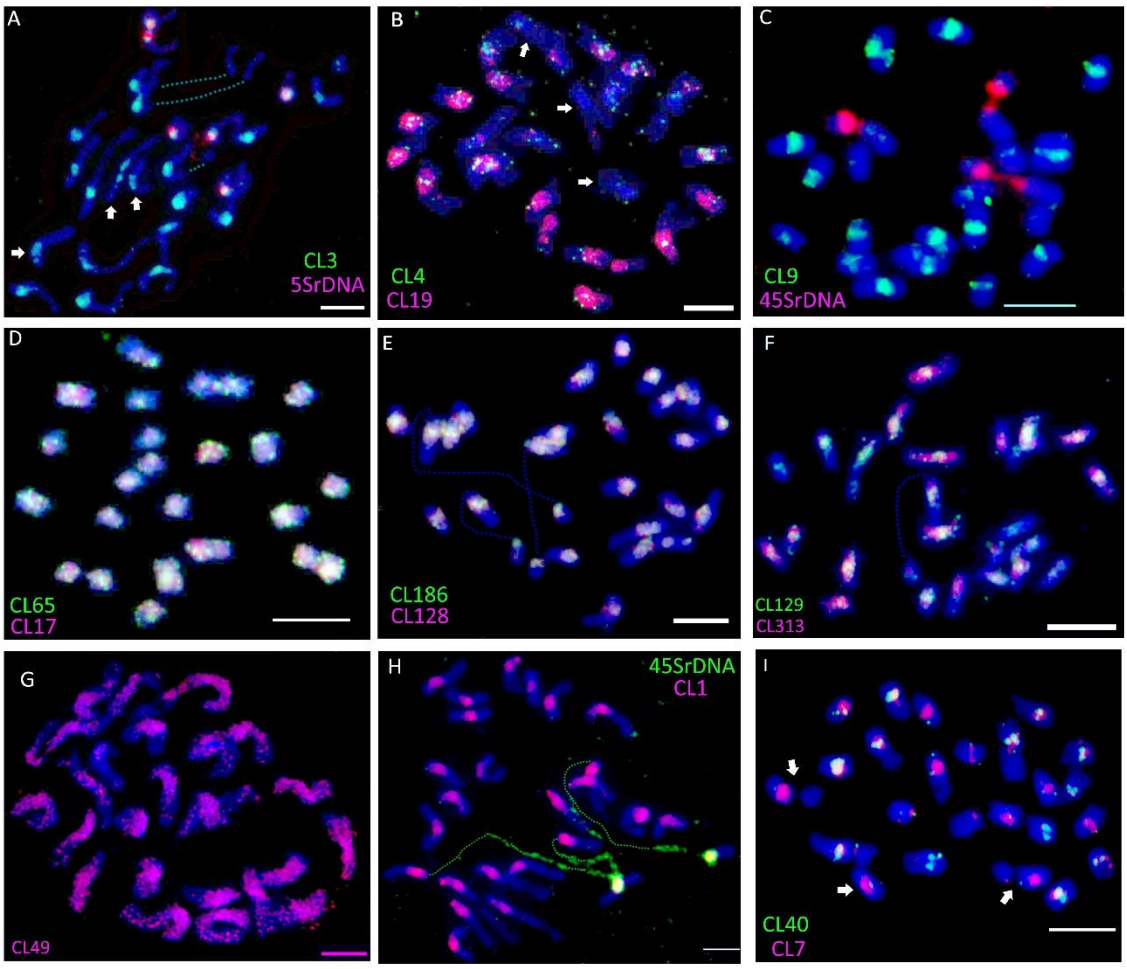


Figure 5.10-1. Genomic distribution of the LTR-gypsy retrotransposons repeats. Most of the clusters annotated with Low_complexity but they were grouped with the group of clusters that belong to LTR-gypsy retrotransposons. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The related clusters that produced the probe from written beside each picture with colour that they labelled with. White arrows: excluding signals from centromere (A), weak signals or absent of the signals on some chromosomes (B), absent of the green signals on the satellite chromosomes (I).

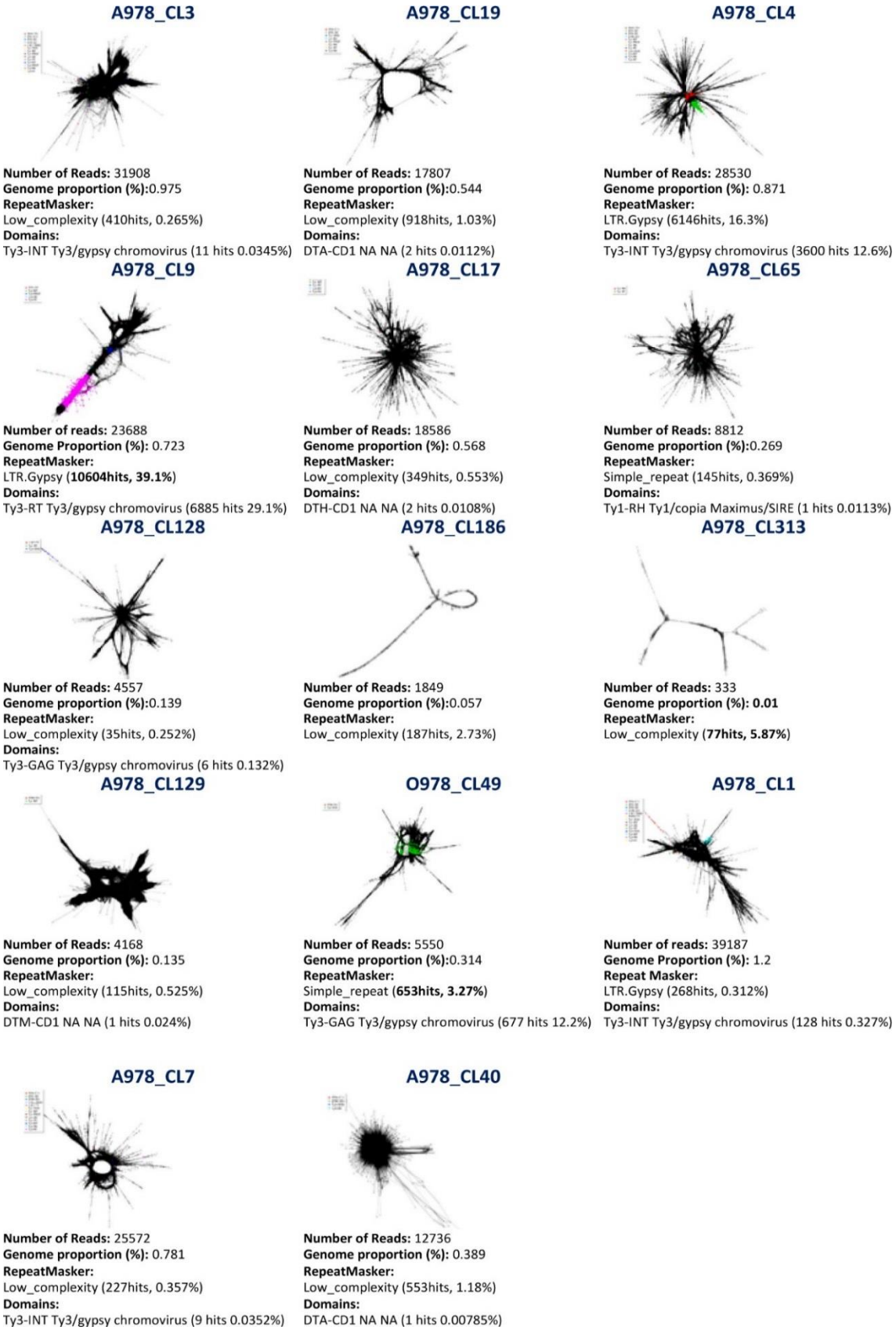


Figure 5.10-2: Pattern from RepeatExplorer outcomes that used to produce the probe for Figure 5.10-1 FISH.

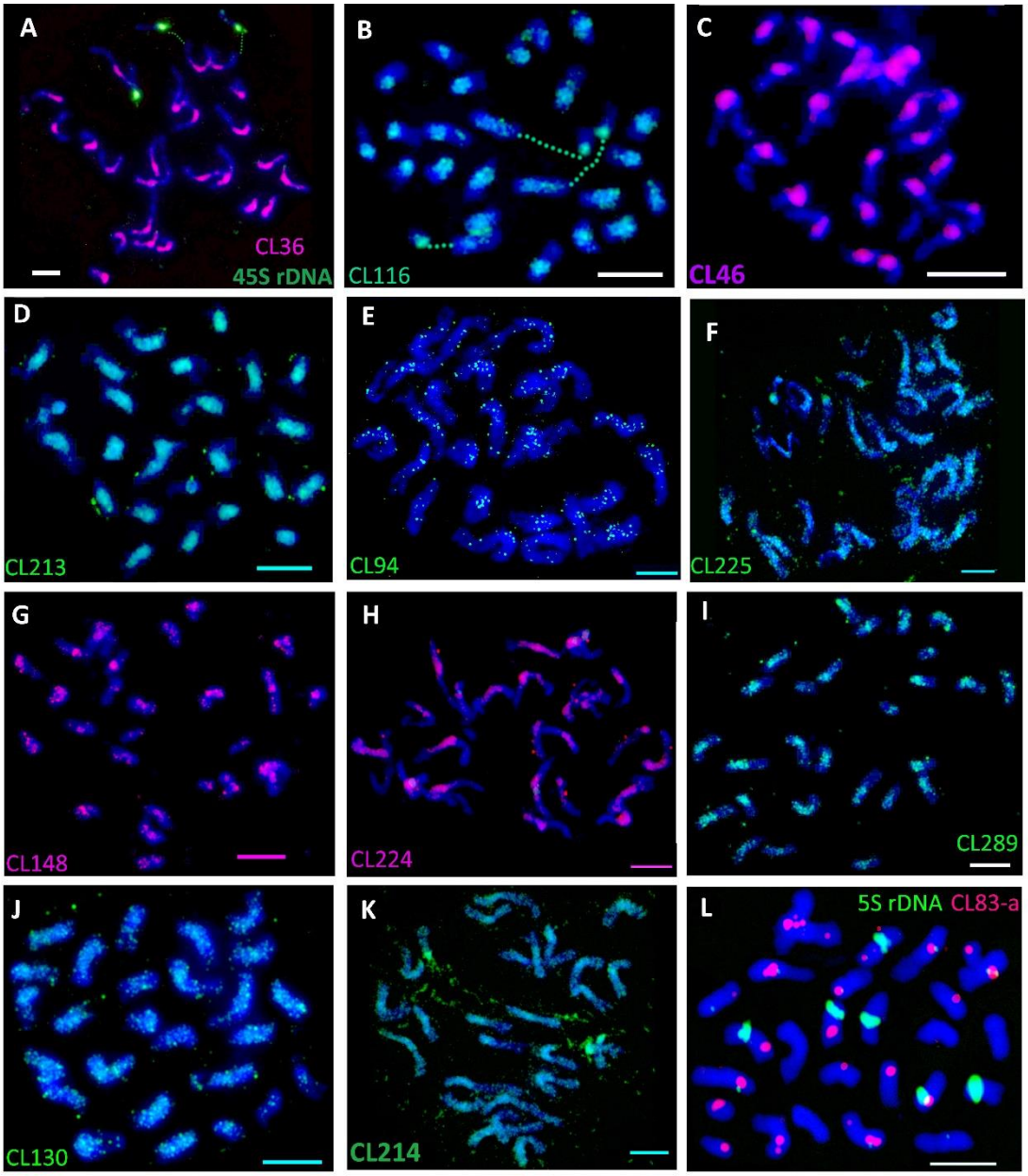
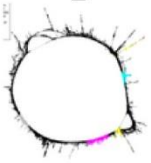


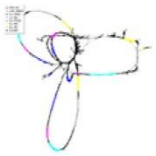
Figure 5.11-1. Genomic distribution of real retroelements. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. (A) Metaphase from O976; (B) Metaphase from A976; (C) Metaphase from S983; (D) Metaphase from O976; (E) Metaphase from S3; (F) Metaphase from S3; (G) Metaphase from S3, CL148; (H) CL224, has 25% of similarity hits to DNA-CMC-EnSpm. (I) Metaphase from A976; (J) Metaphase from S3; (K) Metaphase from S3; (L) Metaphase from A976, CL83a.

A978_CL36



Number of Reads: 13568
Genome proportion (%):0.414
RepeatMasker:
LTR.Copia (**2534hits, 14.3%**)
Domains:
Ty1-RT Ty1/copia Maximus/SIRE (966 hits 7.12%)

A978_CL116



Number of Reads: 5173
Genome proportion (%): 0.158
RepeatMasker:
LTR.Copia (**2197hits, 33.2%**)
Domains:
Ty1-RT Ty1/copia AleII (425 hits 8.22%)

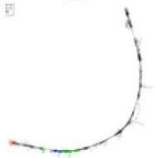
A978_CL46

Number of Reads: 12198
Genome proportion (%): 0.373
RepeatMasker:
LTR.Copia (11032hits, 84.1%)
Domains:
Ty1-RT Ty1/copia Bianca (6119 hits 50.2%)

S3_CL213

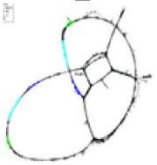
Number of Reads: 1376
Genome proportion (%):0.039
RepeatMasker:
LTR.Copia (**936hits, 55.2%**)
Domains:
Ty1-RT Ty1/copia Tork (230 hits 16.7%)

O978_CL255



Number of Reads: 412
Genome proportion (%):0.023
RepeatMasker:
LTR.Caulimovirus (**140hits, 29.2%**)
Domains: PARA-RT NA NA (57 hits 13.8%)

O978_CL94



Number of Reads: 3410
Genome proportion (%): 0.193
RepeatMasker:
LTR.Caulimovirus (112hits, 1.66%)
Domains:
PARA-RT NA NA (258 hits 6.66%)

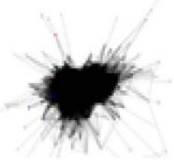
S3_CL148

Number of Reads: 2902
Genome proportion (%):0.083
RepeatMasker:
DNA.CMC.EnSpm (23hits, 0.227%)
Domains:
DHH-CD1 NA NA (13 hits 0.448%)

O978_CL224

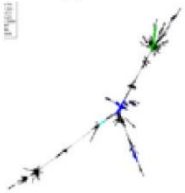
Number of Reads: 414
Genome proportion (%):0.024
RepeatMasker:
DNA.CMC.EnSpm (**139hits, 24.8%**)
Domains:
DTC-CD1 NA NA (133 hits 32.1%)

A978_CL289



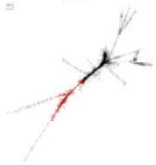
Number of Reads: 390
Genome proportion (%):0.012
RepeatMasker:
DNA.CMC.EnSpm (**100hits, 5.49%**)
Domains: Ty3-INT Ty3/gypsy chromovirus (2 hits 0.513%)

S3_CL130



Number of Reads: 4024
Genome proportion (%):0.115
RepeatMasker:
DNA.hAT.Ac (**532hits, 5.89%**)
Domains:
DHH-CD2 NA NA (394 hits 9.79%)

O978_CL214



Number of Reads: 485
Genome proportion (%):0.028
RepeatMasker:
RC.Helitron (**68hits, 11.2%**)
Domains:
DHH-CD2 NA NA (98 hits 20.2%)

A978_CL83a

Number of Reads: 7252
Genome proportion (%): 0.221
RepeatMasker:
LTR.Gypsy (**4301hits, 51.9%**)
Domains:
Ty3-RT Ty3/gypsy Athila (1289 hits 17.8%)

Figure 5.11-2: Pattern from RepeatExplorer outcomes that used to produce the probe for Figure 5.11-2 FISH.

The patterns with a high proportion of hits to clear retro domains, such as LTR-gypsy, LTR-copia, LTR-caulimovirus, DNA-CMC-EnSpm and Rc-Helitron were chosen for further analysis by *in situ* hybridisation. They were not included in cut-of group lists. However, some of them gave very high similarity hits to retrotransposon types. From these, a young but abundant LTR retroelement is shown in CL36. The sequences in this cluster represent 0.41% of the genome. This is a Ty1-Copia element because the retroelement coding domains are in the order RNaseH - Revese Transcriptase - Integrase – Protease (Figure 5.11-2). *In* situ hybridization showed that this probe is located tightly around the centromeres of all eight chromosome triplets (Figure 5.11-1 A).

CCL94 and CL225 have similarity hits similar to LTR-caulimovirus: both (Figure 5.11-1 E, F) show very dotty diffused signals over all of the chromosomes.

The probe from the DNA-CMC-EnSpm cluster (Figure 5.11-2) showed a rather dispersed and spotty distribution over chromosomes but tending towards the centromeres (Figure 5.11-1 G, I), or in Figure 5.11-1 H showed a broad centromeric signal, perhaps with some gaps around the centromeres. CL130 has 5.9% similarity hits to DNA-hAT-Ac (Figure 5.11-2) which showed *in situ* signals of dots spread over all chromosomes (Figure 5.11-1 J). CL214 (RC.Helitron of 11.2% similarity hits) shown quite dispersed signals over all chromosomes and more broad pericentromeric and the signals seem to be associated with the centromeres (Figure 5.11-1 K). LTR-gypsy-Athila, with 51.9% similarity hits in CL83a, showed *in situ* hybridization signals located tightly on the centromere and it is clearly shown that the signals are absent from six chromosomes (Figure 5.11-1 L).

5.4.6. Tandem repetitive DNA and chromosome specific probes

CL80 showed a tandem repeated pattern after dot plot (Figure 5.12F), which consist of the fragment of 49bp with the copy number of 32.4 and 92% matches. This cluster showed a unique pattern (Figure 5.12G) among the other clusters resulting from RepeatExplorer analysis of the S3 genome. It shows very low similarity hits (0.1 %) to LTR-copia. Further analysis of this cluster by assembling it to the three Taraxacum whole genomes shows it has a very high coverage in the three genomes

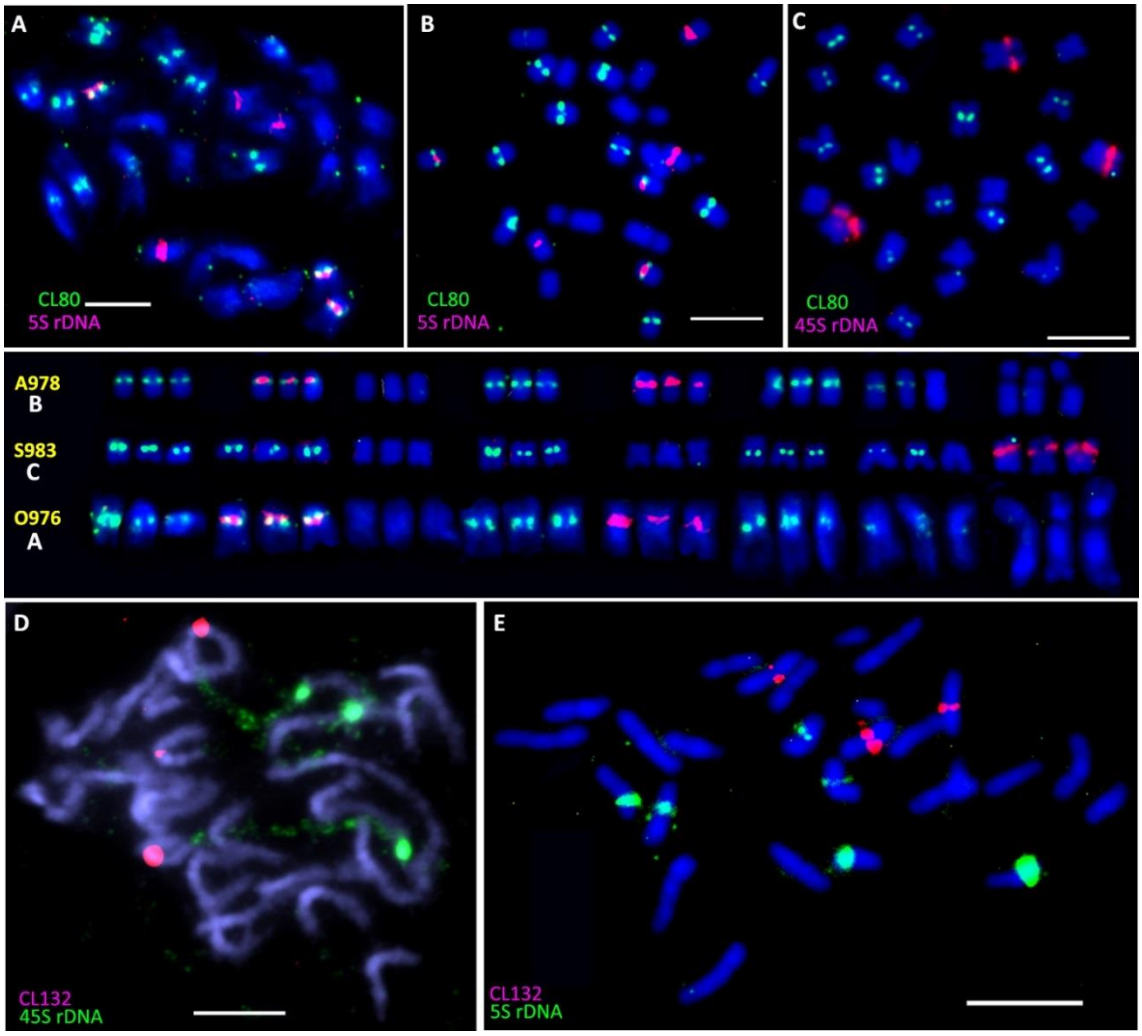
(Supplementary Table 5.2 and Figure 5.7). Interestingly, fluorescence *in situ* hybridization on mitotic chromosomes of the three studied genomes showed that this pattern resulted in a different and unique pattern. It comprised double dot signals in centromeric regions on just 14 chromosomes not all chromosomes or multiple of three (Figure 5.12A-C). Using 5S and 45S rDNA probes along with this (CL80) *in situ* signal, triplet chromosomes were homologised and karyotypes prepared for the three *Taraxacum* microspecies (Figure 5.12 A-C).

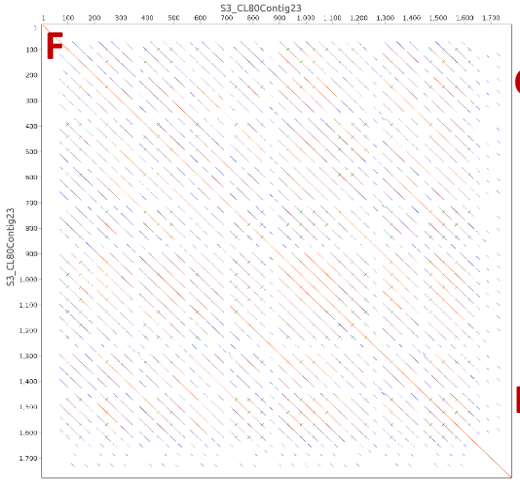
The CL132 from O978 genome showed chromosome-specific signals. The probe from this cluster revealed three signals on the same 45S rDNA chromosome on primary constriction of the chromosomes, which it is rDNA associated primer. It showed very strong signals on two chromosomes and weaker signals on the third chromosome (Figure 5.12D, E, H).

5.4.7. Estimation of repeat proportions, and differences between the three *Taraxacum* microspecies (S, A and O)

Each probe, designed from RepeatExplorer, was tested on the three different *Taraxacum* microspecies to figure out the differences between them in terms of signal shapes and repetitive DNA distributions, along with the primer binds coverages and copies per genome (Table 5.5).

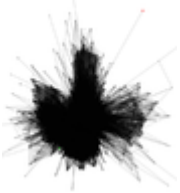
FISH signals showed clear differences in strength between two probes on the same metaphase, according to our data table of primer bind coverage for the studied microspecies (Figure 7). Sometimes probe signals go with the calculated coverages and sometimes not. For example, (Figure 7, Figure 5.13-1 B). Also, the signal from CL135 has a higher copy in S3 than in A978 or O978 (Figure 7). *In situ* signals agree with these copy numbers, being weak on all chromosomes of O978 and A978, and with a stronger signal on broad centromeric areas of the S3 genome. The signal is not present on all chromosomes (probably very weak on c. 3 and absent from satellites) in S3 and O978 (Figure 5.13-1C). The CL38-2 copy number in S3 is much lower than in O978 or A978 (Figure 7). *In situ* signal shows that in A978 and O978 genome the signals are very strong, tight to the centromere but in the S3 genome the signal is very dispersed and dotted over the chromosomes (Figure 5.13-1I).





S3_CL80

G



Number of Reads: 947

Genome proportion (%):0.2480

RepeatMasker:

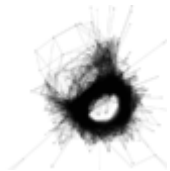
LTR.copia (2hits, 0.103%)

Domains:

Ty1-RH Ty1/copia Bianca (1 hits 0.106%)

O978_CL132

H



Number of Reads: 632

Genome proportion (%): 0.06

RepeatMasker:

Low_complexity (2hits, 0.0909%)

Figure 5.12. Genomic distribution of some clusters from RepeatExplorer outcomes. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). (A) Metaphase O976, (B) A978, (C) S983 show the distribution of signals from CL80 (green signal) along with 5S rDNA in metaphase A978 and O976 and with 45S rDNA (red signal) in S983. (D, E) show the distribution of CL132 signals along with 45S rDNA and 5S rDNA on *Taraxacum* chromosomes. (F) The dot plot pattern is for CL80 tandem repeated pattern, with the RepeatExplorer graph shape for CL80 (G) and CL132 (H).

CL2 probes dose not label all the chromosomes and are absent from some. In O978 the signal is absent or very weak on six chromosomes and in A978 the signal is absent on three chromosomes (white arrow), exactly matching the copy number analysis. On some chromosome arms, there are broad signals (Figure 5.13-1A). CL16 probe shown that O978 genome chromosomes have weaker to no signal compared with A978 and S983 (white arrows-Figure 5.13-1 K).

In other cases, the distribution of the probe signals reflected copy number in the genome. CL171 has very high coverage in A978 compared with S3 and O978 (Figure 7), but this is not reflected by *in situ* hybridization. Signals are dispersed around the centromere in each of the O978 and S3 genomes but show two dots at the centromere in the A978 genome (Figure 5.13-1 D). *In situ* hybridization for CL281 shows the signals tending to the broad centromeric location in S3 and O978 but at a subtelomeric location in the A978 genome (Figure 5.13-1 E). CL27 produces a signal in a broadly centromeric region but with a different distribution on the chromosomes. Signals in O978 are broadly distributed and concentrated at the centromere; in A978 they are also broadly distributed but with gaps and excluding the centromere. In S3 the signals are less distributed on the chromosomes but instead are concentrated at centromere locations (Figure 5.13-1 F).

In addition, I was interested in visualizing the distribution of some clusters that have high genomic proportions, in order to reveal any differences between the three microspecies in location and frequency of these clusters on the chromosomes. CL1 has a high genome proportion 1.2%, with a broad centromeric location in the A978 genome, a more dispersed distribution over the chromosomes in O978, and in S3 more concentrated in the centromere.

CL1 produces broad centromeric signals, but they are much dispersed and dotted over the chromosomes in the A978 and S3 genomes, and gives strong signals in the centromere and in O978 genome distributed over the whole chromosomes absent on some arms (Figure 5.13-1 H). CL38-1 but in the different data set of RepeatExplorer output file, the signal is distributed over the whole chromosomes in S3 but, in A978 and O978, is concentrated in the centromere (Figure 5.13-1 J). According to the above *in situ* hybridization results, there are differences between the genomes of the three related microspecies, in terms of abundance, pattern of the signals, coverage, and distribution of the many repeats.

Figure 5.13-1. Genomic distribution of the repetitive DNA from RepeatExplorer outcomes represents the differences between *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various cluster from RepeatExplorer results. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. White arrow: signal absent.

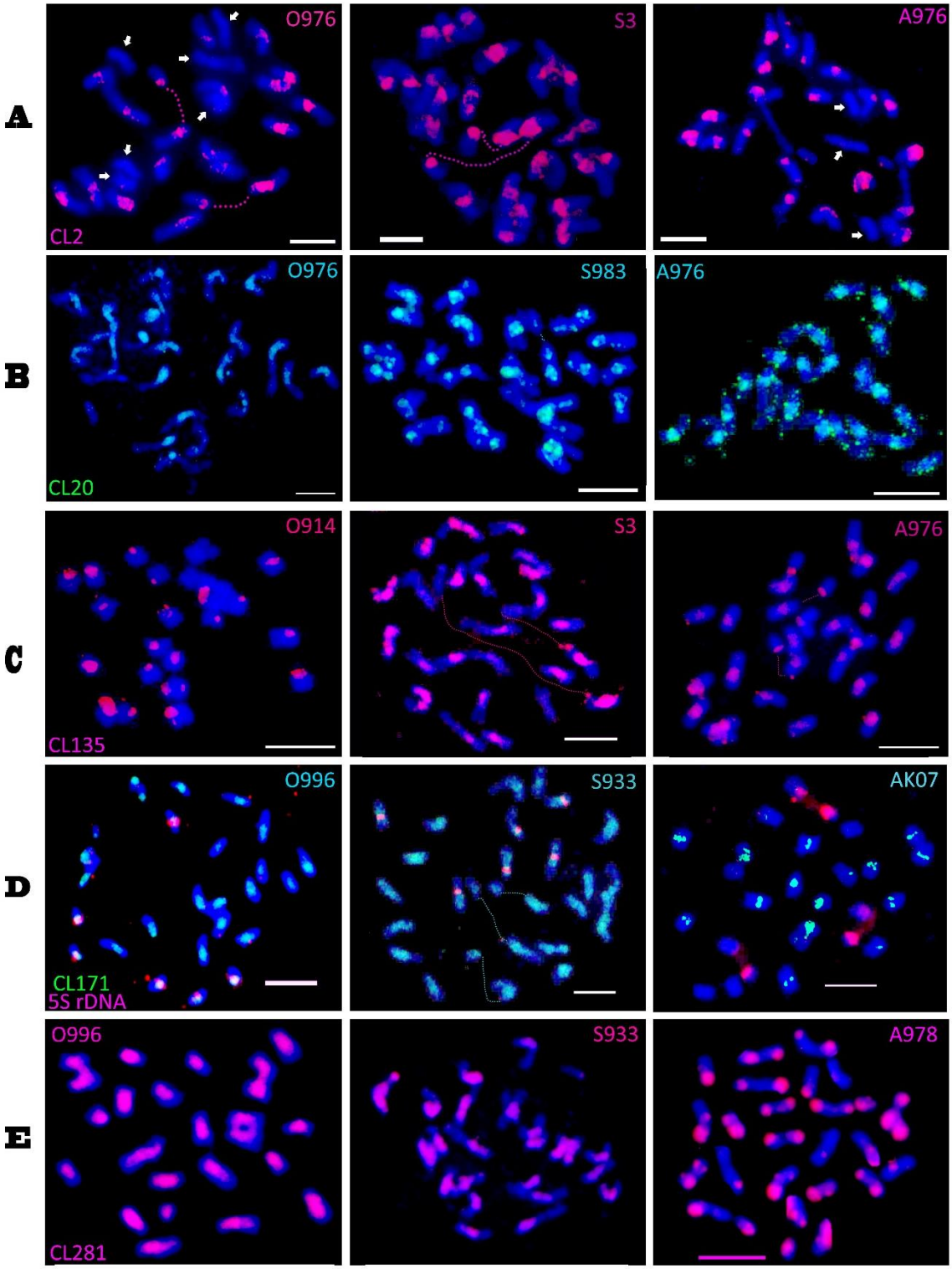
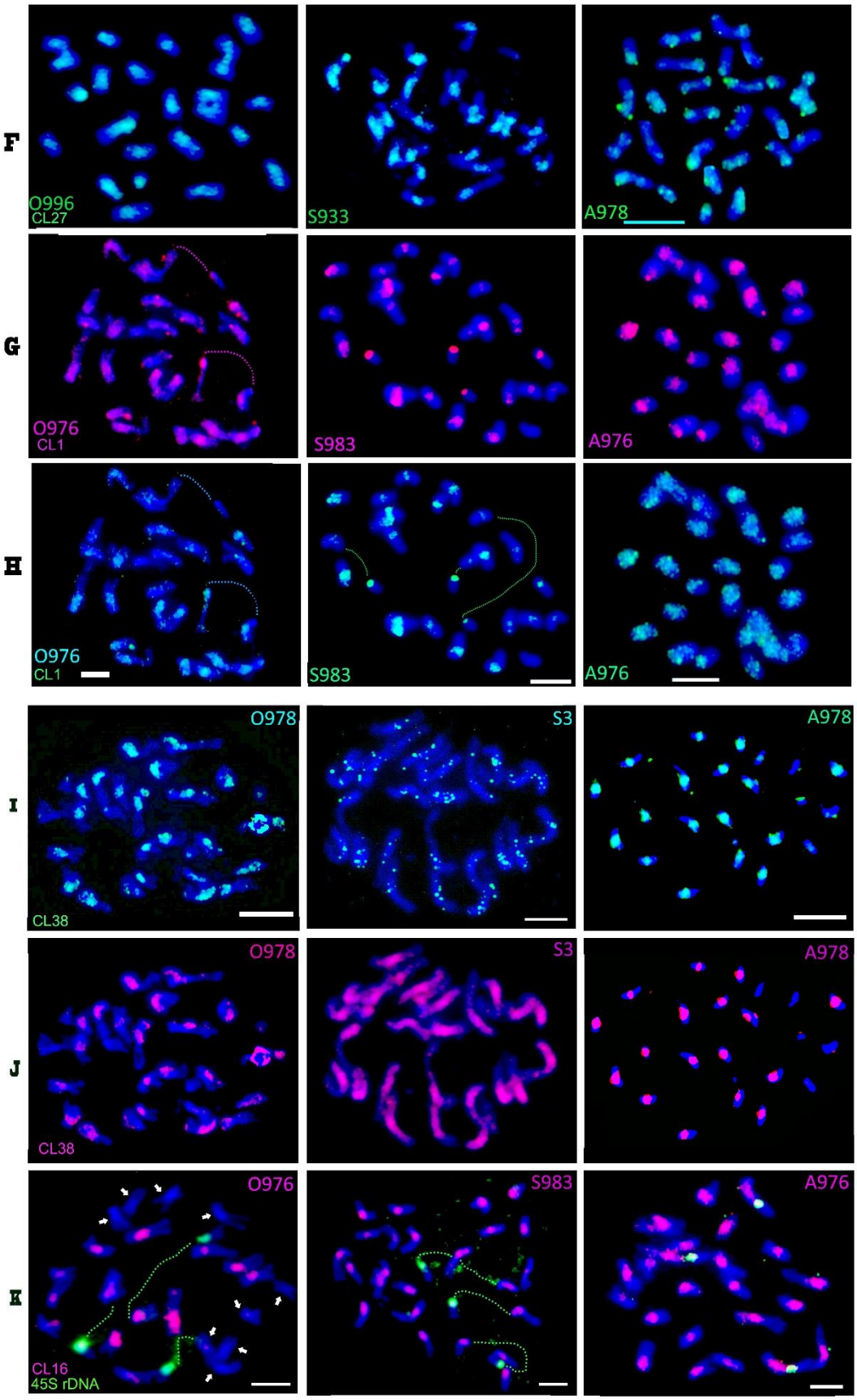


Figure 5.13-1 continue…



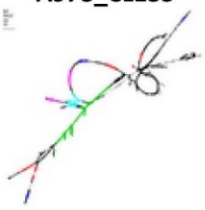
A978_CL2

Number of Reads: 33689
Genome proportion (%): 1.03
RepeatMasker:
Low_complexity (1605hits, 1.03%)
Domains:
Ty3-INT Ty3/gypsy Ogre/Tat (4 hits 0.0119%)

A978_CL20

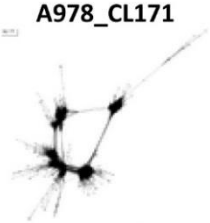
Number of Reads: 17286
Genome proportion (%): 0.528
RepeatMasker:
LTR.Gypsy (**13286hits, 66.2%**)
Domains:
Ty3-INT Ty3/gypsy chromovirus (10507 hits 60.8%)

A978_CL135



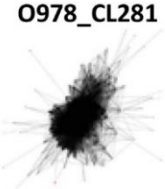
Number of Reads: 4201
Genome proportion (%): 0.128
RepeatMasker:
LTR.Gypsy (**1084hits, 21.1%**)
Domains:
Ty3-INT Ty3/gypsy chromovirus (468 hits 11.1%)

A978_CL171



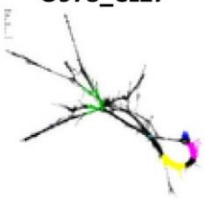
Number of Reads: 2303
Genome proportion (%): 0.075
RepeatMasker:
47.1 Low_complexity (**497hits, 5.64%**)
Domains:
DTA-CD1 NA NA (1 hits 0.0434%)

O978_CL281



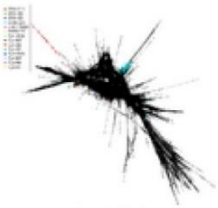
Number of Reads: 194
Genome proportion (%): 0.011
RepeatMasker:
DNA.CMC.EnSpm (**42hits, 4.62%**)
Domains:
Ty1-RT Ty1/copia Angela (1 hits 0.515%)

O978_CL27



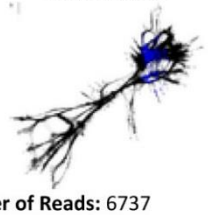
Number of Reads: 7785
Genome proportion (%):0.441
RepeatMasker:
LTR.Gypsy (**2625hits, 28.3%**)
Domains: Ty3-RT Ty3/gypsy chromovirus (1063 hits 13.7%)

A978_CL1-F



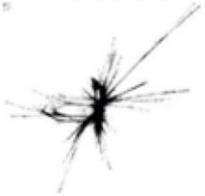
Number of reads: 39187
Genome Proportion (%): 1.2
Repeat Masker: LTR.Gypsy (268hits, 0.312%)
Domains: Ty3-INT Ty3/gypsy chromovirus (128 hits 0.327%)

O978_CL38



Number of Reads: 6737
Genome proportion (%):0.372
RepeatMasker:
Simple_repeat (504hits, 2.12%)
Domains:
Ty3-GAG Ty3/gypsy chromovirus (612 hits 9.08%)

O978_CL38



Number of Reads: 6566
Genome proportion (%):0.372
RepeatMasker:
Simple_repeat (101hits, 0.361%)
Domains:
Ty3-INT Ty3/gypsy chromovirus (5 hits 0.0761%)

A978_CL16

Number of Reads:18854
Genome proportion (%):0.576
RepeatMasker:
Low_complexity (116hits, 0.122%)
Domains:
LINE-RT NA NA (1 hits 0.0053%)

Figure 5.13-2: Pattern from RepeatExplorer outcomes that used to produce the probe for Figure 5.13- FISH.

5.5. Discussion

Recently the study of genomic repetitive DNA has become more common, commensurate with the development of next generation sequencing (NGS) technology. The latter has played an important role in the study of repetitive DNA sequences, particularly transposable elements and especially plant genomic DNA (Tenaillon *et al.* 2011; Macas *et al.* 2007; Wicker *et al.* 2008; Hribova *et al.* 2010; Staton *et al.* 2012).

There are two different groups of transposable (mobile) elements apparently present in all eukaryotic genomes (Heslop-Harrison and Schmidt 1998). They differ in their transposition mechanism: class 1 uses RNA as an intermediate in the transposing process, whereas class2 does not.

Nowadays, low cost of NGS has led to increasing amounts of genomic sequence data, and the development of many computational programs to characterize and assemble the sequences. In this chapter, with the whole genomic DNA sequences from three *Taraxacum* microspecies, I was interested in studying the repetitive DNA component of three closely related microspecies of *Taraxacum* by using the graph-based clustering method of the RepeatExplorer program.

RepeatExplorer cluster based methods of Novak *et al.* (2010) were used to characterize and analyse repetitive DNA components of the *Taraxacum* genome, and estimate the proportions of all repetitive DNA families. These analyses were assessed by using 45Gb of Illumina unassembled reads.

RepeatExplorer takes a large sample of the complete sequence reads and performs a sequence similarity search among the reads, then analyses partial overlap sequences among reads and contigs, which are subsequently put into clusters. The cluster size is a representation of the repeat portion. The sequences of each clusters are then compared against the repeatmasker, a database of interspersed repeats.

In RepeatExplorer outcome files, each repetitive DNA family is represented by a different cluster which can be characterized by a de Bruijn graph composed of connected dots. The protein domains are colour coded and represented in the dataset

graphs. Each of the total number of base pairs, number of reads, and genome proportion are calculated in the analysed dataset, also number and percentage of the hits to known repeats from repeatmasker database are represented. RepeatExplorer outcomes represent only the clusters that are greater than or equal to 0.01% of the genome (Novak *et al.* 2010), so, some repetitive sequence reads with very low coverage cannot be clustered, e.g. in the current study, telomeric repeats, tandem repetitive DNA, and satellite repeats remained unclustered because repeat classifications are determined by the reads number (coverage) and cluster size.

My results show that there are small differences between the genomes of the three Taraxacum microspecies. These differences could be due to the read length of the three genomes (A978 and O978 are 150 bp read length, S3 has 300 bp length). According to Zhang *et al.* (2011), sequences with longer reads result in a better assembly of the sequences. Because of this, I conducted some further subset analysis by taking the first 150bp and the middle 150bp of the 300bp of the S3 genome and comparing the sequences with similar analysis of S3 and A978 and O978 genomes (Figure 5.1 and Supplementary Figure 5.1). However, the RepeatExplorer graphs of cumulative genome proportion represented by the three 150bp segments of S3 are still different from the 150bp reads of O978 and A978 after cluster 100 (35% of the genome-Figure 5.2). It is not clear if this is a real difference because the genomes are 'different' or if it is an analytical difference built into the RepeatExplorer program. Moreover, 5,000,000 reads of 150 bp sequences from S3 genome were shuffled on Ubuntu Linux 13.10, and uploaded to RepeatExplorer, (Supplementary Figure 5.2).

Ideally, every cluster would contain most of the reads from a particular class or type of repetitive element. While this is true for some repeats and clusters, I found that some types of element are separated into multiple clusters. The 45S rDNA is an example, in which their copy number, organised tandemly in the genome, is responsible for the tight circular layout (Figure 5.5). According to Novák *et al.* (2010) this is either caused by a "missing link" where the chain of overlapping reads is interrupted, or by a "weak link" when the number of overlapping reads is low and the element is split into two or more sub-clusters. Low coverage (missing links between the sequences reads) can be caused by a number of factors, including low read depth

which increases the probability of gaps in the coverage, high variation of sequences in that region and subsequently caused the absence of similarity hits, and the total genome coverage of the genomic sequences. However, clustering of abundant repeats in the genome remain unaffected by lower read coverage (Novák *et al.* 2010).

5.5.1. Genome composition and abundant DNA sequence of "low complexity" and "simple repeat" in the three *Taraxacum* microspecies

The most difficult thing to understand when analysing RepeatExplorer outcomes is that each outcome comprises a number of different repetitive clusters with different genome proportions and similarity hits to different repetitive DNA families, and each cluster is made up of a varied composition of sequence types in distinct contigs. Moreover, RepeatExplorer does not identify the number of clusters and they remain as unknown/unclassified repetitive DNA clusters, many with low-complexity or simple-repeats, which make the results complicated to interpret. These clusters also represent a large proportion of genomic repetitive DNA. However, in many other studies these kind of sequences (low-complexity and simple-repeats) are removed before implementing the raw reads with RepeatExplorer (González *et al.* 2012; Staton *et al.* 2012), or are simply regarded as unclassified or unknown sequences (Novak *et al.* 2014). The results from the three *Taraxacum* genomes show that the clusters annotated with low-complexity or as simple-repeats comprise more than 25% of the genomic sequences and 40% of the total repetitive DNA sequence (Figure 5.4A, Table 5.3).

I designed primers from about 35 low-complexity clusters then labelled them. FISH results from these clusters gave very distinct results from each probe. Results from FISH were reflected in the different distribution and chromosomal locations of these probes, and I infer that these differences might be composed of numerous sequences from remnants of other repetitive DNA families in the genome. Low-complexity sequences have been recognised as abundant in eukaryotic proteins with unknown functions. However, Toll-Riera *et al.* (2012) suggested that low-complexity

perhaps contribute in the emergence of novel protein sequences, and Sveinsson *et al.* (2013) referred low-complexity clusters to plastid sequences.

The S3 genome gave a lower percentage of the totals of low-complexity and simple repeats compared with A978 and O978 genomes (Figure 5.4). The results show that the low-complexity proportions increased slightly after analysing 150 bp reads (first half and middle section) of the 300bp S3 and provide further support to the idea that the low-complexity sequences are remnants. The results show that the ratio of unclassified clusters increased dramatically because of the decrease in the amount of reads which make the cluster splits due to low coverage of reads (Figure 5.4A, B).

Not surprisingly, the long reads of 300bp make more sequences into repeat clusters, and single-copy sequences next to repeats get clustered in the repeats. Also a higher proportion of reads is classified into known sequences, as there is more chance to have an identifiable domain within 300bp than within 150bp (Figure 5.1).

A number of authors have noted that AT-content can increase during evolutionary time, and Wang *et al*. (2015) have suggested that "the retention of low complexity A/T-rich genomic 'graveyards' may contribute to the reduced GC-content observed in large plant genomes" (Šmarda *et al.* 2014, Wang *et al.* 20015). In the current study, the GC-content is not reduced (37% of all sequences, and 42% of repeat clustered sequences). The sequences which are put into clusters are not therefore products of a random process of degradation (Table 5.4).

I further analysed the clusters that annotated as simple-repeats, after analysing some of the cluster contigs reads with the tandem repeat finder. It showed that most of these clusters represent microsatellite DNA of di, tri, tetra, penta and hexa nucleotide repeats; and that microsatellite AT and GA di-nucleotide and ACC trinucleotides (Figure 5.6) are the most abundant in the genomes. However, there were two clusters in the A978 and O978 genomes with high proportions of monomeric poly-G and poly-A repeats and they were commonest in RepeatExplorer outcomes with high similarity hits to simple-repeats. Some other scientists also referred to simple repetitive sequences as microsatellites or they used a synonym such as "simple tandem repeats" (Gortner *et al.* 1996; Hearne *et al.* 1992; Poulsen *et al.* 1993; Braaten

et al. 1988; Hamada *et al.* 1982; Schafer *et al.* 1986; Tautz and Renz 1984; Vergnaud 1989).

5.5.2. Abundant repetitive DNA sequences and composition of *Taraxacum* microspecies genomes

In accordance with results from other plant species that have been studied so far, the present analysis showed that the *Taraxacum* genome is composed of LTR-retrotransposons (class I elements) (24-31%), and these are the most abundant repetitive DNA sequences found in the three microspecies of *Taraxacum* (Table 5.3). Of them, Ty1-copia represented 13-16% of the genome while the Ty3-gypsy elements represented 11-15% of the genome (Figure 5.4A). However, data from other sequencing projects indicate the prevalence of Ty3/gypsy retrotransposons in plant nuclear genomes (Macas *et al.* 2007; Bartoš *et al.* 2008; I.R.G.S. 2005; Velasco *et al.* 2007), which makes the results of the current study very interesting.

Novak *et al.* (2010) suggested that highly abundant repetitive DNA such as LTR-copia and gypsy do not affect by clustering methods, in contrast to non-LTR epetitive DNA which is less abundant than LTR repetitive DNA. DNA transposons with other class II elements comprise the minor proportion in the *Taraxacum* genome in current study (Figure 5.1). The large proportions of retrotransposons compared with DNA transposons is perhaps simply indicative of the fact that DNA transposons are more narrowly defined at the superfamily level, or it is perhaps indicative of their transposition mechanism (Heslop-Harrison and Schmidt 1998). LTR-copia frequency is even higher in banana and grapevine (Aubourg et *al.* 2007; Hřibová *et al.* 2010; Cossu *et al.* 2012; Meyers *et al.* 2001), however in other genomes such as papaya, Sorghum and rice, the frequency of LTR-gypsy is much higher than LTR-copia (Ming *et al.* 2008; Paterson *et al.* 2009; Spannagl *et al.* 2009). These differences may be due to the higher activity of LTR-copia retrotransposons than LTR-gypsy during *Taraxacum* evolution. In *Taraxacum* RepeatExplorer HTML output files none of the clusters blasted with real satellite repeat sequences. While tandemly repeated satellite DNA is reported in numerous plant species, some of it has been found to have very few such sequences.

5.5.2.1 *Taraxacum* repetitive DNA markers

Taraxacum nuclear diversity has been measured by SSR and AFLP methods (Van der Hulst *et al.* 2000; Reisch 2004; Majeský *et al.* 2012).

Simple sequence repetitive DNA (SSR) were not clustered in RepeatExplorer. However, some contigs were blasted with microsatellites (Chapter 6). Presumably they are not localized within closely similar genomic contexts and hence not detected by the graph-based clustering approach.

AFLPs have been used to analyse the diversity of the *Taraxacum* microspecies studied here. Majeský *et al.* (2012) observed high genotypic diversity among *Taraxacum* accessions after using the SSR and AFLP markers. Reamon-Büttner *et al.* (1999) used AFLP fragments for *in situ* hybridization in *Asparagus* and showed that many of the AFLP primer amplification products were repetitive DNA sequences, suggesting the primers were adjacent to these repeats. Majeský showed that AFLPs can reveal diversity between accessions, consistent with the result in Figure 5.5, showing that major repetitive sequence clusters differ in their abundance between microspecies.

Analysis of the genomes of the three agamospecies shows some RepeatExplorer clusters with characteristic differences in abundance (by copy number of the reads, and clusters detected, confirmed by strength of *in situ* hybridization) and in genomic location (by *in situ* hybridization) (Figure 5.13). This correlates with the genetic distance between the microspecies calculated from AFLP data (Majeský *et al.* 2012). Repetitive sequences are known to evolve in abundance and sequence between closely related such as the *Drosophila buzzatii* cluster (repeat group), (Kuhn *et al.* 2008). In plants, both satellite and retroelement-related sequences evolve in copy number and sequence at rates that allow subspecies to be distinguished (Saeidi *et al.* 2008; Contento *et al.* 2005). These studies reveal a lot about repetitive DNA evolution at the subspecies or within-genome levels. Somewhat contrasting methods have been used in the various studies. It will be very interesting if the speculative mechanisms of repeat evolution – unequal crossing over; replication slippage; homogenization of repeats; ‘concerted evolution’; ‘molecular drive’ – could be distinguished in sequence analysis of species with contrasting reproductive strategies.

5.5.2.3 Comparison with sexual species of *Helianthus*

In order to compare the repetitive component of the *Taraxacum* agamospermous genome with other sexual genomes, I chose *Hielianthus annus* as a model species in the Asteraceae family (Staton *et al.* 2012). The idea was to test if there is any correlation between mode of reproduction and the repetitive DNA component. The results from Staton *et al.* (2012) were very similar to the study of Natali *et al.* (2013) and Kane *et al.* (2011). Staton *et al*. used same procedure as the current study by using a graph based clustering method to analyse whole-genome shotgun (WGS) reads. Their results showed that the sunflower genome is composed of more than 81% transposable elements, 77% of which were LTR retrotransposons, and LTR-retrotransposons is the most abundant class of repetitive DNA sequences. In the *Taraxacum* genomes of the three microspecies studied here, the proportion of repetitive DNA is much lower (52-62%) and of which 26-36% is composed of transposable elements, with 25-32% LTR elements. These differences, assessed by the same methods in both genera, might be because of differences in the mode of reproduction: sexual in *Helianthus* and asexual in *Taraxacum*. Similar differences were demonstrated earlier in studies of different animal families in which the mode of reproduction also differed. Thus Arkhipova and Meselson (2000), Zeyl *et al.* (1996) and Goddard *et al.* 2001 all found that asexual taxa contain fewer transposable elements than sexual taxa.

Moreover, in the sunflower genome, the ratio of LTR-gypsy retrotransposons (58% of the genome, with the majority comprising chromovirus lineages) is higher than LTR-copia retrotransposons (20% of the genome). However in *Taraxacum* microspecies, the ratios of the two types are almost equal or the LTR-copia slightly exceeded the LTR-gypsy. LTR-copia is 13-16% and LTR-gypsy is 10-15%. Additionally, in other genomes such as papaya, Sorghum and rice, LTR-gypsy is much higher than LTR-copia (Ming *et al.* 2008; Paterson *et al.* 2009; Spannagl *et al.* 2009) and they are higher than in the sunflower genome. Nevertheless, DNA transposons occupied a small proportion (1.3%) of *Helianthus* repetitive DNA, similar to the *Taraxacum* genome in which they comprise 1-3%. LINE elements occupied a very small amount (0.6%) of the *Helianthus* genome, which is the same in *Taraxacum;* in some cases, however, they

were undetectable. It seems that the abundance of LINEs and DNA transposons is specific for plant genomes (Velasco *et al.* 2007; I.R.G.S. 2005; Hřibová *et al.* 2010), (Table 5.6).

In *Taraxacum* the ratio of singletons was different between species and comprised about 10-30% of the genomes (Figure 5.1). Different analysis methods can give very different results in studies of transposons, because of selectivity in data, comparison databases and scales analysis. Even in the data here, there are differences between 300bp and 150bp reads which are largely attributed to read length. Nevertheless, the results suggest that different species, even within Asteraceae, differ in their repetitive DNA structures, as has also been suggested in the Solanaceae (Richert-Polger et al. 2016).

5.5.2.4. Tandem repetitive DNA marker

In the RepeatExplorer outcomes I could identify the highly abundant novel tandemly repeated DNA in the three microspecies of *Taraxacum*. This is represented only by CL80 in the S3 genome output file. It consists of motifs 49 bp in length and repeated tandemly with a copy number of 32.4 in one contig (from tandem repeat finder) sized 1775 bp. It occupied about 1101 copies per S3 genome and was half of this ratio in the A978 and O978 genomes (443; 486 respectively). As shown by the *in situ* results it was visualised as a unique set of double dots at the centromeres of 14 chromosomes (Figure 5.12A-C, F, G). I further assembled the sequences of this tandemly repeated DNA in *Taraxacum* in comparison with the whole sequence of *Helianthus* (Staton *et al.* 2009) but there were no matches found. Except for CL80, no major tandemly repeated satellite DNA motifs were identified in *Taraxacum*. In the Asteraceae, Pires *et al.* (2004), Ruiz-Rejon Crepis (1993; 1995) identified TPRMBO (160 bp) and TGP7 (532 bp) as tandemly organized satellite sequences isolated from *Tragopogon*. After assembling at both low and high stringency, there was no homology to these two satellite sequences. Moreover there were no homology to satellite sequences of banana species which have no conspicuous satellite sequences (Dolezel et al. 1994).

5.5.4. *In situ* hybridization as a tool for identifying repetitive DNA

In order to understand the distribution and organisation of repetitive DNA motifs on the chromosomes, a key method is to use the repetitive DNA sequences as probes and hybridize them by fluorescent *in situ* hybridisation (FISH) to chromosomes (Heslop-Harrison *et al.* 1997). FISH-analysis of selected cluster sequences allowed further insight into the genome organisation of repetitive DNA sequences of the *Taraxacum* microspecies.

Mogie and Richards (1983) did a survey of the occurrence, nature and frequency of rDNA chromosomes in the agamospermous genus *Taraxacum.* They revealed that in most *Taraxacum* sections a characteristic satellite chromosome is seen with a large euchromatic region with an additional distal constriction in some chromosomes which indicates the position of a nuclear organiser region (NOR) (Heslop-Harrison and Schwarzacher 2011). Such chromosomes are characteristic by the presence of primary and secondary constriction sites. The 45S rDNA was located at the shorter arms. This chromosome most usually occurs at the frequency of one per haploid set of chromosomes. Consequently, it has been suggested that the NOR chromosomes correlate with “satellited chromosomes”, and that the filiform region corresponds to the nucleolar organiser region (NOR). They named this chromosome the '*Taraxacum* type' of satellite chromosome (Mogie and Richards 1983). We confirm the presence of six 5S rDNA and three 18S-5.8S-26S rDNA loci. The 45S rDNA sites localized at the secondary constriction at the end of the short arm of the chromosome, and showed a satellite connected by strands of *in situ* signal. This satellite is sometimes misleading in chromosome counts, resulting in counts of 27 instead of 24 chromosomes. It is characterised by a large sub-distal intercalary filiform region. This region is elastic and may vary in length as seen in *in situ* pictures. None of the 5S and 45S ribosomal RNA signal shared the same set of chromosomes.

Furthermore, CL132 from the O978 genome showed chromosome-specific signals. The probe from this cluster showed three signals on the same 45S rDNA chromosomes but on another primary constriction of the chromosomes. It has very strong signals on two chromosomes and a weaker one on the third chromosome. So

this probe is another chromosome-specific probe in *Taraxacum* which can be used for recognizing these chromosome triplets (Figure 5.12D, E, H).

Nevertheless, interestingly, fluorescence *in situ* hybridization (FISH) on mitotic chromosomes showed that elements from distinct repetitive DNA probes gave different signal strengths, locations and patterns of genomic distribution among the three *Taraxacum* microspecies. Thus, both the graph-based clustering and *in situ* hybridization showed that the microspecies, although morphologically closely related, differed.

In situ hybridization showed that LTR-gypsy elements were located in the centromeric region of *Taraxacum* chromosomes although, in some cases, the signal is absent. LTR-copia elements gave mostly broad centromeric signals or were dispersed through all chromosomes (Scmidt and Heslop-Harisson 1998; Santini *et al.* 2002). The different distributions of the different signals from the clusters annotated with LTR-Gypsy and LTR-Copia were probably because of the low similarity hits of cluster sequences that many of these clusters have.

It can be concluded that dispersed repetitive DNA is the major component in the *Taraxacum* genome and, like other angiosperms, these kind of repetitive DNAs could have a role in the genomic evolution and broad range of diversity found in agamospecies. Repetitive DNA sequences, along with cytological approaches, showed important differences among closely related microspecies, more so than more conventional chloroplast and nuclear genome markers. Thus, repetitive DNA markers can be use for comparative purposes to distinguish even closely related agamospecies.

Moreover, RepeatExplorer analysis showed that the repetitive component of *Taraxacum* genome consist of a large proportion of LTR repeats, especially LTR-copia and LTR-gypsy, which made up the major portion of LTR retrotransposons. In contrast, DNA transposons, LINEs, and satellite DNA were detected in very low percentages (some of them nearly undetectable). RepeatExplorer resulted in many clusters annotated as low-complexity; they are unknown sequences with unknown functions too.

CHAPTER 6

Frequency analysis of short DNA motifs (k-mers) to identify and characterize repetitive DNA components in the genomes of *Taraxacum* microspecies

Abstract

This chapter is about studing of the distribution of k-mer frequency from raw-read sequence data of three closely related *Taraxacum officinale* agg. agamospecies. The abundance, organization, and relationships of motifs from 16- to 132- bp long were measured and genomic distribution of major repeats was defined. Most previous work has relied on sampling known repeats or genome assemblies which often collapse or discard repetitive sequences.

Recent sequencing technologies nor very closely related accessions using the k-mer analysis. *In situ* hybridization showed differences between the three *Taraxacum* genomes and some amplified regions showed a similar copy number in the three genomes, others showed differences in relative abundance. A wide range of values of k was evaluated; the S3 reads of 300bp gave a substantially different pattern of occurrence frequency for highly abundant k-mers. This analysis shows the k-mer distribution is somewhat different reflecting different proportions of repetivity as part of the whole genome, and new class of repetitive DNA have been recognized as Passively Amplified DNA Sequences, PADS.

6.1. Introduction

Repetitive DNA is an abundant component of the genome of higher organisms, comprising many repeats (Satellite DNA, rDNA, and simple sequence repeats), and transposable elements (and derivatives such as solo long terminal repeats, solo LTRs), (Schmidt *et al.* 1999; Heslop-Harrison and Schwarzacher 2011). Much eukaryotic genomic diversity is seen in variability in repetitive DNA as a genomic component. Even closely related species may differ in genome size, partly a consequence of amplification of repeats and (Pearce *et al.* 1996), even though chromosomes pair at meiosis (Draper *et al.* 2001). The abundance of repetitive DNA and its variation makes analysis of nature and evolution challenging. Molecular and cytological analysis approaches have been key methods for characterization and isolation of a range of different repetitive DNA sequence elements in plants (Schmidt 1999; Kurtz *et al.* 2008; Biscotti *et al.* 2015).

Analysis of repetitive DNA with bioinformatic tools is becoming increasingly important with non-selective, high throughput (next generation) shotgun sequencing (NGS) technology. The repetitive components of the genomes have been analysed in several angiosperm crops such as banana, pea, soybean, barley and tobacco (Kubis *et al.* 1998; Kuhrová *et al.* 1991; Valárik *et al.* 2002; Kalendar *et al.* 2000; Macas *et al.* 2007*;* Hřibová *et al.* 2010), and model species including Brachypodium (Mur *et al.* 2011). Initially, repetitive elements were identified in genome assemblies, but many condense repetitive DNA (including tandem repeats and transposable elements). Assemblies will often concentrate transposable elements at the ends of scaffolds, in contigs that cannot be assigned to chromosomes or scaffolds, or discards many of the raw-reads of repeated sequences (eg. in oil palm, the assembly includes 57% of the reads, Singh *et al.* 2013, while many repeats are not included; Zaki *et al*. 2017; see also Kubis *et al.* 2003 for repeat distributions). Tandem repeats are collapsed even more frequently (Stacey *et al.* 2010 ; Saha *et al.* 2008; Kurtz *et al.* 2001). Analysis of assembled shorter regions, in particular BACs of 100-200kb long, may be accurate and show particular elements (Menzel *et al.* 2014; Nouroz *et al.* 2015). Graph-based clustering methods using raw-read high-throughput sequence methods have been

developed and identify many major repeat types (by Lysak *et al.* 1999; Macas *et al.* 2007; Novák *et al.* 2010; used eg. by Bomberly *et al.* 2016).

k-mer frequency analysis involves counting the frequency of defined DNA substrings of length k, in raw-reads of DNA sequence data. k-mers may be exploited in measuring genome sizes and in many genome assemblies, including the SOAP-de novo and Abyss assemblers. k-mers are also used in raw read sequence error correction (Pevzner *et al.* 2001; Kelley *et al.* 2010). For repetitive sequence, k-mer analysis is independent of assembly, and therefore an unbiased method to identify repeated DNA motifs in a genome (Bergman and Quesneville 2007; Marçais and Kingsford 2011). k-mers have been used to count the repetitive DNA sequences in bacteria by Williams *et al.* (2013), and Alkan *et al.* (2011) used k-mer approaches in eukaryotic genomes of some mammalian genomes to determine high-order repeat structures from raw read sequences, testing whether these sequences corresponded to functional centromeric regions followed by confirmation by *in situ* hybridization. Krassovsky and Henikoff (2014) confirmed that the most frequent k-mer sequences in *Drosophila melanogaster* were short repetitive DNA motifs and transposable elements. The frequency of k-mers can be counted from raw DNA sequence using several available tools, including Meryl (Myers *et al.* 2000), Tallymer (Kurtz *et al.* 2008), and jellyfish used here (Marçais *et al.* 2011).

The *Taraxacum* genus (Tribe Cichorieae, Asteraceae) is distributed worldwide, often in association with man-made habitats where it is abundant and successful. Apomictic *Taraxacum* species are commonly triploid (2n = 3x = 24), and there are also diploid sexual species (2n=2x=16) (Richards 1973; Nogler 1984; King and Schaal 1990; Singh 2002). The apomictic species produce viable seeds in the absence of fertilization and the generated offspring are identical to their mother (clone). There are thousands of morphological distinct apomictic clones within the *T. officinale* aggregate; a number are recognized as distinct taxonomic microspecies (agamospecies, Majesky *et al.* 2012).

6.2. Aims and objectives

The overall aims of this study are (1) to define the nature, abundance and large-scale genome organisation of repetitive sequences in *Taraxacum*; and (2) to find the diversity, evolutionary mechanisms and consequences of repetitive DNA sequences for the genome. In this chapter, the objectives were:

1. Based on 46 Gbases of Illumina raw reads from three *Taraxacum* agamospecies, what is the distribution of short k-mers (short sequence motifs where k is between 10 and 150 bp long)?
2. What is the nature of the most abundant k-mers?
3. Where are the most abundant k-mers located on the chromosomes of *Taraxacum*?

From the analysis answering these questions, I aimed to address questions about evolutionary biology:

1. What types of repeat reside in the *Taraxacum* genome? Are the abundant repeats related to transposable elements or tandem arrays (satellite types) of repeats with repeat-motifs from 4bp to 10kb long?
2. How does the genomic location of repeats relate to the type of sequence?
3. Are there differences in the most abundant k-mer motifs between three *Taraxacum* microspecies? (As a related question, are the same repetitive DNA sequences are present in the common ancestor of these *Taraxacum* agamospecies?)
4. Do the sequences types, locations and differences suggest evolutionary mechanisms (and time-scales) over which repetitive DNA has amplified?
5. Can k-mer derived probes be used to develop *Taraxacum* chromosome-specific cytogenetic markers or probes to identify chromosomes?

6.3. Materials and methods

6.3.1. Plant materials and sequencing data

Whole genome sequence data of the three closely related agamospecies (2n=3x=24) of *Taraxacum officinale* agg. [section *Taraxacum* (formerly *Ruderalia*), Asteraceae], *Taraxacum obtusifrons* Markl. (O978); *T. stridulum* Trávniček ined. (S3); and *T. amplum* Markl. (A978) (here referred to as A978, O978 and S3) were used in this study.

6.3.2. Illumina sequencing artefact

After we started working with the sequencing data, we found that a sequence (GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG AAAAAAAAAAGGGGGGG) is present in the sequencing data of A978 and O978 genome hundreds of thousands times. This matches the Illumina sequencing primer and is a sequencing artefact so these reads (0.65% of A978 genomic DNA; 1.39% of O978, genomic DNA) were deleted.

6.3.3. k-mer analyses

k-mer analysis was performed on Ubuntu Linux 13.10 (up to later 16.04; 32 Gb memory; OS-type 64-bit; 9 Tb disk), with Geneious version 7.1.4 and later (Kearse *et al.* 2012; available from http://www.geneious.com/). Using raw reads from each *Taraxacum* microspecies, the jellyfish k-mer counting program Version 2.1.3 (Marcais and Kingford 2011) was used to count canonical k-mers where k was equal to 6, each integer from 10 up to 21, 25, 32, 40, 44, 48, 50, 56, 60, 64, 75, 76, 80, 84, 88, 92, 96, 100, 110, 112, 128, 135, 140, 145, and 150 bp.

6.3.3.1. Counting k-mer frequency in raw reads

k-mer analysis were performed on paired-end reads in FASTA formatted sequences, which given the outcomes as k-mer counts (k-mer frequency) in a binary format. This was translated into human-readable text using the “jellyfish dump” command. Count-sequence pairs were processed to extract the most abundant motifs for each value of

k-mer repeated at thresholds of 10, 100, 1000, 10,000, and 100,000 or more times in the raw reads, using “grep” command.

Sequence motifs were imported into Geneious, and, unconventionally, a d*e novo* assembly of k-mer motifs of more than the threshold, overlapped staggered fragments of larger motifs. Resulting contigs were screened with BLASTn against the NCBI database and a customized retroelement motif database based on Hansen and Heslop-Harrison (2004) for conserved retroelement domains and internal protein motifs.

The “jellyfish histo” command was used to extract number of occurrences for each motif analysed from k-mer counts.

6.3.3.2. Genome size estimation

The *Taraxacum* genome size was estimated through the k-mer abundance distribution with the unassembled Illumina c. 45 Gb paired end reads. jellyfish was used to count k-mers with a 13-mer and the frequency distribution of the k-mers was plotted by the Microsoft Excel program.

6.3.4. Primer design and PCR amplification

After assembly of the abundant k-mers, contigs were chosen for PCR primer design. In particular, following the BLASTn analysis, most contigs were chosen from those with no significant similarity to the NCBI databases. Primers were designed within the Geneious program using Primer3. The sequences of primers and probes are listed in Table (6.1), giving forward and reverse sequences, annealing temperature, expected band size and GC% content of the sequence.

6.3.5. Molecular cytogenetic approach

6.3.5.1. DNA probe labelling

For rDNA, 5S rDNA probe labelled from 410bp fragment of the clone pTa794 (Gerlach and Dyer 1980) containing the 5S rDNA repeat unit of *Triticum aestivum*, and 45S rDNA from pTa71 containing a 9kb *Eco*RI fragment of the repeat unit of 25S-5.8S-18S rDNA isolated from *T. aestivum* (Gerlach and Bedbrook 1979).

Table 6.1. List of various primers designed from assembled contigs from k-mer generated sequences. Name of each primer, their forward and reverse sequences, GC%, annealing temperature, and expected band is given.

#	Oligo name	Forward Sequence (5' to 3')	Reverse Sequence (5' to 3')	GC%	annealing Tm	expected band size (bp)
1	128-mer_C10	ATTCATCCACCATCTGGGCC	TTGAGGCCCCATTTTTGTGC	38.8	63.6	514
2	128-mer_C10	CCCCTCCAACTTGCTTTTGG	ACTGGTCTTCCACGCTTCAG	37.6	61.05	520
3	128-mer_C29	GACTGACACATGCTGCGTTG	CCTTAATGCAATTGGGGCCC	38.5	62.45	809
4	128-mer_105	CTCGATCAGAAAAGTGATGC	GTTCGGGTGTTACTATCAGG	29.2	54.3	281
5	128-mer_C112	AGAGAACGAATGAAGACAGC	GGTAGCTTGAGTTTGTATGC	49.1	52.6	218
6	128-mer_C128	CGTACATGACACTTTTCACG	CAATTTCATCTGATCCTCGC	45.5	55.7	238
7	128-mer_C129	CTCTCAGGTTATCTGGATGC	GGTTTTAGAATGTTGTACCGG	35.4	54.4	127
8	128-mer_C154	CACAAGGTTGTTCATGAGC	TTAAGGGTAGGTGGGTATCC	50.4	54.3	137
9	128-mer_C233	TAGAATCAAAGTGCGGAATCC	TTCACCTCTCATATTAGAGC	34.1	53.1	220
10	128-mer_C60	CACTCAGTAAAAACAAGCGG	CAAGTGTTGGATTGTTCACC	49.9	55.3	830
11	128-mer_C62	GGAATTCAAGCGTAACAAGG	CACTCTTCTTTCAATTCGCC	44.2	56.3	303
12	128-mer_C102	CTTTTCGAGCTACAAACACC	ATAACATTCTGGGGGTATGC	44.4	54.6	277
13	128-mer_C31274	ATCCTAGTACTACTCTCGCC	ATTTCGGGTGCGATCATACC	51.7	55	89
14	128-mer_C15	AAGCAGGGATAGTATTTCGG	CATCCTAGTACTACTCTCGC	51.5	51.4	103
15	128-mer_C18	TAAACCCTAAACCCTAAACC	TTTAGGGTTTAGGGTTTAGG	41.4	51.4	116
16	128-mer_C32	GCGAAAAACTCCTATCTTCC	GTAGGTCTTCATCTTCAGGC	38.2	53.6	217
17	128-mer_C89	GCAACTTGGACAAAAAGAAACG	ATTGATTGGTCTGAGGTTATCG	37.1	58.1	70
18	64-mer_C40	GTCATTGTGTGAATCCATGC	TGACCCGTAGTAGCAATCG	49	56.4	200
19	135-mer_C49	AACAACCACATTTAATGAAAATCCACG	ACTCCAGTTGCATAGGACTTATCTAGG	36.6	63.65	352
20	135-mer_C36	CTTTGAACACTAATAACAATTGCATCG	TTTATAAATCTCCTTTTGGTACGTTGG	26.6	62.3	403
21	128-mer_C58	AGCCAGAGATCATTCGTAGGACG	TTTCCCTGGAGTCTTCGTCG	43.1	64.5	706
22	128-mer_C13	TTTTGGACACTATACAAGGAAGTTGG	TGCGTTACCCAGAGAGATGC	35.3	62.55	798
23	128-mer_C49	TCAACGTAGGAGACAAGCCG	GAAATCCCTACTTCGGGGGC	36.4	63.95	656
24	112-mer_C37	GTCTCGAATTTCAGCAGCGG	TTGCTACATGCTTGGTTGCC	42.1	64.05	817
25	112-mer_C34	TGCTGTTCTTTTGACTTTTGAGGG	ACGATTGTTGATTGTCTTCCCG	38.2	64.6	795
26	110-mer_C2	AGACTTAGATTCACAACACTACGG	TCTCTTTAGTGTAGTGTAGTGTTCCC	37.8	57.85	307
27	110-mer_C74	ACACCCCCAAACCAGTATGG	ATTGACATCCCCCACCTACG	33.9	63.2	582
28	100-mer_C99	GTTCGGTAAATACAATGCTTTGGG	TGTGAAGTGTTTGTGACTTGTGG	30.5	62.9	619
29	100-mer_C98	AGCTTAATATGAGTCCAACTGGAGC	CCAAGTGAACAAGATGTACATGGC	35.4	62.9	604
30	100-mer_C28	GGTAGAGGTTCAGGCATCCC	GGCCCTAGCAGTAAGTAACCG	39.6	61.65	608
31	96-mer_C38	ACTACCTCGATTGGATGCGC	AAGCATACGGCCTCTTTCGG	47	64.3	715
32	96-mer_C36	CCTTCTATGCTCAAATTTAGGGGG	CATCACTTTTCTGATCGAGTCCC	32.8	63.55	777
33	88-mer_C83	TGTTATGGTTCAGTTATTCGGATGC	TTCTCAACCCAAATGGAAGGC	25.9	64.3	269
34	88-mer_C74	AAGGGGTTGTTAATGAGAATTGTATGG	AGTATTACCAACATGTCTTGATGCC	33.3	62.55	609
35	84-mer_C93-1	CTTACTTCGTCCTCATTGTGGG	GCTGGAAACTCGAGCTTTGC	30.7	62.85	635
36	84-mer_C93-2	GCAAAGCTCGAGTTTCCAGC	CTGCAACAAGGAGGCATGC	37.2	64.1	514
37	84-mer_C45	CGGATATAGGATCTGAGTATACTGGG	ACGCTGCTCTAGGAGTAAGCC	26.5	61.15	604
38	76-mer_C22	CAAGTTAGTACACGAAACAACAAACC	TGAGATATAGAGACACATAGTGTTTCG	38.4	59.85	242
39	76-mer_C17	AATGGGATAAATGAAAGAACGAGCC	TTAGTGGTTTGTAGTGAAATAGTACGC	30.9	62.35	311
40	64-mer_C83	TGGTTTTGACATTAAGACATGGCC	CAACCAACTCATTAAGATCTTTTACGG	30.7	63.85	593
41	64-mer_C47	GTGCCTCAAAAACACTAAGAACGC	TCTTCCCATTAGTGATCGTATTCCC	37.7	64.15	489
42	60-mer_C76	GGCCCTAGCAGTAAGTAACCG	ACTCTCTGTGGACTCGACCC	34.8	61	515

Table 6.1. continued ….

#	Oligo name	Forward Sequence (5' to 3')	Reverse Sequence (5' to 3')	GC %	annealing Tm	expected band size (bp)
43	56-mer_C66	TCTCACCATCACGGATTAGACC	AATGACTTTCGAGGGTCGGG	34.3	63.65	478
44	56-mer_C65	ACCGAGTGAAAGCAATAGCC	TGTTTGCATGTCAGTGTGTGG	26.3	62	543
45	48-mer_C27	CTTTGAACACTAATAACAATTGCATCG	GGTGAATGCATAGTGATCTGAGAGTGG	28.2	64.8	436
46	44-mer_C10	TCATCTCATGTGATAAGATTTCCTACG	ATGAGAGAGAGTAGATTGGTGAATTGG	30.8	62.65	202
47	32-mer_C100	CACCCACGCTATTTGCAACG	GCTCTCGTTCATCAATACAATTACC	27.9	63.5	340
48	32-mer_C92	TCCCCACTTACGATGGTTAGC	AACTGTTAGGTTCGTTCTCTCG	41.2	60.6	301
49	16-mer_C963	GGATCCCGAATCTAGACATGGC	GTCCACTGACCCTTCTAGACG	43.8	62.65	178
50	32-mer_C71	CGGTTGAAACAAAGTCAAATCCC	ATCCGATAAAAACATACGCGCG	43.5	54.55	391
51	32-mer_C96	TGTGCCGAATCTCTACTGGC	CACTTGTCACACCACTTCCC	55	54.25	441
52	44-mer_C71	CACACTCTGTTCTCATGCCG	CTAGGACAGTATGCGGACCC	60	54.1	448
53	44-mer_C43	TGTATGATCACGCCTTCGCC	TGGGTGTTTGTTTGCATGGC	50	55.35	319
54	64-mer_C100	GGGACAGTCTGAATCACAGC	ATCTTGAGGGCTGGTGTTGG	55	54.15	487
55	64-mer_C71	CCAAATTCGGAAAGCACGGC	CACAATTGGTATCAGAGCGGG	52.4	54.9	650

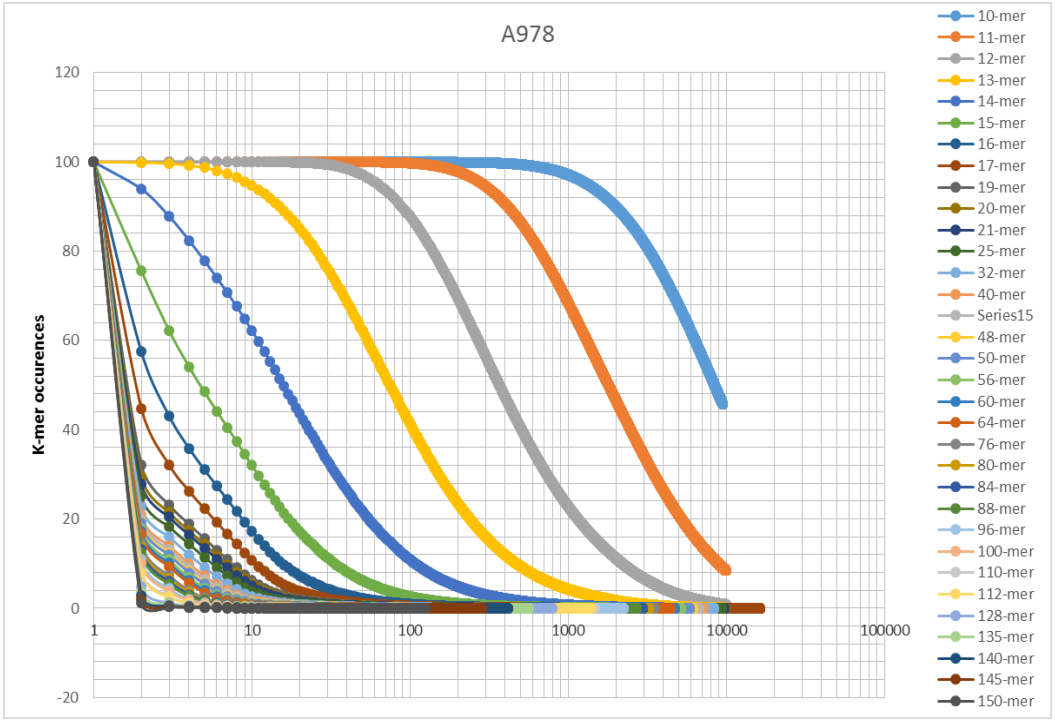
6.4. Results

6.4.1. k-mer frequency analysis in *Taraxacum*

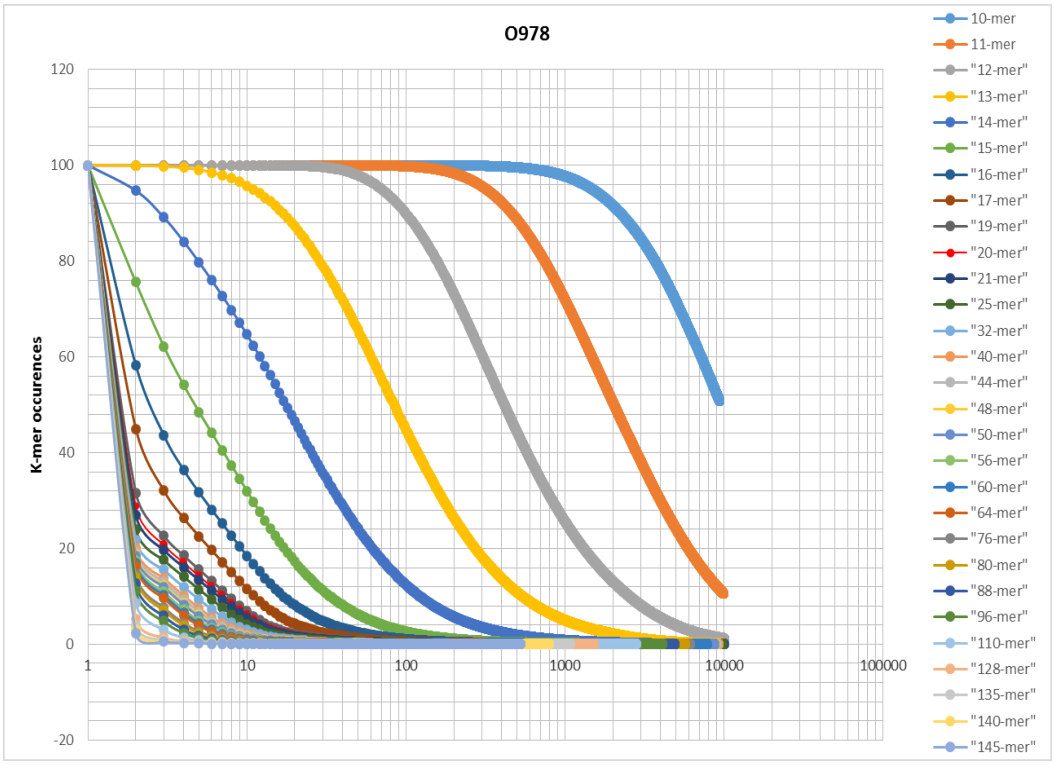
Occurrence of repetitive DNA in the three microspecies of *Taraxacum* genome was assessed using 37 values of k between 6-mers (128 possible canonical motifs) and 150-mers (2.5 x 10^8 motifs) for three *Taraxacum* microspecies. Figure 6.1 shows the graph plotted between numbers of k-mer (X axis) and their occurrences plotted on the Y axis, the slope represents the frequency of repetition in the genome (Bombarely *et al.* 2016). The overall k-mer result studied here through the graph indicates that the difference between the smaller k-mer lengths is higher than the differences present between the larger k-mer in the three *Taraxacum* microspecies. As the graph shown the differences between the k-mers decrease by the increasing of the k-mer size meaning that the information content of the k-mer set increases at a very fast rate from k = 10 to k = 17, beyond this point, increasing k does not significantly caused in increasing the number of unique k-mers, but does decrease the overall resolution of the k-mer set (Figure 6.1). In addition, the differences between the 19-mer up to 150-mer is slightly increased in the A978 and O978 genomes (Figure 6.1a, b), and same in S3 except the 128- and 135-mer which the differences started to increase (Figure 6.1c). Moreover, the slopes that made by the k-mer occurrences graphs indicate that the frequency of the repetitive sequences increase with the increasing of the k-length.

Figure 6.1. Histogram from a variety of k-mers value show *Taraxacum* microspecies repetivity analysed by k-mer frequency in raw reads of (a) A978, (b) O978 and (c) S3.

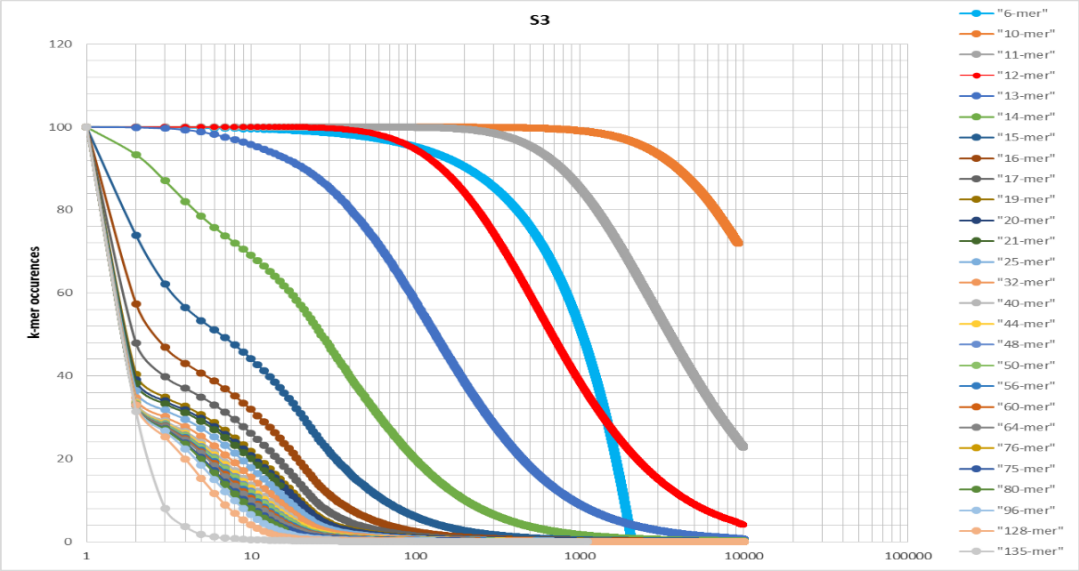
a



b



C



The distribution of all k-mers we have been working with compared among the three *Taraxacum* microspecies (A978, O978, and S3) separately (Figure 6.2), shown that the k-mer distribution in the *Taraxacum* microspecies A978 and O978 have very small differences in k-mer 10-14 but from 15-mer forward the figures belonging to A978 and O978 microspecies are overlapping on each other and exactly similar. However, the S3 k-mer is differ with the A978 and O978 genomes, in the distribution figures showing substantial differences when compared to the A978 and O978 genome in all values of k-mer. Very short values of k, all $k^4/2$ sequences are found in the reads with very high frequency (e.g. Figure. 6.2, 10-mer), and there were no unique k-mers. The analysis does not reveal signatures of the genome sequence under analysis.

The results represented the lower amount of repetitive sequences seen in A978 and O978 comparing with S3 genome (Figure 6.2, 15-mer), for example, all 15-mer occurring ≥10 times account for 32%, for A978, O978 and 44% in S3 genome, all 16-mer occurring ≥10 times account for 18%, 19% in A978, O978 respectively and 32% in S3 genome (Figure 6.2, 16-mer), and all 32-mer occurring ≥10 times account for just 3% in A978, O978 genome and 15% of S3 genome (Figure 6.2, 32-mer), and so on, as the rest of figures shown the repetitive contents decreased with the increase of the k length.

Figure 6.2. Histogram show comparative k-mer analysis between three *Taraxacum* microspecies by frequency of repetivity in raw reads. The account of each k-mer occurrence greater and equal than 10 times represented in a box on the top corner on the right side of each graph.

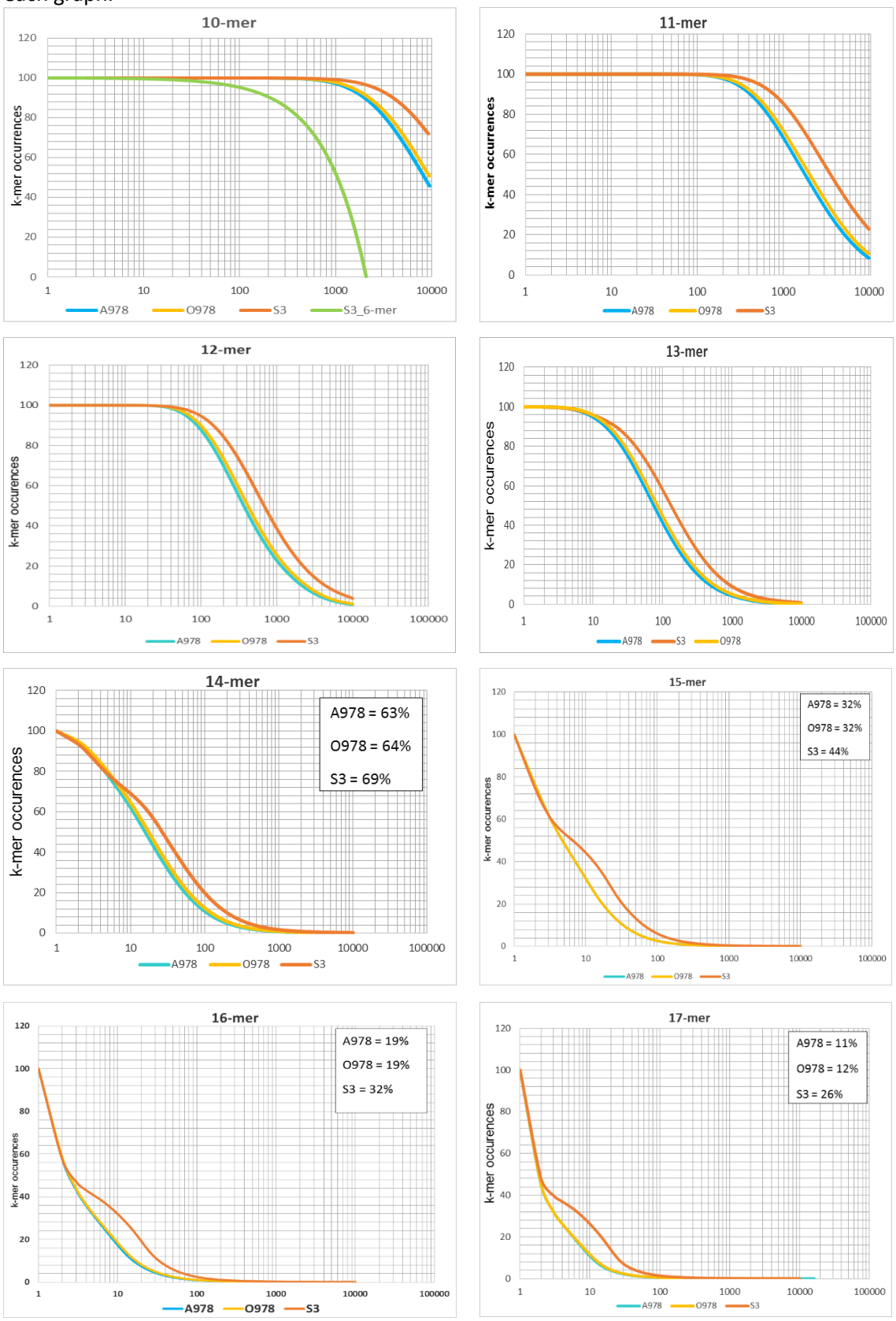
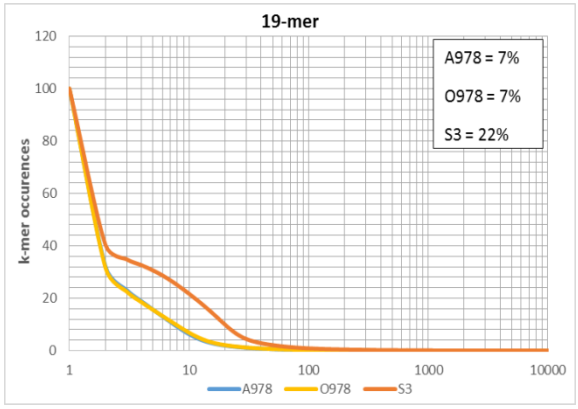
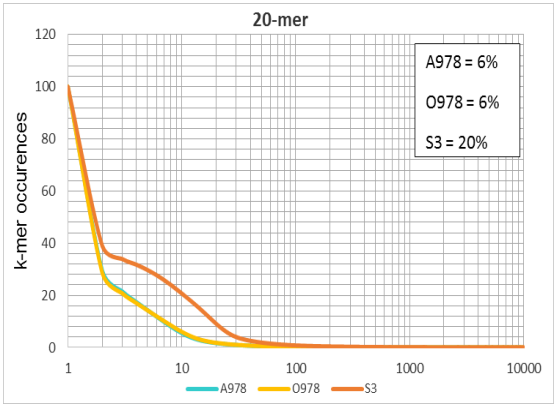
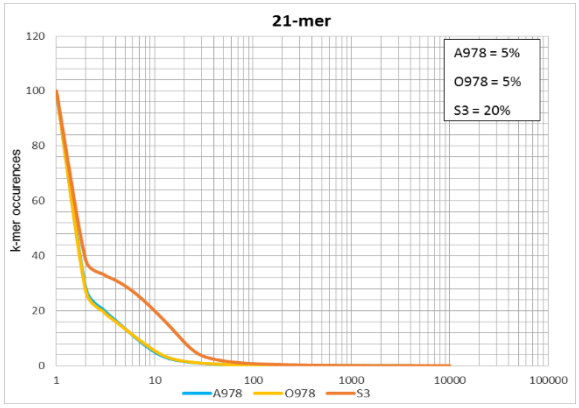
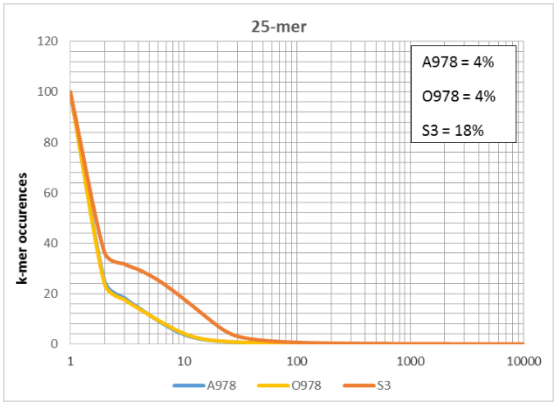


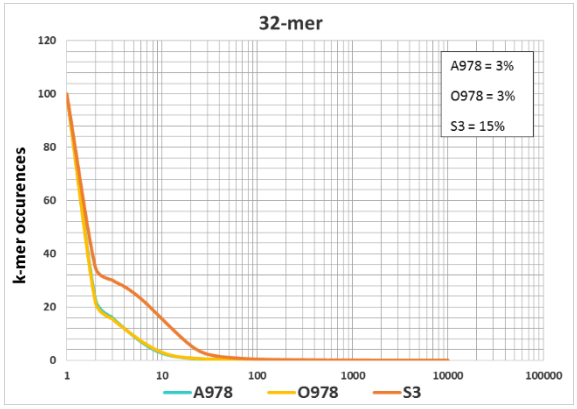
Figure 6.2. continue ….

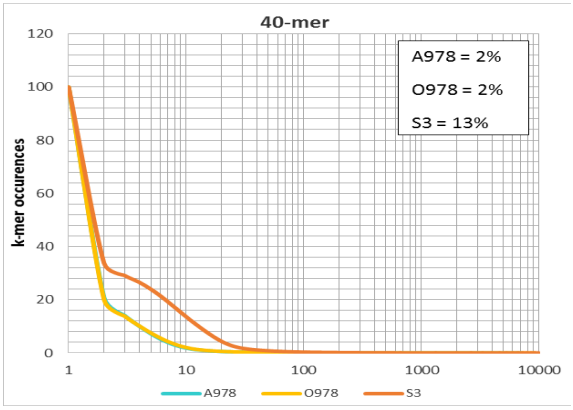


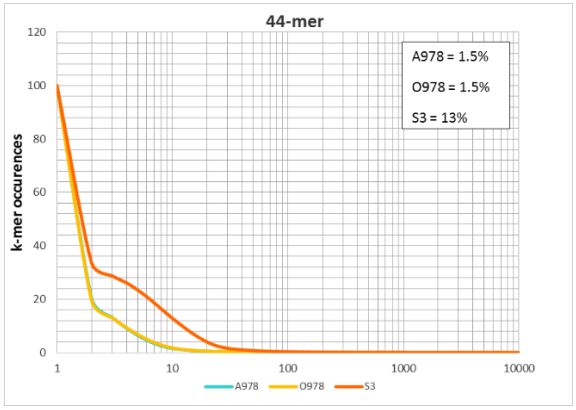












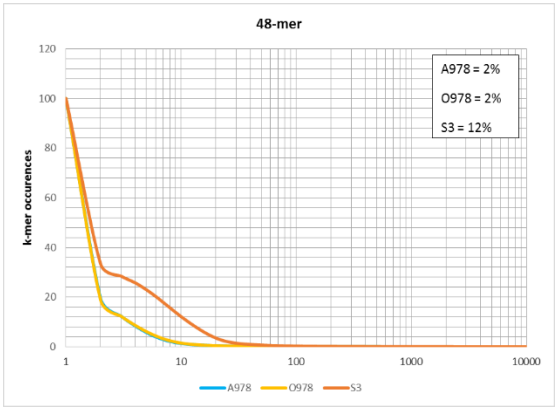
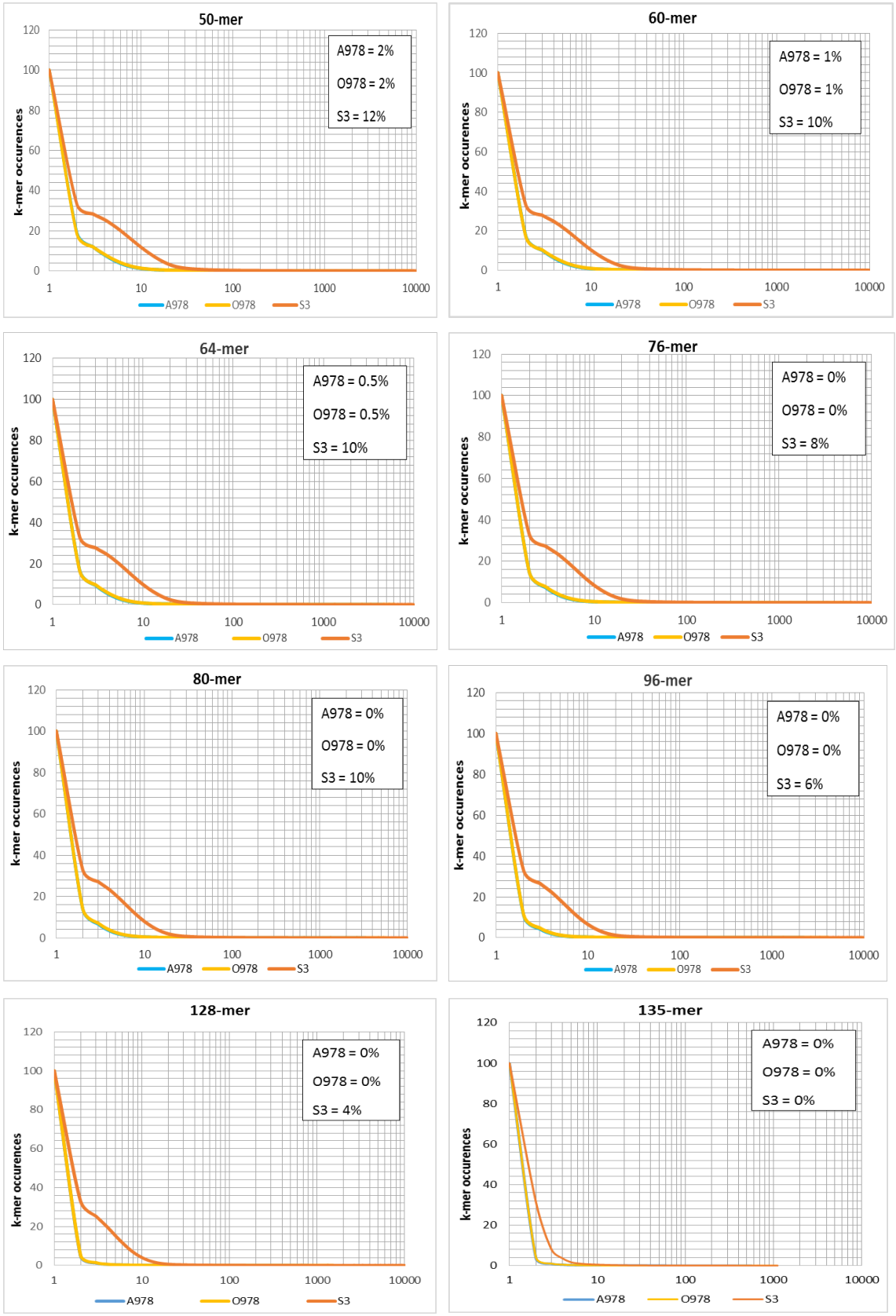


Figure 6.2. continue ….



6.4.2. *Taraxacum* genome sizes estimation

Genome size (base pairs of DNA in the genome) can be calculated from k-mer analysis. The genome sizes of various *Taraxacum* species have also been measured by flow cytometry and microdensitometry, and are recalculated in base pairs (Table 6.2) per single genome taking into account likely errors by the original authors and in databases. The 1Cx *Taraxacum* genome size (unreplicated haploid genome) has been reported as 815-880 Mbp.

Whole genome sequences of the *Taraxacum* microspecies were subjected to k-mer counting using the jellyfish program, with a k-mer size of 13 and the histogram of k-mer frequencies was plotted [X-axis k-mer coverage depth, and k-mer frequency as the Y-axis; so there were 170,000 individual parts of reads (Y-axis) with a particular k-mer that were present 26 times (X-axis)], for the three *Taraxacum* microspecies (Figure 6.3). The value giving the peak (here, at 26x) represents sequences which are present only once per DNA strand in the unreplicated haploid genome – that is the coverage of homozygous single copy sequences; heterozygous sequences would show a lower peak. Genome size was calculated with the formula:

Genome coverage depth = k-mer coverage depth × average read length / (average read length - k-mer size + 1), where the k-mer coverage depth is the maximal peak in the curve. Then, genome size was then estimated as follows: Genome size = total base number/genome coverage depth (for numerical data see Table 6.2, k-mer method).

The 18 Gb of S3 sequence, with a 301 bp read length, was able to show a peak in coverage frequency of 13-mers at 26x coverage (Fig 6.3 a). Repetitive sequences are represented in the 13-mers with greater coverage; taking into account the number of 13-mers per 301 bp read equal to (301-13+1), the 1Cx genome size is estimated at 658 Mbp. This is about 81% of the size estimated by microdensitometry or flow cytometry (815 mbp p- Table 6.2). The lower value may be accounted for by the relatively low coverage of the genome (most whole genome projects with the aim of assembly of the whole genome sequence use 100x or more coverage); the high proportion of repeats, perhaps heterozygosity between the three genomes; and/or evolutionarily recent genome duplication events. These factors also mean that other k-mer sizes do not show clear peaks, and that the two genomes with less sequence coverage did not show distinct peaks in the k-mer analysis (Figure 6.3 b).

Table 6.2. Genome size of *Taraxacum* agamospecies estimated by the Dolezel *et al.* (2003) and k-mer analysis-based method.

Species	ploidy level	2C	unreplicated haploid genome (1Cx_pg)	1Cx (Mbp)	Refrence
Dolezel *et al.* (2003) method					
T. linearisquameum (Ruderalia)	2	1.74	0.87	851	Záveský *et al.* 2005
Taraxacum spp.	3	2.61	0.87	851	Záveský *et al.* 2005
Taraxacum spp.	3	2.7	0.90	880	Záveský *et al.* 2005
T. officinale Weber. (Southern hemisphere)	3	5.3	0.88	864	Bennett *et al.* 1982
T. officinale Weber. (Northern hemisphere)	3	5.1	0.85	831	Bennett *et al.* 1982
T. officinale	2	2.5	1.25	815	Vidic *et al.* 2009
Recalculated as	*3*	*2.5*	*0.83*	*1222.5*	*Temsch et a. 2010*

K-mer method

species	read length (bp)	total illumina read number	haploid genome fold	k-mer length	total k-mer number	read depth	genome size (Mbp)
Taraxacum stridulum (S3)	301	59,258,642	20.9	13	17,125,747,538	26	658.7

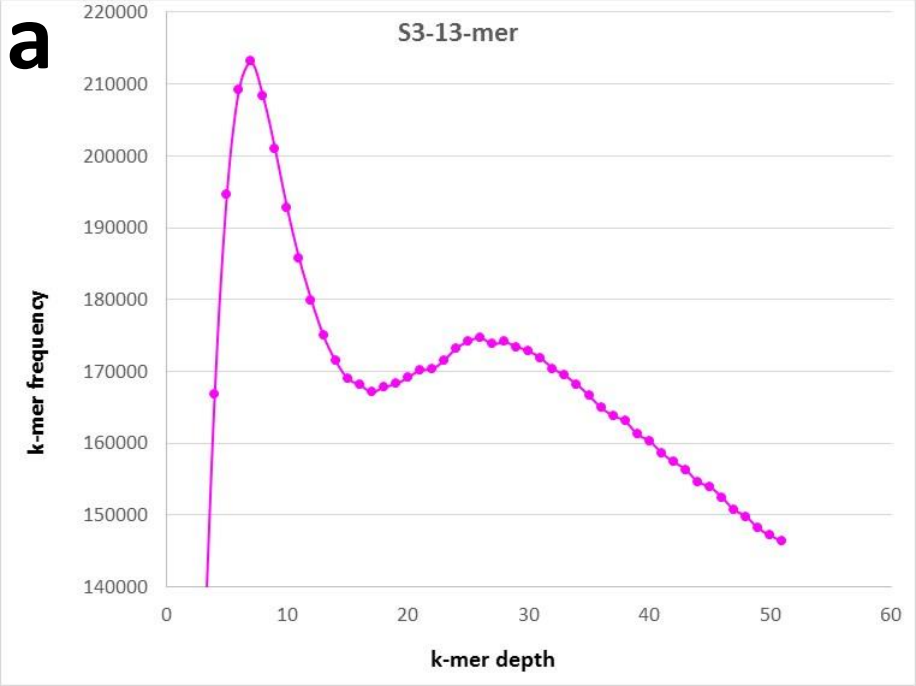
Whole genome sequences of the *Taraxacum* microspecies were subjected to k-mer counting using the jellyfish program, with a k-mer size of 13 and the histogram of k-mer frequencies was plotted [X-axis k-mer coverage depth, and k-mer frequency as the Y-axis; so there were 170,000 individual parts of reads (Y-axis) with a particular k-mer that were present 26 times (X-axis)], for the three *Taraxacum* microspecies (Figure 6.3). The value giving the peak (here, at 26x) represents sequences which are present only once per DNA strand in the unreplicated haploid genome – that is the coverage of homozygous single copy sequences; heterozygous sequences would show a lower peak. Genome size was calculated with the formula:

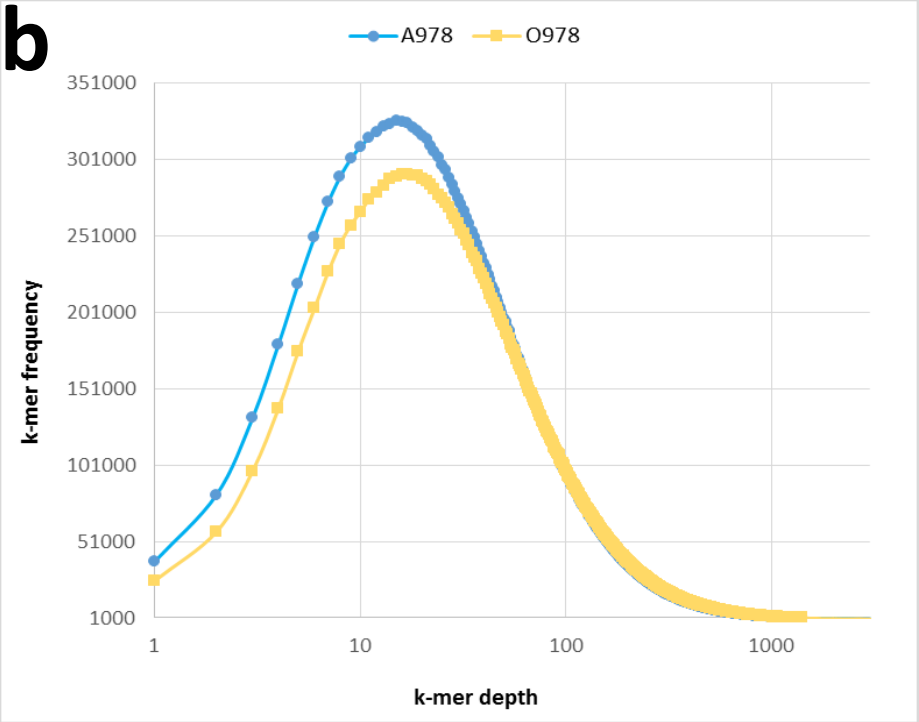
Genome coverage depth = k-mer coverage depth × average read length / (average read length - k-mer size + 1), where the k-mer coverage depth is the maximal peak in the curve. Then, genome size was then estimated as follows: Genome size = total base number/genome coverage depth (for numerical data see Table 6.2, k-mer method).

The 18 Gb of S3 sequence, with a 301 bp read length, was able to show a peak in coverage frequency of 13-mers at 26x coverage (Fig 6.3 a). Repetitive sequences are represented in the 13-mers with greater coverage; taking into account the number of 13-mers per 301 bp read equal to (301-13+1), the 1Cx genome size is estimated at 658 Mbp. This is about 81% of the size estimated by microdensitometry or flow cytometry (815 mbp p- Table 6.2). The lower value may be accounted for by the relatively low coverage of the genome (most whole genome projects with the aim of assembly of the whole genome sequence use 100x or more coverage); the high proportion of repeats, perhaps heterozygosity between the three genomes; and/or evolutionarily recent genome duplication events. These factors also mean that other k-mer sizes do not show clear peaks, and that the two genomes with less sequence coverage did not show distinct peaks in the k-mer analysis (Figure 6.3 b).

A k-mer analysis to estimate coverage of the genome and hence genome size could not be carried out for A978 and O978 with the more limited 12 Gb of sequence (10-12x coverage, within the residual peak of low coverage reads so no major peak from single copy DNA could be identified; Figure 6.3 b).

Figure 6.3. k-mer frequency distribution curve. A. represent k-mer curve for *Taraxacum stridulum* (S3), the curve show two peaks which the first peak is residual peaks and the secound one is major peak from single copy DNA.





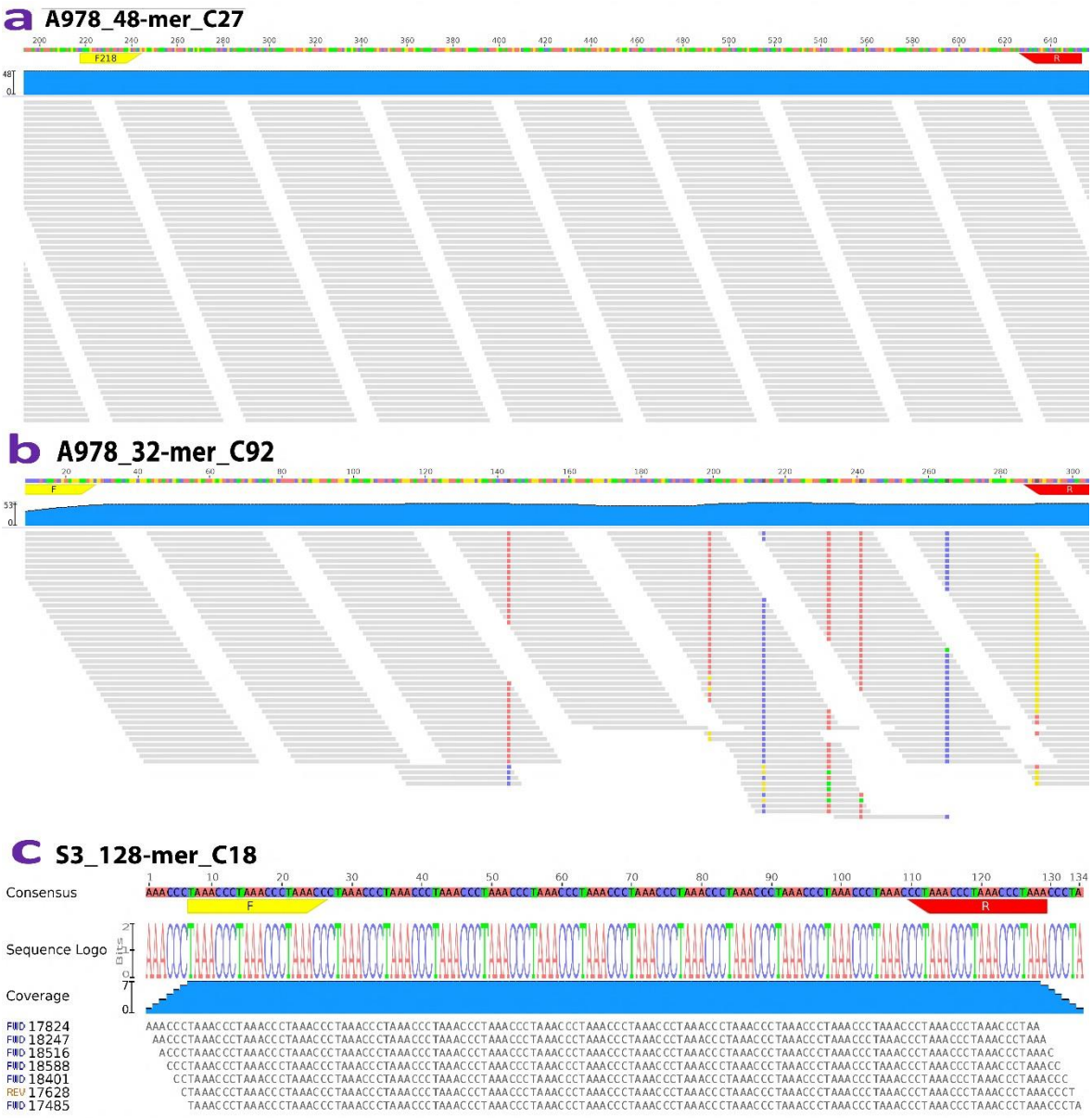
6.4.3. Assembly of k-mer outcomes and analyzing the resulted contigs

k-mer sequences with high abundances (>10, >100, >1 kbp, >10 kbp, >100 kbp) were extracted from the list of k-mers ("dump" file). These were then put into the assembly algorithm in Geneious to reassemble all the k-mers with similar or overlapping sequences. Several cut-offs were selected for 'high-abundance' for each k-mer length, typically to select the most abundant few hundred and few thousand k-mers. After assembly, all representative contigs were compared by BLASTn against the NCBI-GenBank sequence dataset to find any similarity hits with previously published repetitive sequences, and to avoid analysing of mitochondrial and chloroplast sequences.

Table 6.3 shows an example of an Assembly report and the subsequent choice of primers and then probes to amplify and label the abundant motif from genomic DNA.

The number of k-mers in the raw genome sequencing reads can be calculated as (4^k) with half being the reverse complement direction (canonical). For example, with 48-mers from A978, there were 143156 (of the possible $4^{48}/2= 3.96 \times 10^{28}$) sequences which occurred 1000 times or more. Of the 143156 sequences, 143096 were related and assembled into 1511 contigs between 49 bases long (i.e. reads overlapping by 46 or 47 bases, or with a small number of internal variant bases) and 9313 bases long (Table 6.3-10). Most of the contigs were compared by BLASTn to different known sequences present in the GenBank. Many of the contigs resulted in "No significant similarity" after BLAST search. Moreover, because the raw reads included 150 copies of chloroplast genome, chloroplast sequences were usually detected in the largest contigs (Figure 6.4); while pre-processing of the raw reads could have been carried out to remove chloroplast (and mitochondrial) sequences, there presence did not affect the analysis and provided a reference for the behaviour of a known abundant non-nuclear sequence class in the pipeline (including their fragmentation and isolation as k-mers, the process of contig assembly, coverage determination and homology analysis).

Figure 6.4. Assembly and primer bind for some of k-mer analysis from different *Taraxacum* microspecies genomes.



After BLAST, a representative range of sequences of abundant k-mers and k-mer derived contigs were chosen to isolate and label, so as to find their locations and distribution on the chromosomes by *in situ* hybridization, and investigation of the organization of their repetitive sequence types.

In an example of the k-mer analysis, we chose the 716 bp consensus sequence from 669 overlapped reads for primer design; Figure 6.4a shows the assembly picture with primer binding regions. Another example is shows 32-mers from A978: there were 330271 reads (of possible $4^{32}/2$= 9.2 x 10^{18}) sequences, which occurred 1000 times or more. Of the 330271 reads, only 364 reads were not assembled and did not relate to other sequences reads and overall the de novo assembly resulted in 4680 contigs between 33bp smallest (only 2, 32-mers assembling with a one base overlap at each end) and 20619bp as the largest contig length (Table 6.3.8). Contig 92 was chosen for designing primer to identify the repetitive sequences present in these reads from the 420 bp consensus sequence; there were no significant similarity hits to GenBank nor in the retrotransposons protein domains built by (Hansen and Heslop-Harrison 2004). Figure (6.4b) shows the assembly picture with primer binding regions.

Moreover, in attempt to identify activation of telomere repeat, 128-mers (Table 6.3.21), there were 5039 (of possible $4^{128}/2$=5.8 x 10^{76}) sequences which occurred 10000 times or more, were subjected to a *de novo* assembly. Of the 5039 reads assembled into just 19 contigs with the largest of 1017 bp and smallest with 130bp, and only one read not assembled and did not related to other sequences reads. The contig of 134 bp was found to be telomeric sequence, which consist of the 7 bp of telomeric sequences CCCTAAA/TTTAGGG repeated 19 times in the consensus sequence, Figure (6.4c) shows assembly picture with primer bind regions.

Similar analysis for a few other k-mer lengths and 'no significant similarity' BLAST searches were took place and the other tables from Table 6.3 are the assembly result of the different k-mer that used to design primers and hybridization probes used in current study.

Finally, from the selected contigs and relevant probes, the whole genome reads were mapped back to the probe sequence to give the coverage and an absolute value of the abundance of the particular probe (Table 6.4).

Table 6.3. List of tables show the assembly report of chosen k-mer sequences.

(1) A978_100-mer > 100

372,021 of 372,050 reads were assembled to produce 1,662 contigs; 29 reads were not assembled

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	29	1662	1662	64
Min Length (bp)	100	101	101	1006
Median Length (bp)		184	184	1296
Mean Length (bp)	100	358	358	3145
Max Length (bp)	100	49041	49041	49041
N50 Length (bp)		523	523	11901
Number of contigs >= N50		201	201	4
Length Sum (bp)	2900	595252	595252	201294

(2) A978_110mer > 100

263,252 of 263,278 reads were assembled to produce 973 contigs; 26 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	26	973	973	44
Min Length (bp)	110	111	111	1020
Median Length (bp)		194	194	1664
Mean Length (bp)	110	412	412	3863
Max Length (bp)	110	21575	21575	21575
N50 Length (bp)		628	628	7537
Number of contigs >= N50		83	83	7
Length Sum (bp)	2860	401582	401582	170005

(3) A978_110mer > 1000

1,215 of 1,216 reads were assembled to produce 21 contigs; 1 read was not assembled

Statistics	Unused Reads	All Contigs	Contigs >=100 bp
Number of	1	21	21
Min Length (bp)	110	111	111
Median Length (bp)		154	154
Mean Length (bp)	110	181	181
Max Length (bp)	110	545	545
N50 Length (bp)		180	180
Number of contigs >= N50		7	7
Length Sum (bp)	110	3804	3804

(4) A978_112mer >100

245,550 of 245,566 reads were assembled to produce 851 contigs;

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	16	851	851	45
Min Length (bp)	112	113	113	1013
Median Length (bp)		189	189	1472
Mean Length (bp)	112	433	433	3723
Max Length (bp)	112	21573	21573	21573
N50 Length (bp)		809	809	7535
Number of contigs >= N50		64	64	7
Length Sum (bp)	1792	368733	368733	167560

(5) A978_128mer >100

119,359 of 119,368 reads were assembled to produce 331 contigs; 9 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	9	331	331	38
Min Length (bp)	128	129	129	1000
Median Length (bp)		255	255	1869
Mean Length (bp)	128	580	580	2688
Max Length (bp)	128	18260	18260	18260
N50 Length (bp)		1181	1181	2800
Number of contigs >= N50		33	33	9
Length Sum (bp)	1152	192200	192200	102159

(6) A978_135mer >100

46,795 of 46,811 reads were assembled to produce 243 contigs; 16 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	16	243	243	20
Min Length (bp)	135	136	136	1090
Median Length (bp)		250	250	1432
Mean Length (bp)	135	468	468	2252
Max Length (bp)	135	8897	8897	8897
N50 Length (bp)		728	728	2726
Number of contigs >= N50		34	34	4
Length Sum (bp)	2160	113843	113843	45049

(7) A978_16mer >1000

662,813 of 663,149 reads were assembled to produce 9,535 contigs; 336 reads were not assembled

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	336	9535	2361	9
Min Length (bp)	16	16	100	1044
Median Length (bp)		49	162	1326
Mean Length (bp)	16	84	211	1274
Max Length (bp)	16	1551	1551	1551
N50 Length (bp)		135	227	1326
Number of contigs >= N50		1546	652	5
Length Sum (bp)	5376	809860	499254	11467

(8) A978_32mer >1000

329,907 of 330,271 reads were assembled to produce 4,680 contigs; 364 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	364	4680	961	37
Min Length (bp)	32	33	100	1002
Median Length (bp)		54	157	1753
Mean Length (bp)	32	102	293	2577
Max Length (bp)	32	20619	20619	20619
N50 Length (bp)		136	384	2820
Number of contigs >= N50		589	117	8
Length Sum (bp)	11648	478259	282497	95377

(9) A978_44mer >100

126,240 of 135,376 reads were assembled to produce 20,381 contigs; 9,136 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp
Number of	9136	20381	3637
Min Length (bp)	44	44	100
Median Length (bp)		66	125
Mean Length (bp)	44	77	138
Max Length (bp)	44	506	506
N50 Length (bp)		78	135
Number of contigs >= N50		6919	1427
Length Sum (bp)	401984	1583242	504525

(10) A978_48mer >1000

143,096 of 143,156 reads were assembled to produce 1,511 contigs; 60 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	60	1511	572	16
Min Length (bp)	48	49	101	1002
Median Length (bp)		82	164	1258
Mean Length (bp)	48	148	281	2244
Max Length (bp)	48	9313	9313	9313
N50 Length (bp)		203	347	3488
Number of contigs >= N50		219	99	3
Length Sum (bp)	2880	224402	160988	35907

(11) A978_56-mer >100

1,741,449 of 1,741,697 reads were assembled to produce 13,483 contigs; 248 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	248	13483	8222	105
Min Length (bp)	56	57	100	1000
Median Length (bp)		117	186	1216
Mean Length (bp)	56	194	270	2449
Max Length (bp)	56	42029	42029	42029
N50 Length (bp)		272	327	2324
Number of contigs >= N50		2475	1819	9
Length Sum (bp)	13888	2616598	2225462	257190

(12) A978_60-mer >100

1,539,822 of 1,540,003 reads were assembled to produce 11,227 contigs; 181 reads were not assembled.

Statistics	Unused eads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	181	11227	7204	121
Min Length (bp)	60	61	100	1004
Median Length (bp)		123	189	1187
Mean Length (bp)	60	207	280	2268
Max Length (bp)	60	74870	74870	74870
N50 Length (bp)		292	345	2166
Number of contigs >= N50		2024	1536	11
Length Sum (bp)	10860	2330939	2020696	274443

(13) A978_64-mer >100

1,356,674 of 1,356,818 reads were assembled to produce 9,241 contigs; 144 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	144	9241	6200	117
Min Length (bp)	64	65	100	1004
Median Length (bp)		132	190	1254
Mean Length (bp)	64	223	293	2352
Max Length (bp)	64	74869	74869	74869
N50 Length (bp)		320	372	1901
Number of contigs >= N50		1610	1258	11
Length Sum (bp)	9216	2063633	1822056	275261

(14) A978_76-mer >1000

30,248 of 30,253 reads were assembled to produce 198 contigs; 5 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	5	198	118	8
Min Length (bp)	76	77	100	1118
Median Length (bp)		113	152	1799
Mean Length (bp)	76	240	344	2182
Max Length (bp)	76	5283	5283	5283
N50 Length (bp)		420	635	1992
Number of contigs >= N50		19	12	3
Length Sum (bp)	380	47679	40673	17460

(15) A978_84-mer >100

683,819 of 683,883 reads were assembled to produce 3,624 contigs; 64 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	64	3624	2997	92
Min Length (bp)	84	85	100	1000
Median Length (bp)		163	190	1232
Mean Length (bp)	84	294	336	2673
Max Length (bp)	84	49052	49052	49052
N50 Length (bp)		431	464	3758
Number of contigs >= N50		552	487	5
Length Sum (bp)	5376	1066322	1008621	245917

(16) A978_88-mer >100

593,348 of 593,396 reads were assembled to produce 2,984 contigs; 48 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	48	2984	2622	84
Min Length (bp)	88	89	100	1005
Median Length (bp)		171	188	1271
Mean Length (bp)	88	311	340	2805
Max Length (bp)	88	53550	53550	53550
N50 Length (bp)		442	474	21320
Number of contigs >= N50		439	401	4
Length Sum (bp)	4224	928092	894089	235703

(17) A978_96-mer >100

441,878 of 441,919 reads were assembled to produce 2,052 contigs; 41 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	41	2052	1972	73
Min Length (bp)	96	97	100	1001
Median Length (bp)		180	189	1295
Mean Length (bp)	96	340	350	2982
Max Length (bp)	96	49049	49049	49049
N50 Length (bp)		507	519	13384
Number of contigs >= N50		264	256	4
Length Sum (bp)	3936	698235	690382	217739

(18) O978_32-mer >10000

2,327 of 2,339 reads were assembled to produce 84 contigs; 12 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp
Number of	12	84	15
Min Length (bp)	32	33	103
Median Length (bp)		49	301
Mean Length (bp)	32	95	305
Max Length (bp)	32	570	570
N50 Length (bp)		265	364
Number of contigs >= N50		11	6
Length Sum (bp)	384	8061	4575

(19) O978_64-mer >100

1,627,445 of 1,627,983 reads were assembled to produce 11,410 contigs; 538 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	538	11410	7655	137
Min Length (bp)	64	65	100	1000
Median Length (bp)		128	188	1195
Mean Length (bp)	64	217	285	2217
Max Length (bp)	64	42036	42036	42036
N50 Length (bp)		308	357	2396
Number of contigs >= N50		2040	1588	17
Length Sum (bp)	34432	2486559	2187314	303854

(20) S3_128-mer >1000

8,864 of 8,914 reads were assembled to produce 307 contigs; 50 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	50	307	307	56
Min Length (bp)	128	129	129	1027
Median Length (bp)		283	283	1984
Mean Length (bp)	128	728	728	2610
Max Length (bp)	128	9545	9545	9545
N50 Length (bp)		1899	1899	2935
Number of contigs >= N50		31	31	15
Length Sum (bp)	6400	223725	223725	146179

(21) S3_128-mer >10kb

5,038 of 5,039 reads were assembled to produce 19 contigs; 1 read was not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs>=1000 bp
Number of	1	19	19	1
Min Length (bp)	128	130	130	1017
Median Length (bp)		312	312	1017
Mean Length (bp)	128	405	405	1017
Max Length (bp)	128	1017	1017	1017
N50 Length (bp)		441	441	1017
Number of contigs >= N50		6	6	1
Length Sum (bp)	128	7698	7698	1017

(22) S3_64-mer >10000

6,466 of 6,474 reads were assembled to produce 67 contigs; 8 reads were not assembled.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	8	67	37	5
Min Length (bp)	64	65	102	1136
Median Length (bp)		107	192	2451
Mean Length (bp)	64	539	911	5186
Max Length (bp)	64	15420	15420	15420
N50 Length (bp)		5718	5718	15420
Number of contigs >= N50		2	2	1
Length Sum (bp)	512	36158	33736	25930

In the bioinformatics analysis, numerous variables could have been adjusted, including thresholds of k-mer repeat number for assembly, parameters for the overlap and number of mis-matches or indels allowed during assembly, and nature of GenBank sequence comparisons (type, e.g. Megablast, BLASTn, BLASTx, discontiguous megablast; and sensitivity). The aim was to identify repetitive DNA motifs in an unbiased manner from the reads, so all the parameters were selected empirically based usually on program defaults. The copy number of resulting sequences was determined accurately from the reads. Notably, empirical values are also used in genome assembly algorithms (for example, to choose the k-mer length giving best assembly, or for similarity scores before joining reads into clusters, into contigs, and then into scaffolds) where there is no verification of the assembly, leading to major revisions to whole genome sequences when reanalysed or when new approaches become available.

6.4.4. Chromosomal localization of high-frequency k-mer-derived contigs

PCR primers were designed from the k-mer derived contigs, and those showing amplification are given in Table 6.1. Amplified products were labelled with biotin or digoxigenin and used for *in situ* hybridization to metaphase chromosome spreads from the three *Taraxacum* microspecies (O978, S3 and A978). While each probe showed generally similar organization in the three microspecies, there were significant differences, described below, for some probes.

The abundance of each probe in the three genomes was measured by counting the number of reads assembling to the sequence spanned by the primers and the copy number was calculated per genome for the three *Taraxacum* microspecies (Table 6.4), and ploted to show the differences between the three genomes (Figure 6.4). While some amplified regions showed a similar copy number in the three genomes, others showed differences in relative abundance (number of raw reads in the amplified contig region) between the S3, O978 and A978 genomes (Figure 6.1). However, for most sequences, copy numbers differed by less than 2-fold.

Figure 6.6 A, B, C shows the *in situ* hybridization pattern of a 128-mer_C129. In A978 the probe is terminal on some of the chromosomes and some centromeric regions but signal strength is relatively low. In O978, the probe is more centromeric in location, but absent or very weak on three chromosomes (not satellite chromosomes). In agreement with the bar-chart of sequence abundance (Figure 6.5) the probe is more abundant in the S3 genome (Figure 6.6 C), where it is present on all chromosomes with a broad centromeric location.

The probe from 64-mer_C32 also shows differences between the abundance of the probe in the three genomes by *in situ* hybridization (Figure 6.6 D, E, F) and copy number analysis (Figure 6.5). The *in situ* hybridization picture shows different locations of the probes on the chromosomes of the three genomes: in A978 (Figure 6.6 D) there are some locations excluded from centromeres, while the probe labelled one or both arms, and some other locations and chromosomes are weakly labelled. In O978 (Figure 6.6 E), the probe labels about two-thirds of the chromosomes with very strong signals; however there are about 9 chromosomes have very weak signals or the signal is absent. In S3, the probe shows signals on all of the chromosomes except satellite chromosomes which have very weak signals: the signals locations are mostly on the centromeres with some gaps. In contrast, the copy number analysis shows that S3 has the highest abundance.

Table 6.4. coverage and an absolute value of the abundance of the particular probe (primer regions) assembled to whole genome reads.

#	Genome	repeated more than	Name	length (bp)	assembled to "A978" genome				assembled to "O978" genome				assembled to "S3" genome			
					No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome	No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome	No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome
1	A978	100	100-mer_C28	608	19,986	92.7	4964	477	23704	93	5887	483	17899	90	8861	424
2	A978	100	100-mer_C98	604	19,929	94.5	4982	479	22504	93	5626	461	20440	91	10186	487
3	A978	100	100-mer_C99	6019	26,762	91.3	671	65	34321	92	861	71	27916	87	1396	67
4	A978	1000	110-mer_C2	307	43,666	93.4	21477	2065	42326	93	20818	1706	73725	86	72284	3459
5	A978	100	110-mer_C74	582	5,384	80.0	1397	134	34976	91	9075	744	33192	84	17166	821
6	A978	100	112-mer_C34	795	49,103	89.7	9326	897	58327	89	11078	908	52554	88	19898	952
7	A978	100	112-mer_C37	817	35,931	93.7	6641	639	41246	93	7623	625	35460	92	13064	625
8	A978	100	128-mer_C49	656	24,886	94.0	5728	551	24494	95	5638	462	23643	89	10848	519
9	A978	100	128mer_C13	798	68,203	94.1	12906	1241	69171	94	13089	1073	51945	95	19593	937
10	A978	100	128mer_C58	706	29,109	92.6	6226	599	30098	92	6437	528	31237	89	13318	637
11	A978	100	135me_C49	656	56,876	84.6	13092	1259	57164	87	13158	1079	74871	78	34354	1644
12	A978	100	135mer_C36	403	18,434	83.7	6907	664	19901	83	7457	611	18783	73	14029	671
13	A978	1000	16-mer_C963	178	7,078	92.6	6004	577	4677	92	3968	325	2461	85	4162	199
14	A978	1000	32-mer_C100	340	21,080	87.5	9362	900	25012	87	11108	911	19963	80	17673	846
15	A978	1000	32-mer_C92	301	22,704	82.6	11390	1095	26561	82	13325	1092	17075	68	17075	817
16	A978	100	44-mer_C10	202	20,595	84.9	15395	1480	50812	88	37983	3113	19583	73	29181	1396
17	A978	1000	48-mer_C27	436	21,056	93.9	7292	701	21923	94	7593	622	22122	65	15272	731
18	A978	100	56-mer_C65	543	2,036	95.5	566	54	2063	96	574	47	2055	92	1139	55
19	A978	100	56-mer_C66	478	16,856	80.6	5325	512	20847	77	6586	540	21324	66	13428	642
20	A978	100	60-mer_C76	515	20,946	88.9	6141	591	23825	90	6986	573	21996	82	12856	615
21	A978	100	64-mer_C47	489	19,004	85.2	5868	564	24268	86	7494	614	18431	70	11345	543
22	A978	100	64-mer_C83	593	23,245	88.2	5919	569	22074	87	5621	461	34560	88	17542	839
23	A978	1000	76-mer_C17	311	37,050	90.8	17989	1730	33941	92	16479	1351	32259	83	31222	1494
24	A978	1000	76-mer_C22	242	12,470	95.2	7781	748	12757	95	7960	652	14027	92	17447	835
25	A978	100	84-mer_C45	604	45,338	89.8	11335	1090	52522	90	13131	1076	24654	72	12286	588
26	A978	100	84-mer_C93-1	636	30,183	91.1	7166	689	35439	91	8414	690	29071	85	13758	658
27	A978	100	84-mer_C93-2	514	54,108	92.5	15896	1528	60478	92	17767	1456	60197	87	35252	1687
28	A978	100	88-mer_C74	609	39,561	93.0	9809	943	49539	93	12283	1007	31067	92	15355	735

Table 6.4. continued ….

#	Genome	repeated more than	Name	length (bp)	assembled to "A978" genome				assembled to "O978" genome				assembled to "S3" genome			
					No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome	No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome	No. of reads assembled	Pairwise identity (%)	sequence depth	copies per genome
29	A978	100	88-mer_C83	648	39,006	92.1	9089	874	43749	92	10195	836	37324	88	17337	830
30	A978	100	96-mer_C36	777	37,738	90.7	7334	705	42648	90	8288	679	40862	82	15829	757
31	A978	100	96-mer_C38	715	22,703	91.9	4795	461	26363	92	5568	456	31590	89	13299	636
32	O978	100000	32-mer_C71	391	36,696	87.4	14172	1363	42572	89	16441	1348	47518	83	36580	1750
33	O978	100000	32-mer_C96	441	1,345	95.1	461	44	1696	95	581	48	2267	94	1547	74
34	O978	100	44-mer_C43	319	78,083	83.8	36961	3554	65144	85	30836	2528	93701	73	88414	4230
35	O978	100	44-mer_C71	448	55,929	86.8	18851	1813	60210	88	20294	1663	75509	79	50733	2427
36	O978	100	64-mer_C100	487	11,025	94.5	3418	329	12546	95	3890	319	11247	89	6951	333
37	O978	100	64-mer_C71	650	20,641	90.5	4795	461	24504	91	5692	467	21263	84	9846	471
38	S3	1000	128-mer_C10-6	521	57,539	94.0	16676	1603	65005	94	18840	1544	65763	91	37994	1818
39	S3	1000	128-mer_C10-5	534	54,835	94.2	15506	1491	60793	94	17191	1409	60332	93	34007	1627
40	S3	1000	128-mer_C10-7	476	68,892	93.6	21854	2101	76369	94	24226	1986	76785	92	48555	2323
41	S3	1000	128-mer_C102	277	21,123	90.3	11515	1107	24212	90	13199	1082	33678	82	36596	1751
42	S3	1000	128-mer_C105	281	13,360	90.4	7179	690	15209	90	8173	670	18591	87	19914	953
43	S3	1000	128-mer_C112	218	29,499	83.6	20433	1965	33101	85	22928	1879	49766	78	68714	3288
44	S3	1000	128-mer_C128	238	32,949	87.1	20905	2010	38681	87	24541	2012	65102	78	82335	3939
45	S3	1000	128-mer_C129	127	18,916	91.5	22491	2163	20556	92	24441	2003	30682	86	72719	3479
46	S3	10000	128-mer_C15	103	47,152	78.4	69126	6647	54329	76	79647	6528	204941	83	598905	28656
47	S3	1000	128-mer_C154	137	13,360	83.4	14725	1416	34540	81	38070	3120	32806	77	72077	3449
48	S3	10000	128-mer_C18	116	52,009	64.0	67701	6510	21523	60	28017	2296	35842	66	93004	4450
49	S3	1000	128-mer_C233	220	13,378	72.2	9182	883	14429	70	9904	812	13082	56	17899	856
50	S3	10000	64-mer_C32	217	11,418	97.1	7945	764	14740	97	10257	841	32745	95	45420	2173
51	S3	1000	128-mer_C60	830	64,205	94.5	11681	1123	145212	95	26418	2165	87789	95	31837	1523
52	S3	1000	128-mer_C62	303	17,397	94.4	8670	834	19749	94	9842	807	23765	92	23608	1130
53	S3	10000	64-mer_C80	70	6,543	96.0	14114	1357	8403	96	18126	1486	25572	97	109960	5261
54	S3	10000	64-mer_C40	281	23,020	92.5	12370	1189	47629	89	25594	2098	40451	87	43330	2073
55	S3	10000	128-mer_SHC31274	89	56,864	78.5	96477	9277	52809	78	89597	7344	200151	83	676915	32388

sequence depth = (number of reads assembled * average read length (bp)/ reference sequence length
es per genome = sequence depth / genome sequence fold

Figure 6.5. The figure from numerical data from (table 6.4) show abundance of probe (primer regions) in the three *Taraxacum* genomes.

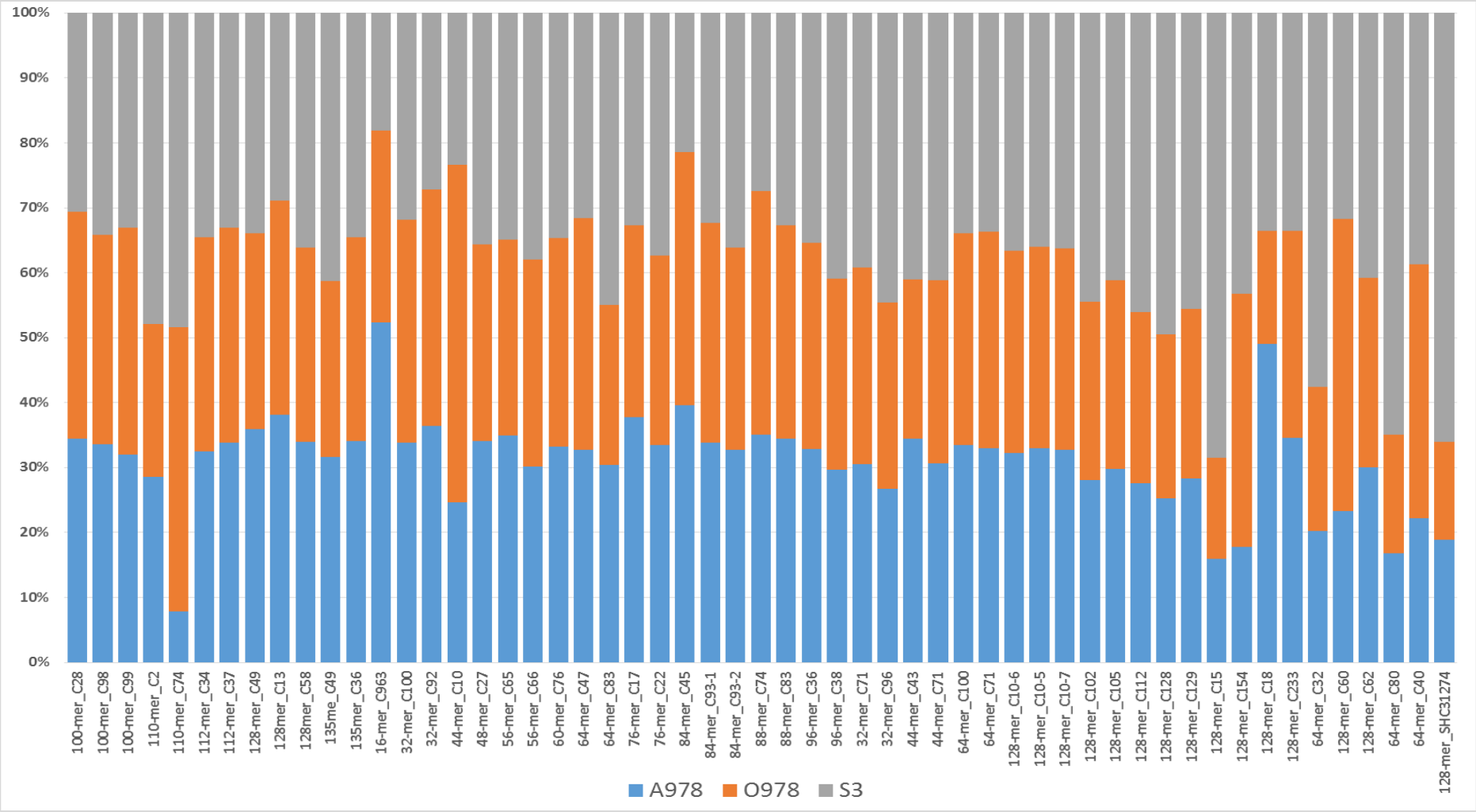


Figure (6.6 G, H, I) shows 128-mer_C58, where overall the probe pattern shown mostly centromeric location on the metaphase of the three genomes, with some differences in the abundance between the three genomes and a lower strength of hybridization to A978. Figure 6.5 shows that the probe is similar but slightly more abundant in S3 than the other two genomes.

In Figure 6.7 A, B, C, the genome of S3 shows a relatively uniform distribution of the 128-mer_C105 all over chromosomes with no clear blue regions without the probe; the pattern is around centromeres of all chromosomes in O978 and A978 genome which represent some different strength among chromosomes and the signals is absent on some of the chromosomes and strong on others of O978. Copy number (Figure 6.5) reflects hybridization strength with relatively most copies in O978.

The probe of 44-mer_C10 is abundant in O978 and the signal is centromeric with a few gaps seen on some chromosomes. In comparison to A978, signals are absent or very weak on some of the chromosomes, and in S3, the signals is weaker with gaps at many centromeres. The sequence is in highest abundance in O978 genome and less in S3 and A978 (Figure 6.5).

The 64-mer_C83 probe is highly abundant in S3 and has lowest abundance in O978. But in the *in situ* picture (Figure 6.7 G, H, I) the probe shows a diffused pattern on the chromosomes of the three genomes; in O978 there was one group of three chromosomes with very strong signals that were not present in the S3 and A978 genomes.

Figure 6.6. Probes used to show differences in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH pattern comes from the k-mer written beside the picture, and the *Taraxacum* microspecies of the metaphase.

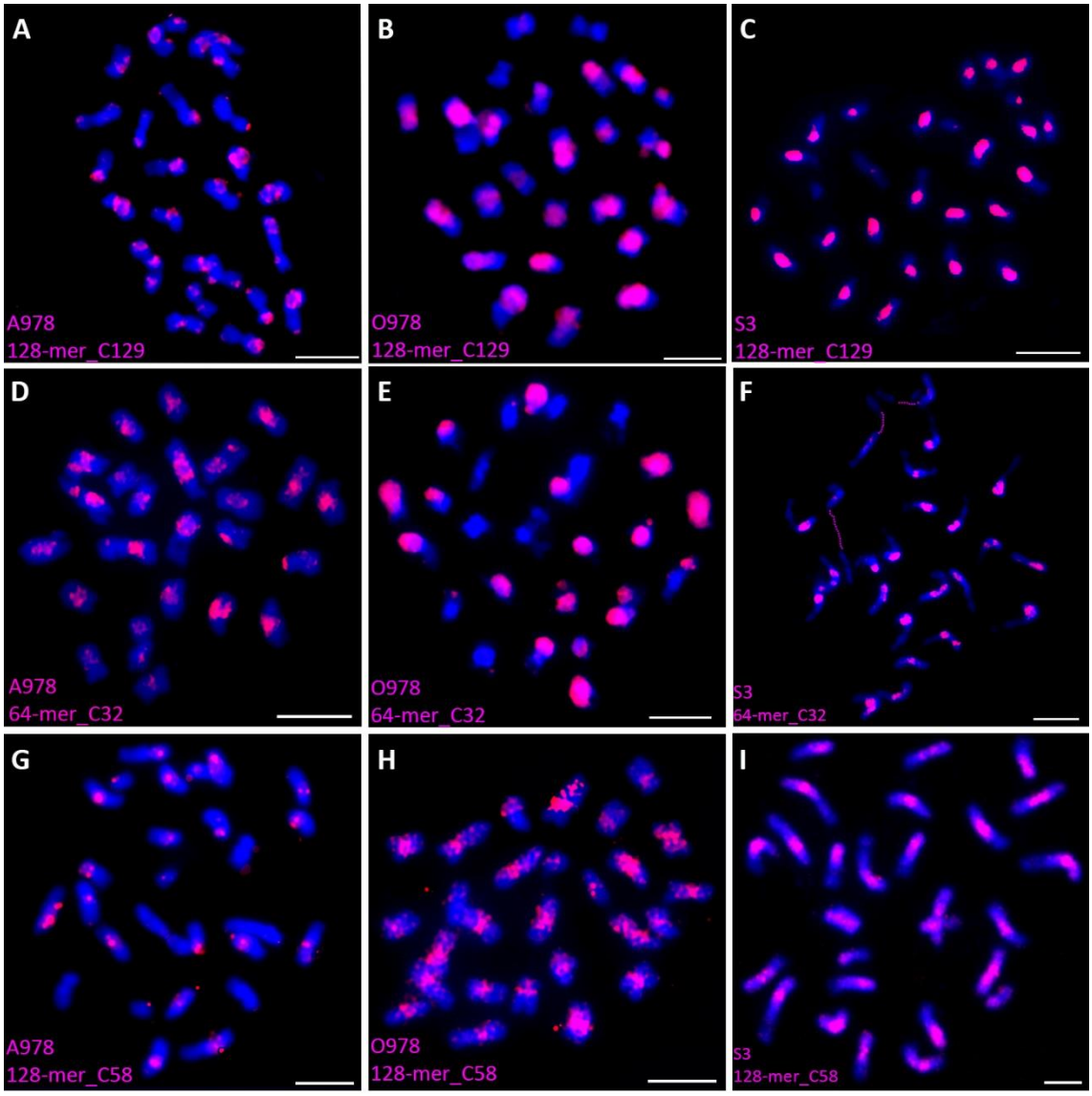
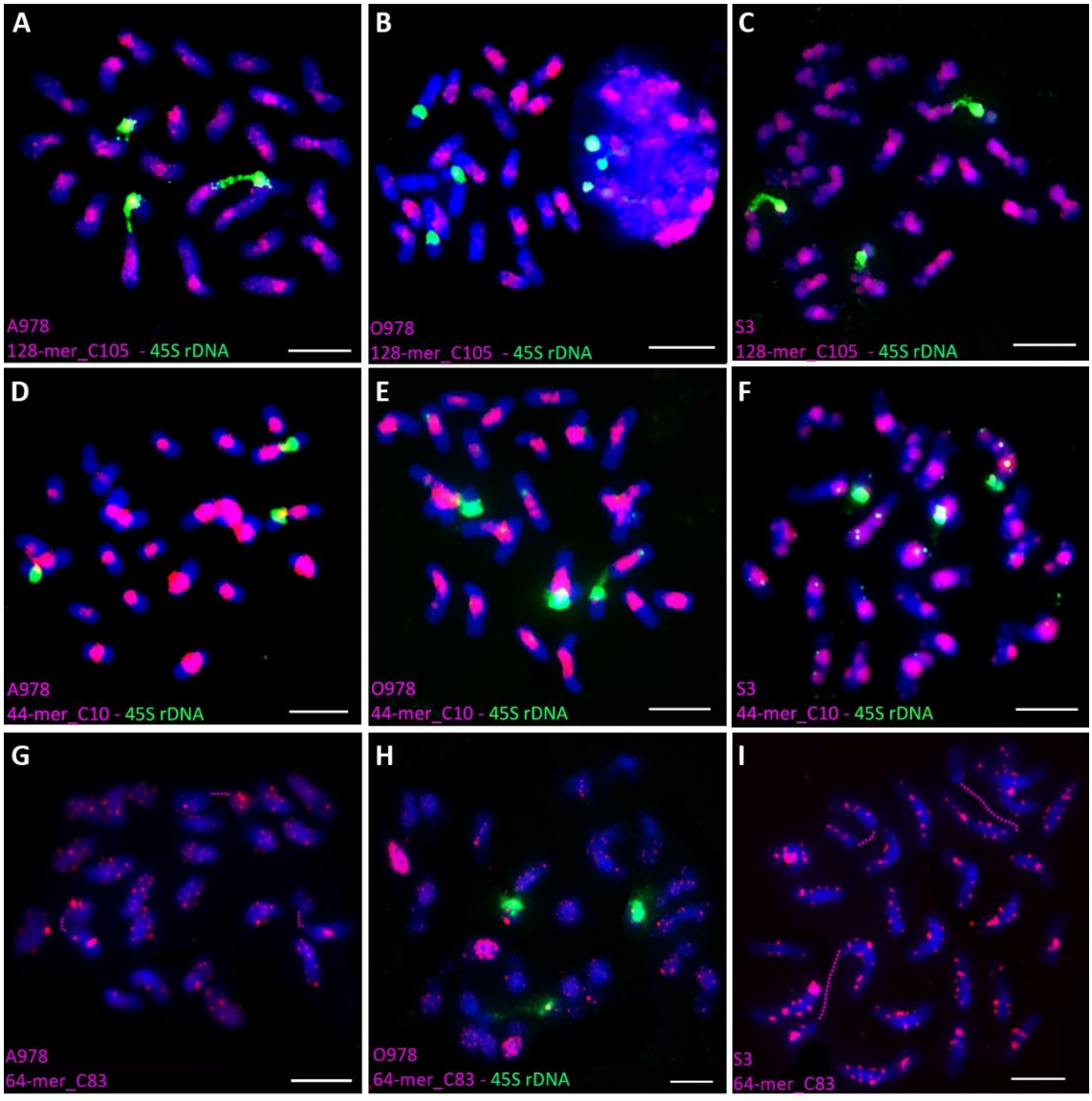
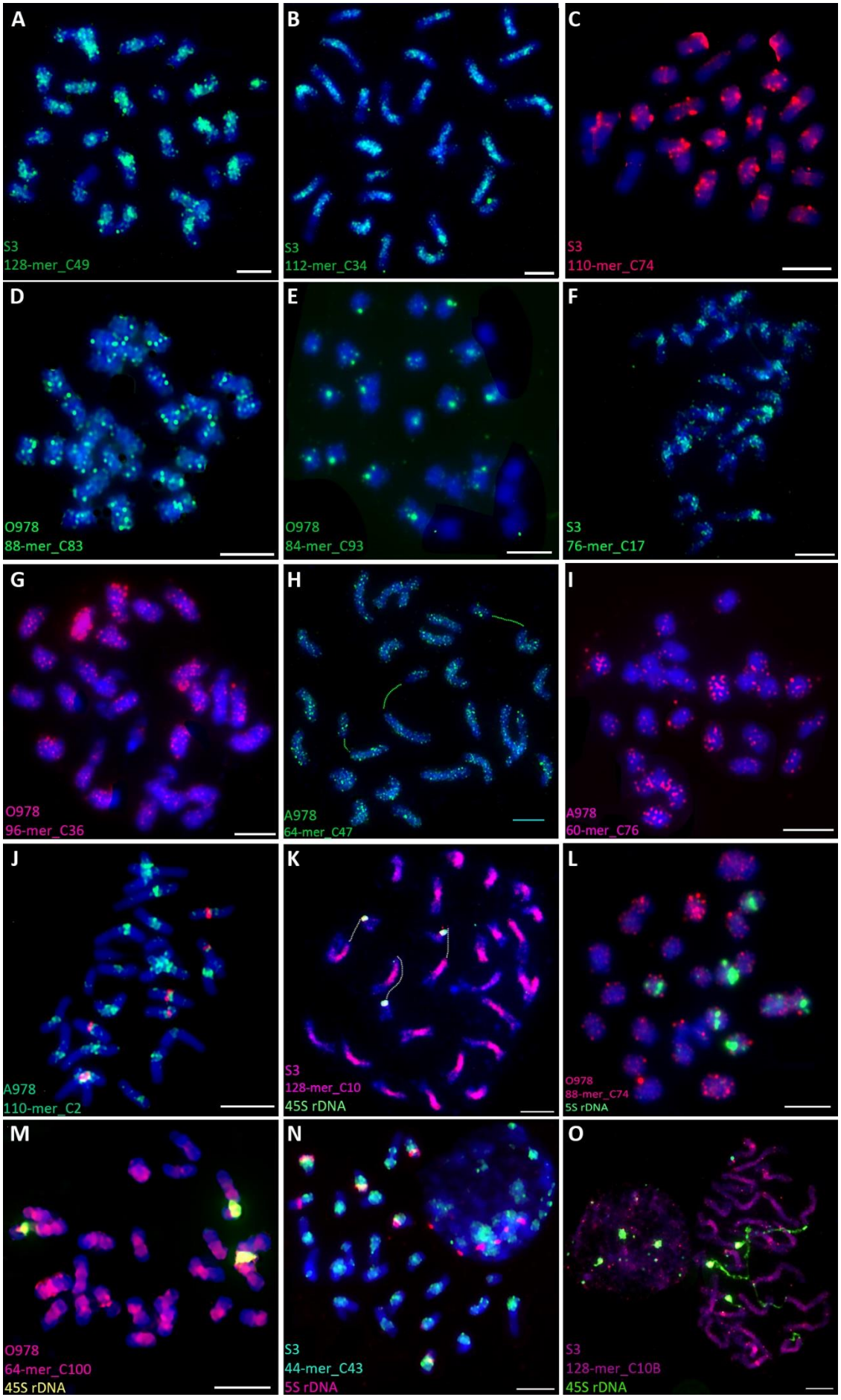


Figure 6.7. Probes used to show differences in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH pattern comes from the k-mer written beside the picture, and the *Taraxacum* micro species of the metaphase.



For many probes, the genome distribution patterns were highly similar between the three microspecies. Figure (6.8 A and F) represents probes with multiple strong pairs of dots, with some telomeric, some intercalary, superimposed on broad centromeric locations or hybridization to whole chromosome arms. Figure (6.8 B, D and H) shows probes with widespread labelling all over the chromosomes. Figure (6.8 C) from 110-mer_C74 shows many double dots of arms (rather than the broader band or diffuse signal of many other probes) on centromeric regions of all chromosomes. Figure (6.8 E) from 84-mer_C93 shows different patterns of terminal double dots near the end of chromosomes and some intercalary sites. Figure (6.8 G, I and L) shows a dot pattern over most chromosomes with some chromosomes having stronger signals and some other chromosomes with no signals or very weak signals. Figure (6.8 J) shows a pattern of rather stronger sites at centromeres and some diffuse signal along arms. Figure (6.8 K, M) shows hybridization in broadly centromeric regions and with much dispersed signals to include some arms and exclusion from some other arms. Figure (6.8 N) shows the most frequently observed k-mer distribution pattern with broad centromeric hybridization, some chromosomes stronger than others, with some arms unlabelled or with stronger signals at centromeres. Figure (6.8 O) shows a probe with a more dispersed signal, with some centromere bands but no excluded arms. Probes for 45S and 5S rDNA (either reference probes of those derived from k-mers) were test to label the satellite and 5S chromosomes (Figure 6.8 J – O).

Figure 6.8. Probes used to show picture of probes that have different probe distribution and locations in the chromosomes in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH pattern comes from the k-mer written beside the picture, and the *Taraxacum* micro species of the metaphase. ➔


A
S3
128-mer_C49
B
S3
112-mer_C34
C
S3
110-mer_C74
D
O978
88-mer_C83
E
O978
84-mer_C93
F
S3
76-mer_C17
G
O978
96-mer_C36
H
A978
64-mer_C47
I
A978
60-mer_C76
J
A978
110-mer_C2
K
S3
128-mer_C10
45S rDNA
L
O978
88-mer_C74
5S rDNA
M
O978
64-mer_C100
45S rDNA
N
S3
44-mer_C43
5S rDNA
O
S3
128-mer_C10B
45S rDNA

Figures (6.9 and 6.10) shows k-mer patterns where there is high similarity in distribution of each probe (locations on the chromosomes) on A978, O978 and S3, despite the probes differing in k-mer length and sequence (also confirmed by sequence dot-plots). The pattern can be described as showing widespread distribution over of all chromosomes with some concentration at or around most centromeres (Figure 6.9 C, F, D, L), lack of signals on a few chromosomes (Figure 6.9 A, B, D, E, J), and the presence of some gaps at centromeres (Figure 6.9 G, H, I, K) and largely excluded from 45S rDNA (Figure 6.10 A-J, not K): some chromosomes have stronger signals concentrated on centromere, and hybridization signals may be largely absent from one arm or diffuse on most of an arm. While this group of probes mostly have similar patterns of distribution on the chromosomes, as shown in Figure 6.10 with double-hybridization with two k-mer probes, there are small differences in location, confirming hybridization to different chromosomal sequences. Figure 6.10K, L (green detection) shows a slightly different pattern, hybridizing in the broad centromeric pattern (red and green, J, K, L) described above, but in addition labelling the three 45S rDNA sites (seen in green only; K, L).

6.4.5. Highly abundant and unique *in situ* patterns and chromosome classification karyotype

Fragments of both 45S rDNA and 5S rDNA were included among abundant k-mers, and could then be assembled into contigs. From the k-mer contigs, a 5S rDNA sequences was extracted from very highly abundant 128-mer (128-mer_SHC31274): this was repeated more than 31000 times in the S3 genome. After primer design, PCR amplified two different fragment sizes. The expected size was 89 bp, while the gel picture (Figure 6.11) showed two bands, the lower band as expected at <100 bp (named here as “128mer5S.1”) and a higher one at c. 600 bp (named here as “128mer5S.2”). After labelling both recovered bands from the gel, *in situ* hybridization showed two different signal distributions with these two probes from the same PCR. The probes showed clearly differences between the three *Taraxacum* microspecies. The 128mer5S.1 probes showed the expected 5S rDNA signals with six sites on two triplets (3x) of chromosomes, one triplet with strong signals and the other triplet with slightly weaker signals.

Figure 6.9. Probes used to show more default FISH pattern from different primer amplification and different k-mer analysis in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH pattern come from the k-mer written beside the picture, and the *Taraxacum* micro species of the metaphase.

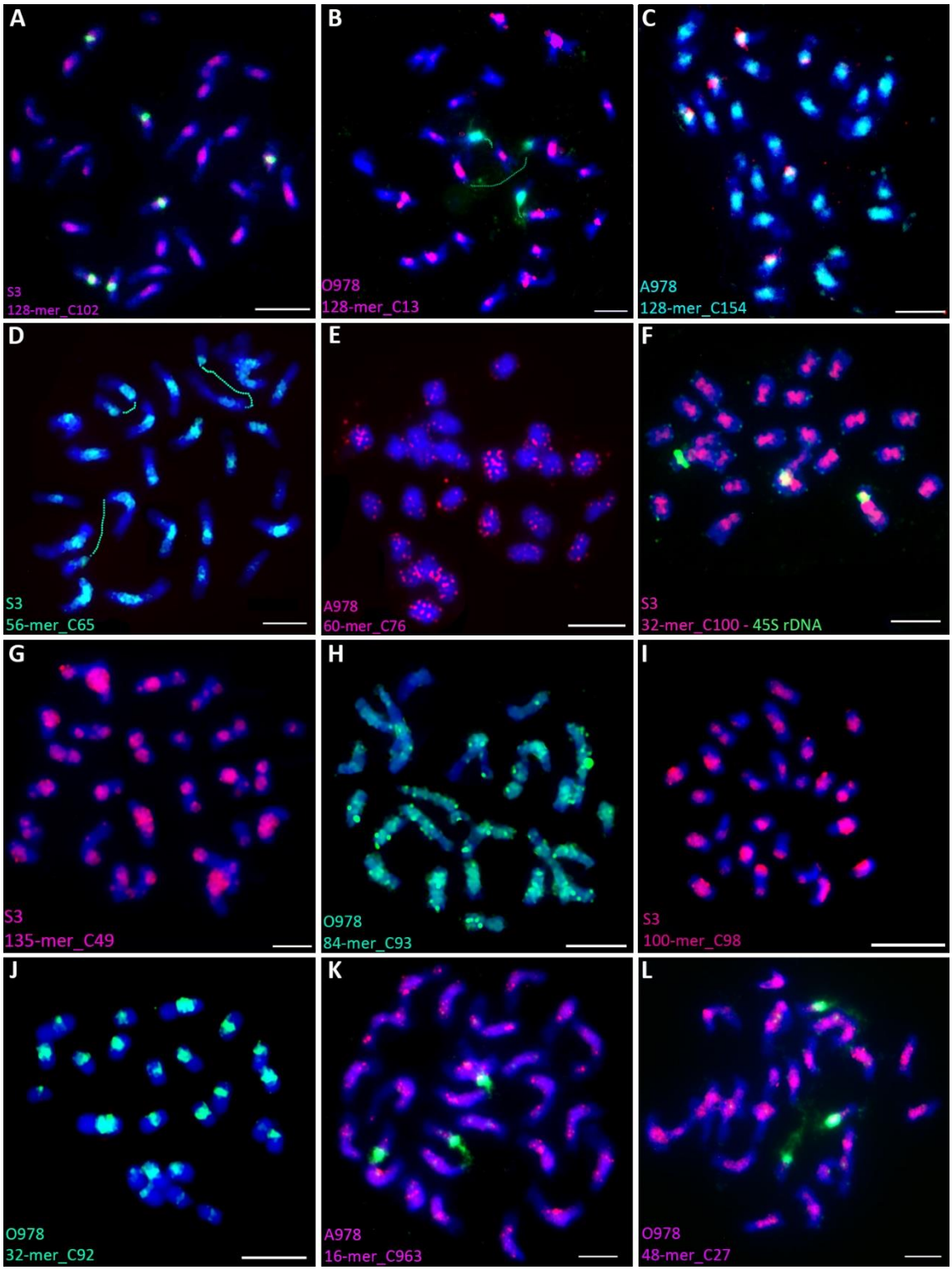
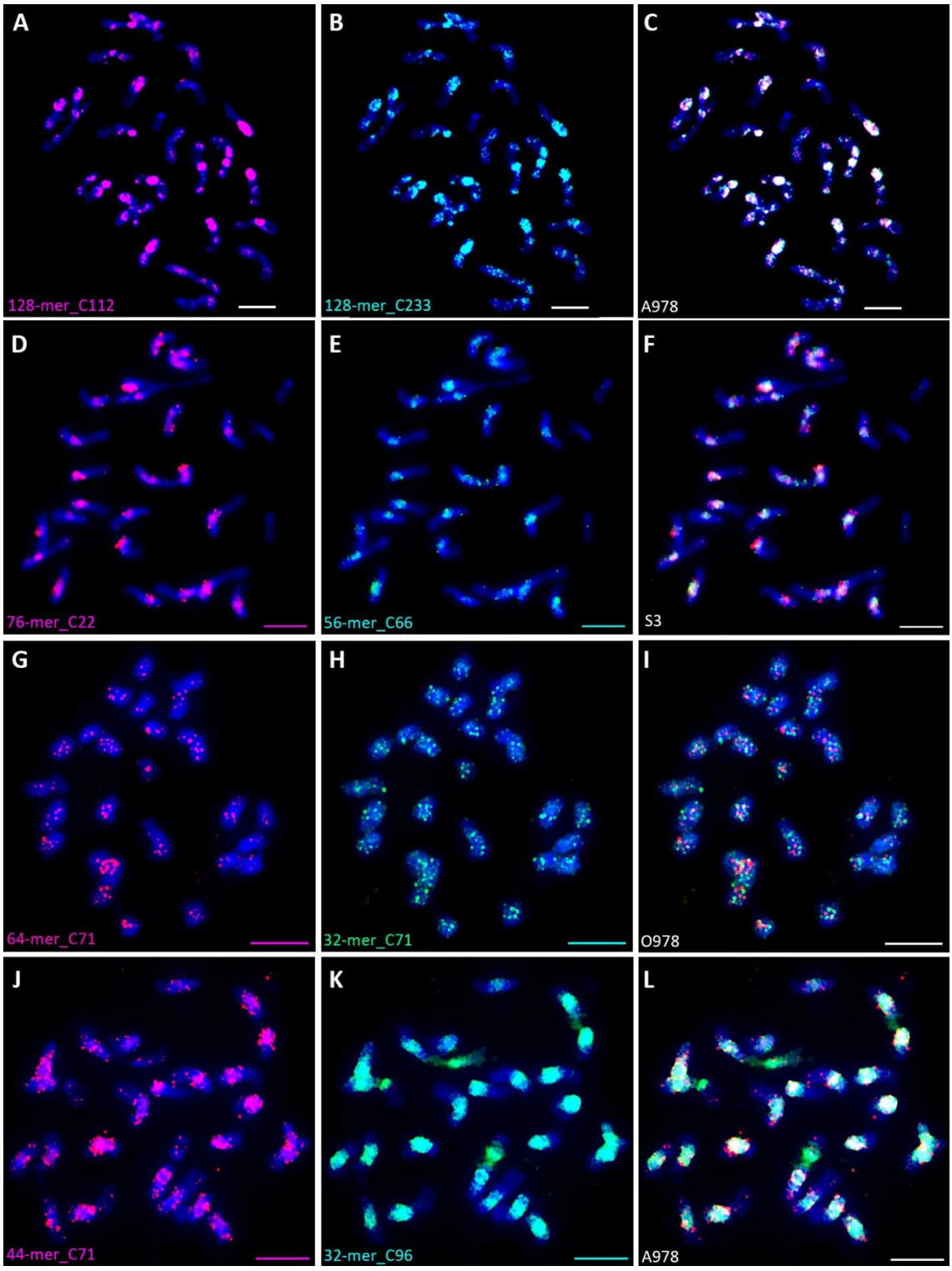


Figure 6.10. Probes used to show default FISH pattern from different primer amplification and different k-mer analysis, the patterns show similarity between the two probes on the same metaphase in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH pattern comes from the k-mer ritten beside the picture, and the *Taraxacum* micro species of the metaphase.



A similar pattern and strength was seen with wheat 5S rDNA (pTa794), widely used as a reference probe in many species. The larger 128mer5S.2 probes showed a different distribution of signals which included most of the 6 sites of 5S rDNA but with additional sites. In O978 and S3, there were signals on 10 sites, with the 6 sites from 5S rDNA all collocating with the 128mer5S.2 product. The location of the remaining four signals is different between the two accessions: in S3 the signals are located in sub-telomeric regions (Figure 6.12 A, B, C) while in O978 the signals are located in sub-telomeric and centromeric regions (Figure 6.12 G, H, I). However, in A978, there are only six overlapping 5S sites and no trace of the 128mer5S.2 signal at localized sites on other chromosomes (Figure 6.12 D, E, F). Thus, the probe 128mer5S.2 shows clear differences between the three *Taraxacum* agamospecies summarized as present only at 5S sites in A978, and distinctive sites present only on some chromosomes at distinct sub-telomeric location in S3, and sub-telomeric and centromeric sites in O978.

Figure 6.11. Gel picture of amplification of probe 128-mer_SHC31274, showing the amplified band of 128mer5S.1 (>600) and 128mer5S.2 (>100). Q-step 2 ladder (YorkBio), were used on both side of the gel.

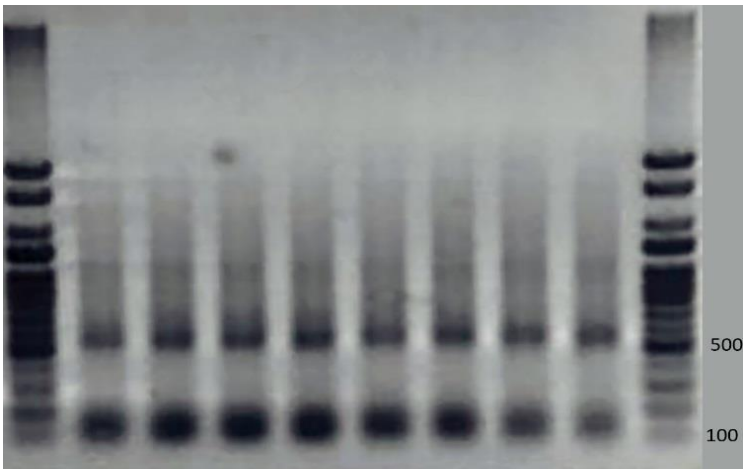


Figure 6.12. Probes used to show distribution of 5S rDNA (128mer5S.1-green) and (128mer5S.2-red) from 128-mer_high copy sequence and its differences between the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes.

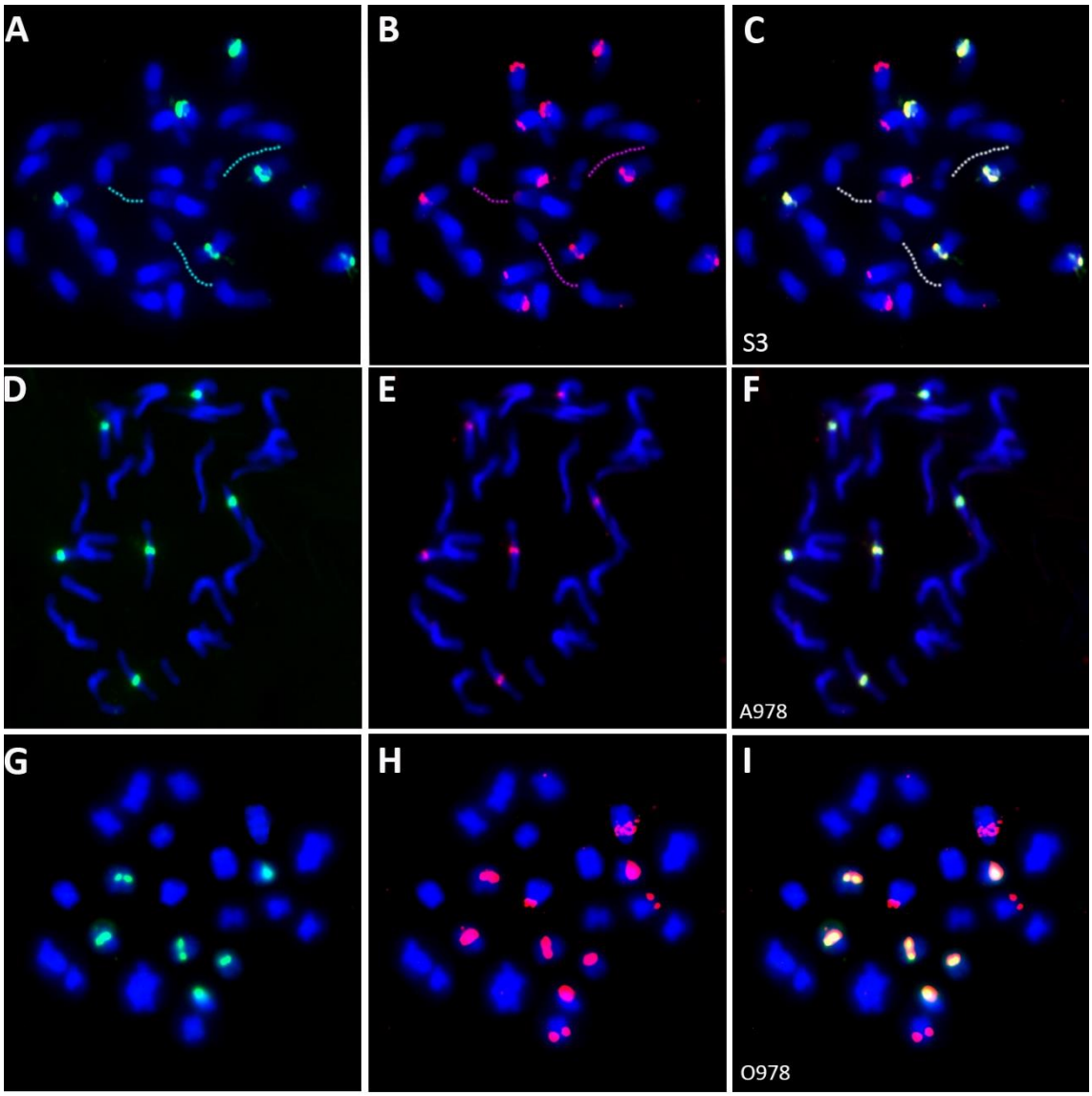
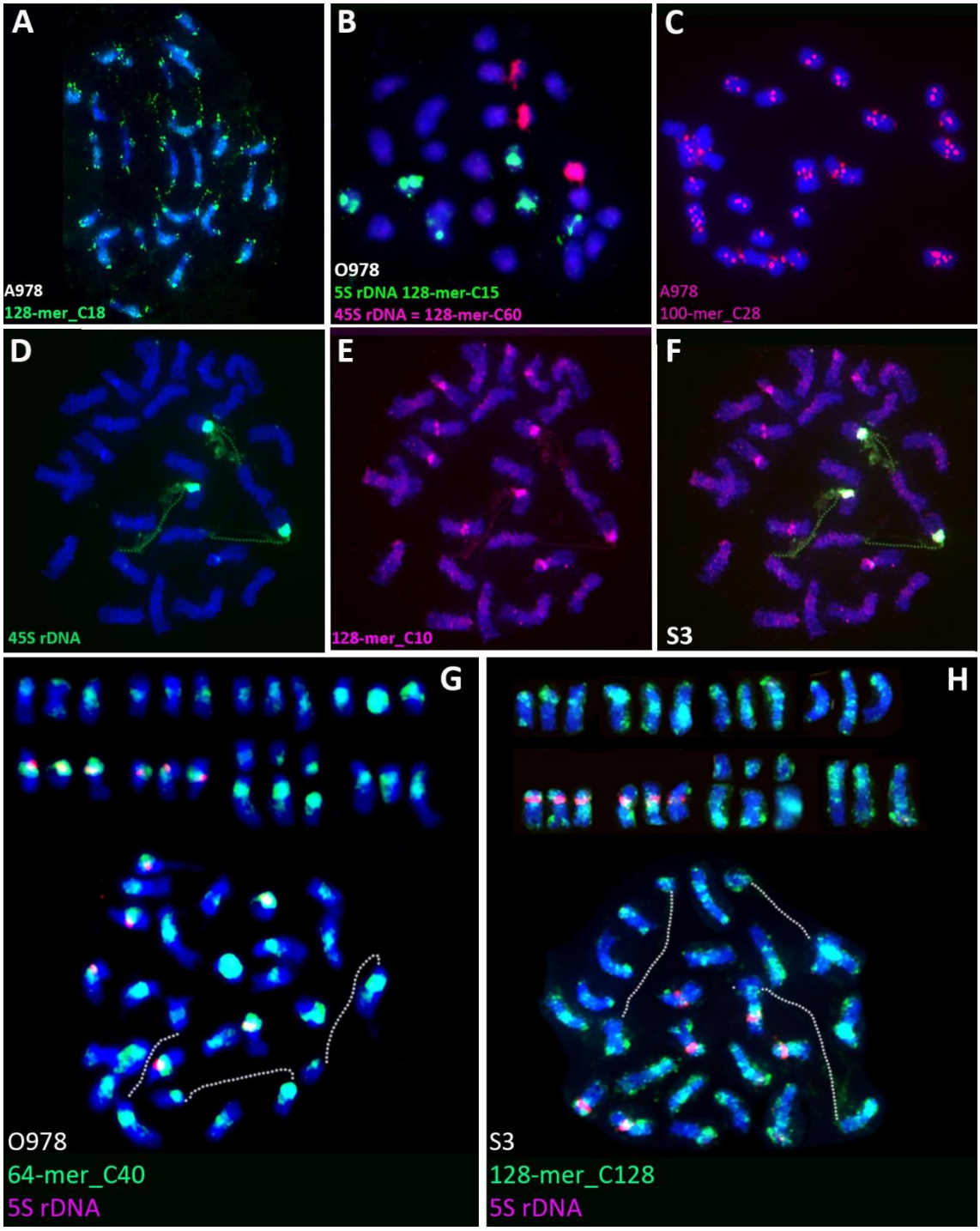


Figure (6.13B) shows the location of the 45S rDNA probe on one triplet of chromosomes. The primer was designed to the k-mer contigs and again confirms the ability of the analysis to reveal this expected tandem repeat with a larger repeat motif (c. 9.5kb). In prometaphase and early metaphases, the 45S showed the long distance separation of the satellite from the rest of its chromosome.

Another of the abundant sequence motifs expected in the analysis was the telomeric sequence (CCCTAAA) 19 canonical sequences CCCTAAA and TTTAGGG and seven canonical sequences in the 128-mer (Figure 6.4a). Figure (6.13A) shows very strong telomeric signals which appeared as double dots at telomere which sometimes appeared stretched out, giving an extended dotted fibre coming out from the end of about half the chromosomes. A few chromosomes have the telomere sequence in the centromeric regions with a slightly dispersed organization, an occasional feature of telomeric sequences in other species (eg. in Arabidopsis thaliana).

Other probes in Figure 6.13 show characteristic chromosomal locations. The probe from 100-mer_C28 (Figure 6.13 C) is localised as distinct sites at the centromere of all chromosomes; BLASTn reported a sequence homology to a *Lactuca* microsatellite sequence. 128-mer_C10 labelled nine sites, three of them related to the three satellite (NOR) chromosomes with the 45S rDNA sites, and the other six sites related to the 5S rDNA sites. It is unusual to see all 45S and 5S rDNA sites labelled just by one probe (Figure 6.13 D, E, F). Two probes give signals that are helpful chromosome karyotyping. 64-mer_C40 (green Figure 6.13 G) shows one triplet with no green signal and strong 5S rDNA, one triplet with medium-to-weak green signals with medium 5S rDNA, two morphologically distinct triplets with very strong 64mer_C40 probe, and other locations of green signal well-defined at centromere or whole-arms and variable in strength. 128-mer_C128 (Figure 6.13 H; green signals) showed many sub-telomeric sites, some along whole arms and some bands of chromosomes.

Figure 6.13. Probes used to show some k-mer sequences with unique site on the chromosomes which given chance to make a karyotype from these probes in the three *Taraxacum* microspecies. Mitotic metaphase spreads of *Taraxacum* microspecies (2n = 24) after FISH with probes for various k-mer length. The chromosomes counterstained with DAPI (blue). Bar = 10 µm. Drown dot lines in some figure indicate satellite separated region connected to another part of the chromosomes. The FISH patterns come from the k-mer written beside the picture, and the *Taraxacum* micro species of the metaphase.



6.5. Discussion

6.5.1 k-mer analysis

The k-mer analysis revealed many individual k-mers with very high copy number within the raw sequence reads, across all 37 values of k from 10 to 150 that have been tested here. As an overview, the results from the analysis supported those using different techniques to analyse the genome. To set out in the introduction above, the aim was to build a complete picture of the repetitive DNA content of the *Taraxacum* genome. Therefore, a broad survey of the abundance and organization of a wide range of the k-mer-derived sequences was carried out. More detailed analysis of individual classes identified here will be needed to understand fully the evolution of the individual sequences, the copy number and chromosomal distribution.

6.5.2 *Taraxacum* genome size

There have been several experimental methods that were used for analysing or estimating genome size such as Feulgen densitometry and flow cytometry, giving genome size as C-value (Bennett and Smith 1976) which can be calibrated and recalculated as base pairs. k-mer frequency analysis by using unassembled genomic sequence data can be used to estimate genome size of the sequenced organism (Li *et al.* 2010; Potato Genome Sequencing Consortium 2011; Huang *et al.* 2009). We estimated genome size of *Taraxacum* (Cx) by k-mer analysis and compared this value with other published *Taraxacum* genome sizes. Bennett *et al.* (1982) reported *Taraxacum officinale* Weber as 5.3 pg (Southern hemisphere) or 5.1 pg (Northern hemisphere) for the 2n=3x=24 cytotype by Feulgen densitometry (microdensitometry), thus the 1Cx genome sizes is 831.3 Mbp (831 Mbp per unreplicated haploid genome). This is similar to the 851 Mbp and 880 Mbp 1Cx sizes reported for *Taraxacum* 2n=3x=24 section Ruderalia by Zavesky *et al.* (2005). However, Vidic *et al.* (2009) reported a 2n=16 *Taraxacum officinale* as having 2.5 Gbp 2C DNA content (1Cx=1250 Mbp; remeasured but not karyotyped by Temsch *et al.* 2010 as 1.254 pg 1Cx = 1226 Mbp). However, if they had measured a triploid and not a diploid, the DNA content would have averaged 825 Mbp 1Cx (Table 6.2). The 1Cx genome size is estimated at

658 Mbp by k-mer, similar at about 81% of the size estimated by microdensitometry or flow cytometry; given the presence of organellar sequences and the high proportion of repeats, the k-mer approach is expected to be less accurate. The lower k-mer size has been seen in other analyses (e.g. Kim *et al.* 2014).

In previous studies, 17-mer or larger sizes were used to estimate genome size by determining the peak of single copy DNA in the sequence data. Kim *et al.* (2011) reported that for estimating genome size by using k-mer analysis, a short k-mer could underestimate genome size (<20) and choosing low k-mer depth could overestimate genome size (<20). However, in current study with the genomic sequencing data of 21x, 10x and 12x haploid genome fold we could not show a clear peak of single copy DNA sequences for k-mer larger than 13-mer, possibly due to low coverage of sequencing, or heterozygosity (between the three genomes). A k-mer analysis to estimate coverage of the genome and hence genome size gave no clear peak for A978 and O978 with the more limited 12 Gb of sequence (12x coverage; Figure 6.3 b; Kim *et al.* 2014).

6.5.3 Comparison of k-mers for different values of k between 10 and 150

Many sequence assembly algorithms exploit k-mers for the making contigs from short sequence reads (eg. SoapDenovo, Abyss; although not Newbler, typically used for 454 sequencing with limited recent updates; nor Geneious assembler). At the moment, the value for k is determined empirically to maximize the contig lengths and number of reads incorporated into contigs for each combination of genome and sequencing technology. k is typically 15 to 50 (although not always reported in publications). Various values of k are also tested for genome size estimates to maximise the resolution and distinctness of peaks from single copy DNA (Figure 6.1 here). However, there are no reports of analysis of multiple lengths of k-mers to evaluate repetitive DNA sequences in genomes; furthermore, computer hardware (memory) and algorithms used up to 2012 were not appropriate for use outside bioinformatics research labs where there has been minimal interest in repetitive DNA. Therefore, a wide range of values of k was evaluated within the work reported here. In general

terms, most k-mer lengths revealed the expected high-copy sequences in the raw reads: chloroplast and mitochondrial sequences, 5S and 45S rDNA and telomeres.

6.5.4 Unexplained sequencing technology differences

Three *Taraxacum* microspecies genomes were used to compare repetitive DNA results from two sequencing technologies, between independent samples, and to examine evolutionary differences between the microspecies. As far as we know, no previous work has compared different recent sequencing technologies nor very closely related accessions using the k-mer analysis. The microspecies were obtained from Majeský *et al.* (2012) who genotyped them and placed them in separate taxonomic groups by morphology. The microspecies *T. amplum* (A978) is placed in one morphological series, the AMP group. O978 and S3 are placed in a second series OSP, with three morphologically divergent taxa O, S and P, which are also robustly distinguished by AFLPs based on nuclear DNA polymorphisms (Majeský *et al.* 2012).

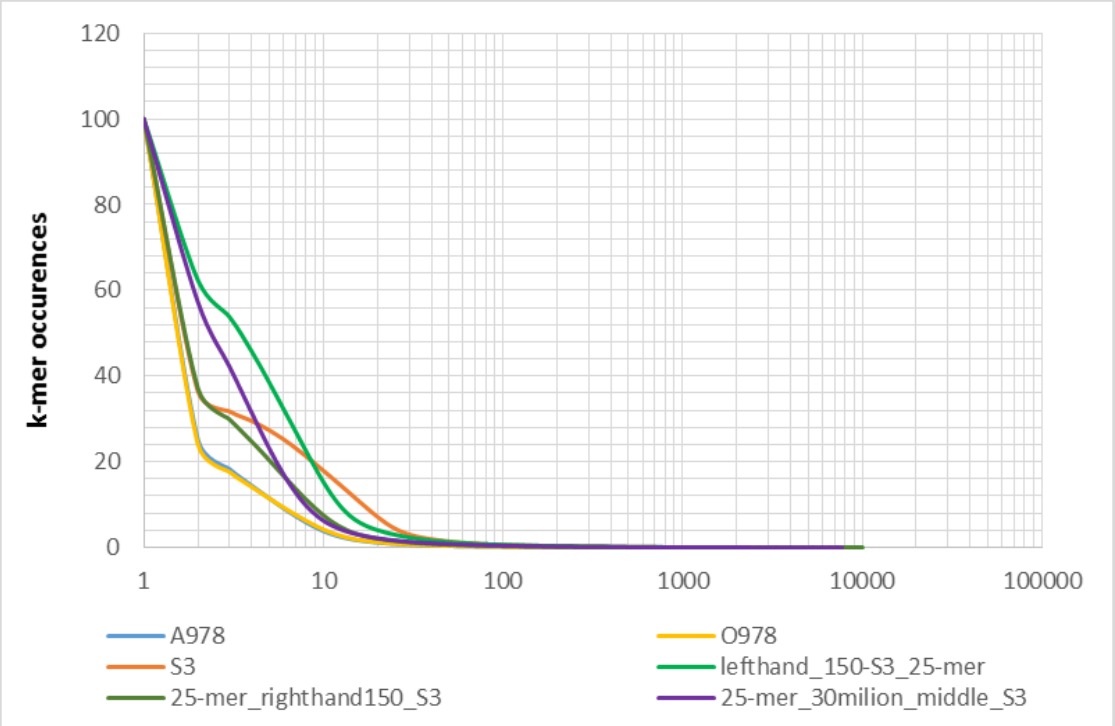
Unexpectedly, the S3 reads of 300bp gave a substantially different pattern of occurrence frequency for highly abundant k-mers. The consequence of this was that the three genomes could not be compared directly without ending up analysing the features of sequencing technology, but at the same time it meant that significant results are not being reported which are a consequence of the technology. It is well-known that there are many sequence artefacts as a consequence of chemistry, optics and analysis algorithms in sequencing machines, and all aspects are continuously being modified. Here, differences between S3 genome and A978, O978 genome was surprising: the sequencing library construction method is the same for Miseq and Hiseq, and the kit that was used to prepare the library in both platforms is the “Nextera kit”, although carried out 6 months apart. A new chemistry from Illumina, "TruSeq Nano" promises "high genome coverage and reduced bias", suggesting that the previous technologies have "low genome coverage and bias", although nothing like this is stated (http://www.illumina.com/techniques/sequencing/ngs-library-prep.html).

The genetic (Majeský *et al.* 2012), IRAP data (Chapter 3), nuclear and chloroplast (Chapter 4) data show that S3 and O978 are genetically the most similar accessions. These results may, or may not, reflect differences in repetitive DNA which

can evolve at different rates and with different mechanisms to low-copy sequences analysed as markers, although the expectation would be that there is some correlation.

This analysis shows the k-mer distribution is somewhat different reflecting different proportions of repetivity as part of the whole genome. The distribution of all k-mers here was compared with species we have been working with. Further analysis took place to look for the reasons for such differences. We expected that the differences may relate to the differences in the sequencing read lengths, which for the A978 and O978 is 151 bp read length and S3 is 301 bp read length. We reanalysed the S3 genome by cutting the sequencing length to the same as A978 and O978 genome, 151 bp lengths. However, k-mers with first, middle and last 150bp fragments of the original 301bp read from S3 gave a similar pattern to the 301bp analysis, so it is not an artefact of analysing from 300bp reads against 150 bp (Figure 6.14). It seems unlikely that any artefact arising from 'ends' of sequence reads affect anything for k< 50, but S3 shows a different graph shape. Without the two 151bp Hiseq reads, it would have been easy to make comparisons with evolutionary implications.

Figure 6.14. Analysing of different length positions of 301 bp of S3 genomic sequence reads and compared it with A978 and O978 genome and S3 301bp 25-mer occurrences.



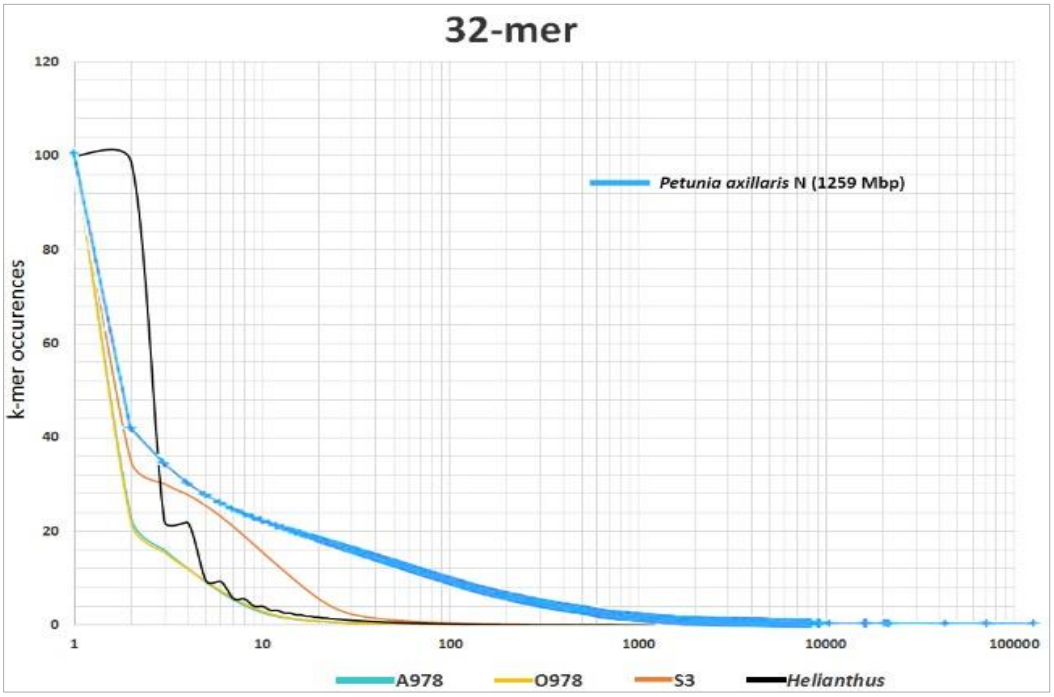
6.5.5 Genome repetitivity compared across taxa

Genome repetitivity was assessed via 27-33 different k-mer frequency (Figure 6.1) in the three *Taraxacum* microspecies as in the tomato genome (Tomato Genome Consortium 2012. Supplementary Figure 42). Graphs were overlaid onto the tomato, potato and sorghum patterns, in addition to *Petunia axillaris* N (Bombarely *et al.* 2016, Supplementary Figure 10a and b). As reported in (Bombarely *et al.* 2016) that *P. axillaris* genome size is 1259 Mbp and it is larger than the tomato, potato and sorghum in addition to our result in *Taraxacum* so there will be the expectation of larger repetitivity. It is clear from the Figure (6.15a) that the *Petunia* curve laid over the three *Taraxacum* microspecies, for instance all 16-mer occurring ≥10 times account for 18% for A978, 19% for O978 and 32% for S3 when it represent 24% of the tomato, 22% of the potato genome, 41% of the Sorghum genome, and 50% of Petunia genome. These results indicating that the *Taraxacum* microspecies are closer to the tomato and potato genomes by repetitive content and genome size, while the Petunia and Sorghum have twice the frequency of repetitivity (Figure 6.15b). In addition, we tried to compare our results with the results of analysing of 1.7 Gb sequence data from *Helianthus annus* from Staton *et al.* (2012), but because of the limited amount of sequences, the data are not robust enough to allow meaningful comparison (Figure 6.15). We can suggest that the 12Gb in 150bp reads (c. 15x genome coverage) is suitable to analyse repeat content, 21 Gb in 300bp reads (as for S3) is in excess, while 4Gb is most likely to be too limited.

Another analysis took place to find where are these differences located in the k-mers results in S3 genome compared with the A978 and O978 genomes. We looked at the 32-mer, and we started with removing the top 5, 10, 20, 30, 40, and 50 of most frequent 32-mers. By removing these reads, the graph of S3 become closer to A978 and O978 genome, and at the end when removing top 50 reads the graph of the three genomes overlapped (Figure 6.16). Thus, I can say that the differences between the three microspecies are within the top 50 most abundant k-mers at 32-mers. The GC content of the two sequencing technologies was closely similar, and k-mer analysis of randomized sequences showed identical graphs (data not shown) so GC/AT preferences were not involved.

Figure 6.15. Assessment of genome repetitivity *via* 16-mer and 32-mer frequencies. The cumulated coverage of the genome in base pairs is plotted against increasing 16-mer and 32-mer frequencies. **(a).** 32-mer **(b).** 16-mer data overlaid onto graph from *Petunia axillaris* N (Bombarely *et al.* 2016, Supplementary Figure 10) and tomato (Tomato Genome Consortium 2012, Supplementary Figure 42).

a



b

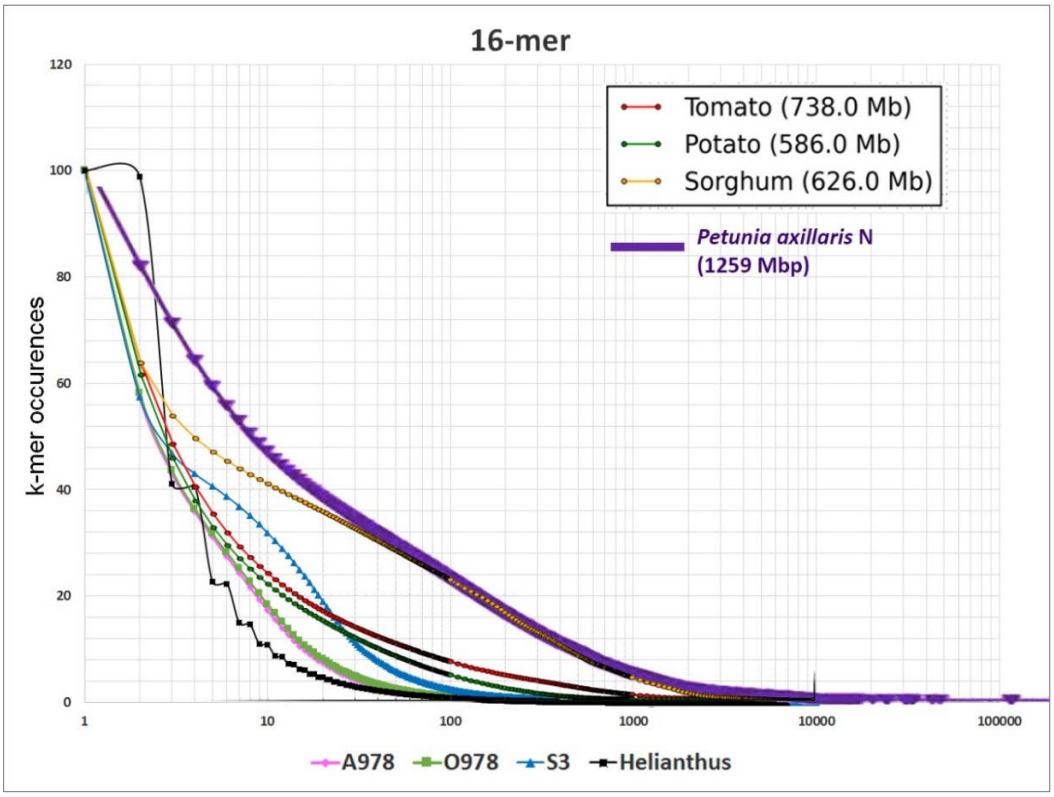
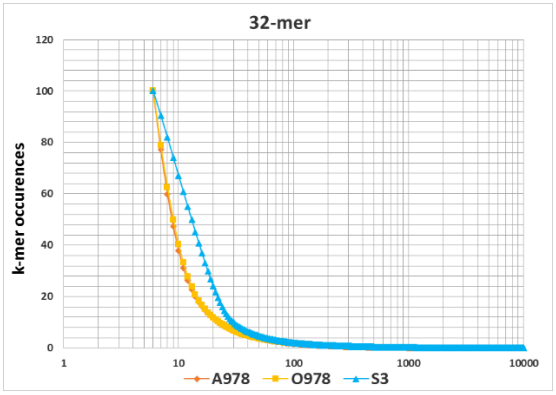
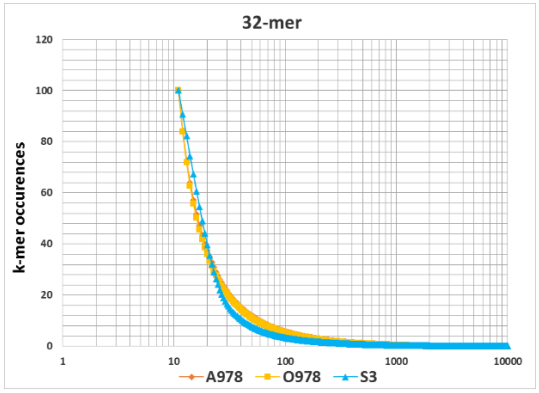


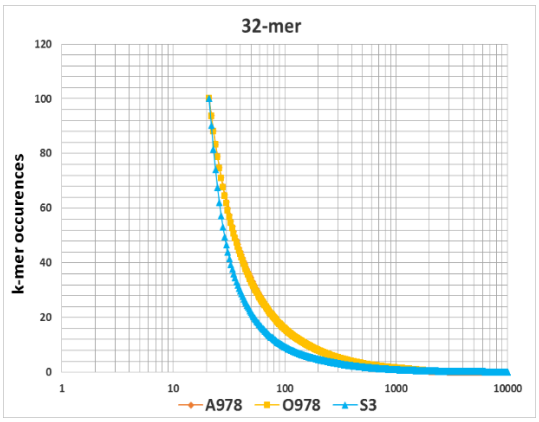
Figure 6.16. Plotting 32-mer frequency after removing top 5-50 reads.



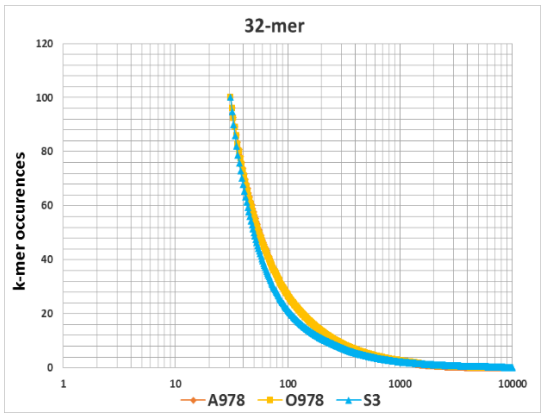
Removing top 5



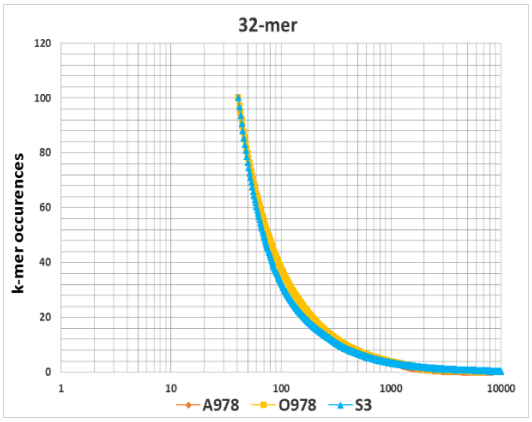
Removing top 10



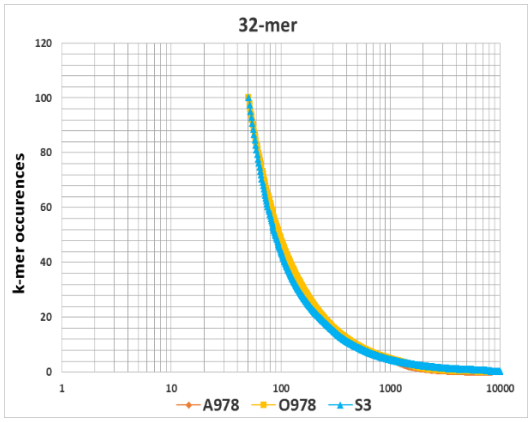
Removing top 20



Removing top 30



Removing top 40



Removing top 50

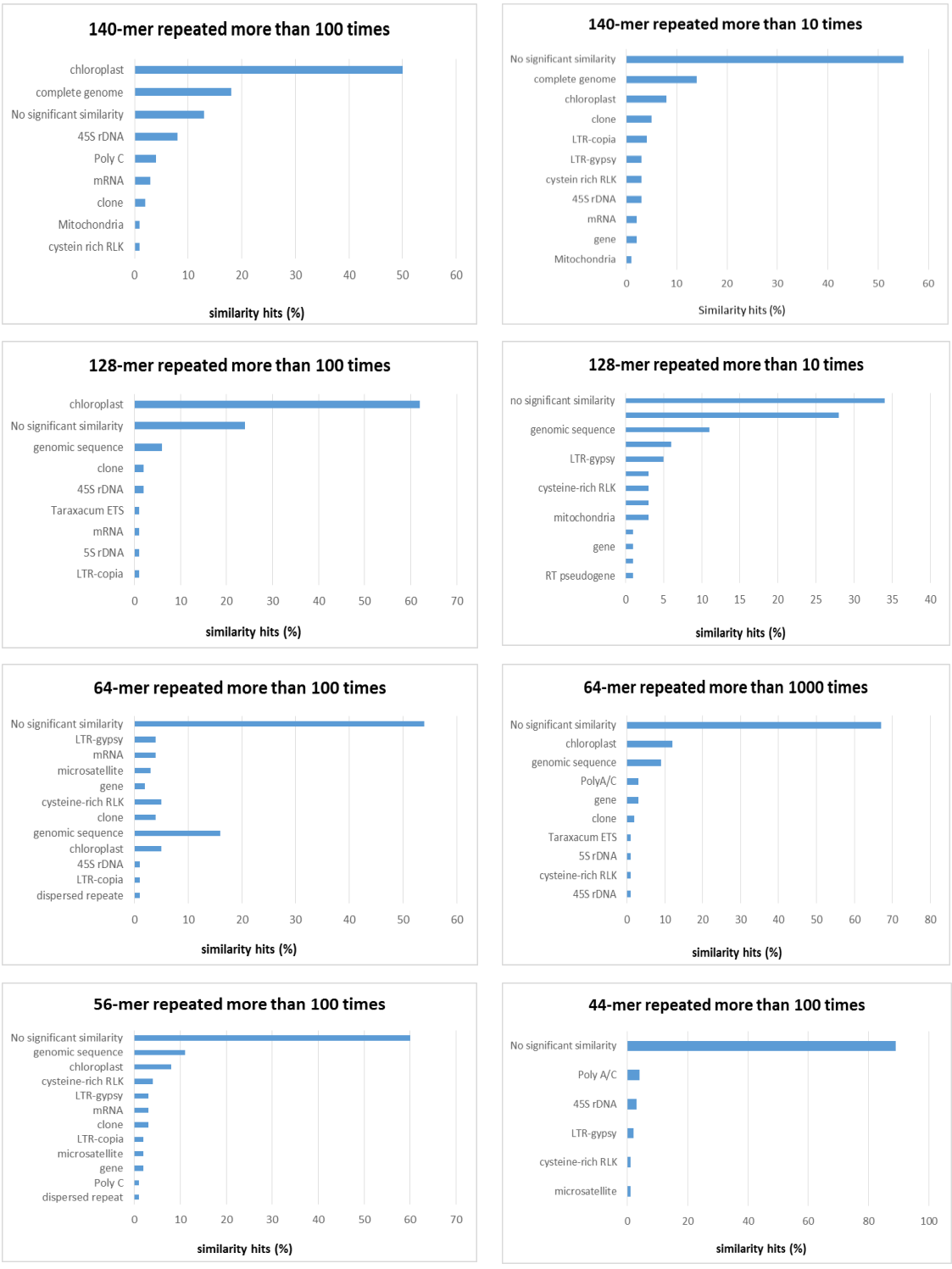
6.5.6 Types of repeat in the *Taraxacum* genome

BLAST searches against the DNA nucleotide and protein motif databases using the top k-mer derived contigs revealed a range of known repeated sequence classes, demonstrating that the analysis is isolating motifs which are expected in any plant genome. Chloroplast sequences were abundant, as expected, and was discarded from the analysis of nuclear repetitive DNA (see Chapter 4 for chloroplast genomes). Mitochondrial sequences were also identified and not further analysed.

45S and 5S rDNA, as would be expected from the known abundance of these sequences, were quickly and robustly identified as larger and smaller satellite repeats by the k-mer assemblies with almost all values of k, before they were shown to be rDNA by sequence comparison. In our bioinformatics analysis, we could not identify any other sequences with satellite patterns except one of the probe which was related to *Lactuca* and has homologies with one of the Rauscher and Simko (2013) microsatellite-based markers used to develop lettuce SSRs. Some *in situ* pictures (see below) showed strong centromeric signals of probes, but the edges of the signal were rarely as well-defined as seen for rDNA or satellites in other species. For smaller satellites, some telomere sequences were identified as multimers of TTTAGGG, but among the high abundance motifs, other microsatellites only featured once (in the 44mer>100 assembly). Searches of the reads suggested many microsatellite and compound microsatellite arrays were present, but were below the length and abundance thresholds of the repeat analyses. Thus we conclude that the overall abundance of satellite DNA in *Taraxacum* is low compared to many other species where major satellites are present at the centromeres or telomeres of many chromosomes.

Both copia and gypsy LTR retroelement motifs were identified from most k-mer lengths analysed. A major family of repetitive genes in many plants are the receptor like kinases (RLKs) (Tör *et al*. 2009), regulating many processes in plants. One kinase, the cysteine-rich RLK (RECEPTOR-like protein kinase) featured in many searches and it is identified in all five different values of k in Figure 6.17 and is known to having roles in the pathogen defence regulation and programmed cell death (Acharya *et al.* 2007; Idänheimo *et al.* 2014).

Figure 6.17. Represent the percentage of top 100 contigs high similarity hits (BLASTn) in NCBI GenBank for different k-mer length.



A high proportion of contigs for all values of k and repeat thresholds was found to have no significant similarity to any database sequence, or were reported as matching otherwise unannotated mRNA / complete genome / genome sequence / gene / "clone" accessions (Figure 6.16). Therefore, a number of these representing different k-mers and thresholds were synthesized as labelled probes or isolated from genomic DNA by PCR and labelled before localization of chromosomes by *in situ* hybridization.

6.5.7 Chromosomal localization of abundant sequence motifs identified by k-mer analysis

FISH results showed all repeats had characteristic and distinct hybridization patterns on metaphases chromosomes. The results also demonstrated that there are small, but important and arguably evolutionarily significant, differences in abundant sequence localization between the microspecies (Figure 6.9, 10).

Ribosomal 45S (18S-5.8S-26S) and 5S RNA genes (rDNA) are tandemly repetitive DNA present in hundreds or thousands of copies distributed on one or more chromosome triplets in the triploid. The nucleolus organizer region (NOR) is the location of the 45S rRNA gene repeats on the chromosomes. The 45S rDNA were observed on the short arm of the nucleolar organizing chromosomes (NOR) where the satellite sometimes separates from chromosomes. The 45S rDNA loci showed that their chromosomal localization on three chromosomes on *Taraxacum* microspecies (24=3x=24) is useful tool for determining the ploidy level even if the chromosome preparations were not very good or an interphase spread was used. 5S rDNA was also used: it identified two triplets of chromosomes in all of the three *Taraxacum* microspecies, with different locations and signal strengths on the chromosomes (Figure 6.13 B). Interestingly, one of the 128mer5S.2 probes has 10 sharp bands on chromosomes, four not associated with 5S in O978 and S3 but different in A978 which does not shows these extra sites with 128mer5S.2 probes. Therefore, the genome can include something that looks like a tandem-repeat array a localized site. These are not very abundant, and could be associated with a retrotransposon inserted in the 5S rDNA. Similar results have been reported in Musa by Teo (2007-PhD thesis) where amplification of the wheat 5S rDNA sequence gave smaller band size (~ 400 bp)

compared with amplifications of 5S rDNA sequences from *M. acuminata* (~ 600 bp). It was shown that retrotransposon-related sequences isolated from *M. acuminata* led to expansion of the size of the PCR band of amplified 5S rDNA. These results agree with the results from (Kumar and Bennetzen 2000; Nouroz *et al.* 2015) suggesting that some species have more active and abundant elements than others. It is clear that several insertions of mobile elements take place and are active in the genome: these insertions could include insertion to the genes or near to the genes leading to altering functions of that gene (Kumar and Bennetzen 2000). In addition, rDNA probes can be used solely or in combination with other repetitive DNA probes to identify individual chromosomes; the rDNAs have been used as significant markers to identify homeologous chromosomes in *Brassica* (Heslop-Harrison 1993).

6.5.8 Chromosome-specific cytogenetic markers

Following *in situ* hybridization, in accordance with the triploid nature of the species, chromosomes showed patterns of probe distribution that could normally be grouped into triplets, suggesting the autotriploid nature of the hybrid. This contrasts to, for example, the triploid *Crocus sativus*, also 2n=3x=24, where hybridization patterns of repetitive DNA did not form clear triplets (Alsayied *et al.* 2015).

In situ hybridization is not a quantitative method, although it does show major variations in signal strength, so conclusions regarding differences between microspecies by *in situ* hybridization strength and counts of copy number in raw reads were in some, but mostly weak, agreement. It is most likely that 2- to 4-fold hybridization strength differences are not seen clearly by *in situ* hybridization because exposures for micrographs are adjusted to give clear signal; forms of fluorescent image cytometry are not effective. Many authors have tried quantification, but the lack of follow-ups to positive publications suggests quantification is not reproducible. In contrast, an accuracy of 2% or better is achieved by either flow cytometry or absorbance microdensitometry (Bennett and Smith 1976; Leitch and Leitch 2013; Dolezel 2010). Other factors may also be involved in the weak correlation of *in situ* hybridization strength and read count. In the absence of high homology between probe and chromosomal target, hybridization may occur to less similar sequences. The difference may reflect family size of sequences related to the probes since the copy

number counting requires a high homology between reads and sequence, compared to the use of a more heterogeneous PCR product between primers and a lower homology required the PCR product to show hybridization to the chromosomes. Nevertheless, both the bioinformatics and *in situ* methods suggest that there are differences in abundance of repetitive sequences between A978, O978 and S3 and there are differences in genomic distribution determined by *in situ* hybridization.

6.5.9 The nature of repetitive DNA in *Taraxacum*

Many of the probes identified from k-mer analysis of the genomic sequence reads that were used for *in situ* hybridization showed a somewhat similar pattern of chromosomal location or genomic distribution. They localized to many broad centromeric sites, some chromosomes had centromeric gaps, and some whole arms showed stronger, weaker or no signal (Figure 6.10). When two of these probes were hybridized together, they showed slightly different hybridization patterns. Therefore, the results were interpreted to be showing specific hybridization of the probes to homologous chromosome sequences, and were not related to differential denaturation of chromatin, nor to be GC-content-related, nor to high similarity between domains within probes. The sequences were not identified as having any similarity to characterized sequences in the GenBank nucleotide database. They are abundant repetitive DNA families that accumulate in potentially inactive or inert domains of the chromosomes where they can amplify and recombine but have no genetic effects. Few showed motifs related to any transposable element domains. Most previous reports of plant repetitive DNA recognize defined transposable element derivatives, along with satellite sequences, as being the major component of repetitive DNA. We therefore suggest that the most abundant repetitive sequences in *Taraxacum*, which do show changes in abundance and chromosome distribution during evolution of the microspecies, belong to a new class we recognize as Passively Amplified DNA Sequences, PADS. The behaviour and amplification of PADS contrasts with the active amplification of retroelements; the amplification through unequal crossing-over and recombination of satellite DNA; and the genomic sweep mechanisms that lead to homogenization of repeats in other species.

CHAPTER 7

GENERAL DISCUSSIONS

7.1 Genetic variation in agamospermous *Taraxacum*

Apomictic reproduction has been considered as an evolutionary dead end (Chapman *et al.* 2003; Hörandl and Hojsgaard 2012). Theoretically, the offspring produced in apomictic plants are presumed to be identical to their maternal parent. But there is genetic variation in apomictic plants, including morphological variation (sometimes aneuploidy) (Sørensen and Gudjónsson 1946), and variability in size and morphology of pollen grains (Chiguryaeva 1976). There have been a number of studies interesting at the genetic variation within populations (e.g. population of *Taraxacum*), using isozyme or allozyme markers (Menken *et al.* 1995) and DNA markers (King and Schaal 1990; Van der Hulst *et al.* 2000). The genotypic variation contrasts with the triploid species *Crocus sativus*, where minimal genetic variation has been detected (Alsayied *et al*. 2015).

The source of the variation seen in agamospecies includes mutation, in the absence of meiosis: genetic variation must be somatic mutations in apomictic plants (such as producing eggs from somatic cells), or variation may come from rare hybridization since hermaphrodite apomict plants can act as pollen donors in cross fertilization sympatric regions, autosegregation in the areas that sexual plants are absent, and subsexual reproduction is a potential source of variation in *Taraxacum* (Darlington and Mather 1950; Introduction). Also, repetitive DNAs are considered as the source of genetic variation, because of their mobility and differences in abundance and their nature (McClintock 1984; Heslop-Harrison *et al.* 1997; Hilbrict *et al.* 2008). There have been many studies investigated genetic variation in *Taraxacum* especially

in the populations with obligate apomixis including study of Mogie and Richards 1983; Lyman and Ellstrand 1984; Ford and Richards 1985; Mogie 1985; van Oostrum *et al.* 1985; Hughes and Richards 1988; Majeský *et al.* 2012, which mostly consider that the observed variation in *Taraxacum* microspecies come from mutations. So far, the majority of the studies have been performed to investigate transposable elements diversity and evolution has been achieved in sexual reproduced eukaryotic organisms. A few studies on animals (Arkhipova and Meselson 2000; Zeyl *et al.* 1996; Goddard *et al.* 2001) suggested that asexual taxa may contain less transposable elements diversity than sexual taxa.

Active transposable elements could cause deleterious mutations; accumulation of deleterious mutation cause in asexual organisms leads to the extinction of that organism (Ozias-Akins and Van Dijk 2007), so, theoretically asexual lineages may become extinct due to transposon accumulation (Dolgin and Charlesworth 2006). Docking *et al.* (2006) reported that in *Taraxacum* substitution rates at non-synonymous sites is much lower than at synonymous sites. Zeyl *et al.* (1996) suggested that spreading and ability of increase in frequency of genomic active transposable element in asexual populations is much lower compared with their sexual relatives. Because asexuality is a derived state in higher eukaryotes, there is the possibility that transposons transferred from sexual ancestors to asexual species because it reduces the need for initial spread (Introduction). It was demonstrated that in apomictic *Taraxacum* there is a possibility of generating heritable variation by means of increased transposon activity by somatic recombination (Richards 1989; King and Schaal 1990). Retrotransposons may differ in expression level in sexual and apomictic organisms, such as the two highly expressed retrotransposons, N17 and N22, in sexual *Paspalum notatum* compared with its apomictic relative (Ochogavia *et al.* 2011). Transposable elements from maize have been experimentally introduced to the *Hieracium* genome. The functionality of the maize transposon system in *Hieracium* offers the potential for its use in a mutagenesis screen to "turn off" apomictic genes and to search for completely sexual progeny (Bicknell 1994a).

Genus *Taraxacum* Wigg. is complicate genus in the aspect of taxonomic and phylogenetic study, this could be because of a low level of morphological structural differentiation, predominant asexual reproduction, co-concurrent of sexual and asexual plant in the same population, polyploidization, and complex and ancient hybridization (Kirschner *et al*. 2003).

7.1.1 Variation in chloroplast and nuclear genome

The pattern of genetic variation in agamospermous groups leads to understanding of the evolutionary nature, variations, and origin of the apomictic complexes of this agamospermous group.

We investigate the taxonomic status of agamospecies and analysing the nature of their relationships (reticulate vs divergent) conducted by study of some chloroplast region and nuclear DNA to investigate the relationship between *Taraxacum* (*Hieracium*) members. The results from chapter 3 showed that results from nuclear (45S rDNA) and chloroplast (coding and noncoding region) showed variability but the supporting bootstrap values were very low between *Taraxacum* (*Hieracium*) agamospecies. The *Taraxacum* accessions were investigated by Majeský *et al.* (2012) by checking the ploidy level of the original plants used in this study. Majeský also checked their reproduction mode, by emasculation and screening of DNA content by Flow Cytometry in seeds.

Chapter 4 investigates PCR amplification and sequencing of multiple regions of chloroplasts and nuclear plastomes. This is time-consuming and requires optimization, while whole plastome sequencing only requires DNA extraction and a service provider (next generation sequencing). Plastome comparisons showed that among the three *Taraxacum* accessions two of them O978 and S3 are identical in contrast with A978 genome. Sequenced individuals represent agamospermous progeny (O978, S3) belong to the OSP group and A978 belongs to the AMP group. Despite their clear and robust nuclear differentiation, sequencing of the chloroplast *trn*L–*trn*F intergenic spacer showed they shared the cp*1a* haplotype. This suggests haplotype cp*1a* might be derived from the most recent common ancestor of many derived *Taraxacum* sections

(Chapter 4). Nuclear markers (AFLP) showed genetic boundaries between the O978 and S3 sub-groups, whole chloroplast sequences did not (Majeský *et al.* 2012), This may point to the young evolutionary age of the two microspecies (*T. obtusifrons* and *T. stridulum*): they have not accumulated any chloroplast mutations between each other and their most recent common ancestor. Morphologically, they are well-defined as separate morphological units with a low number of observed genotypes within investigated individuals from the O and S microspecies: two (*T. obtusifrons*) and four (*T. stridulum*) multilocus genotypes were detected by six nuclear SSRs (simple sequence repeats) among 21 and 23 genotyped individuals, while AFLPs showed only one AFLP-phenotype among 10 fingerprinted individuals of both microspecies (Majeský *et al.* 2012). Definitely, whole plastome sequences provide far more discrimination power than individual markers, for phylogeny reconstruction.

These results show the point of recent origin of these accessions (probably post-Pleistocene). The sequenced plastome (A978) may represent the most common recent chloroplast type involved in origin of many evolutionarily younger *Taraxacum* taxa. The results confirm also that genotyping variation could be generated in asexual species, albeit at a lower rate than sexual species. The processes that generate genetic variation may themselves be under selection; selection for optimization of mutation rates has been proposed for asexual organisms. Further, asexual *T. officinale* has high fecundity and viability, and large numbers of individuals colonize widespread habitats. Therefore, even a low rate of nonmeiotic recombination would result in the accumulation of genotypic variation. The process of clonal selection among the asexual genotypes may subsequently influence the distribution of clonal genotypes in different environments over time, and such processes may facilitate adaptive evolution (King and Schall 1990).

7.1.2 Variation in repetitive content

7.1.2.1 Nature and abundance of repetitive DNA sequences

We employed bioinformatics analysis and repeat clustering methodology of large scale sequence data for comparative analysis of three microspecies of *Taraxacum*

officinale agg. (Cichoridae, Asteraceae), using ~45 Gb genomic DNA sequence data to get deep insight into repeat composition, identify and quantify major groups of repetitive DNA family sequences including transposable elements. In Chapter 5, In accordance with results from other plant species have been studied so far, the present analysis showed that *Taraxacum* genome is mostly composed of LTR-retrotransposons (class I elements) 24-31%, and the most abundant repetitive DNA sequences found in the three microspecies of *Taraxacum* genome were LTR-retrotransposons. Out of them, Ty1-copia represented 13-16% of the genome while the Ty3-gypsy elements represented 11-15% of the genome. The number of the Ty1-copia lineages is more than the presented lineages by Ty3-gypsy. Data in others species have indicated prevalence of Ty3/gypsy retrotransposons in plant nuclear genomes (Macas *et al.* 2007; Bartoš *et al.* 2008; I.R.G.S. 2005; Velasco *et al.* 2007), contrasting with *Taraxacum*.

7.1.2.2 In situ hybridization as a tool for identifying repetitive DNA

In order to understand the repetitive DNA motifs distribution and organisations on the chromosomes the key method is to use the repetitive DNA sequences motifs as probe and hybridize them by Fluorescent *in situ* hybridisation (FISH) on chromosomes (Heslop-Harrison *et al.* 1997). FISH analysis of selected cluster sequences allowed further insight in to the genome organisation of repetitive DNA sequences of the *Taraxacum* microspecies.

In most *Taraxacum* sections a characteristic satellite chromosome is seen with a large euchromatic region distal additional constriction in some chromosomes which indicates the position of nuclear organiser regions (NOR) (Mogie and Richards 1983; Heslop-Harrison and Schwarzacher 2011), as these chromosomes are characteristic by present primary and secondary constriction sites. The 45S rDNA located at the shorter arms. This chromosome most usually occurs at the frequency of one per haploid set of chromosomes. Consequently, they have been suggested that the NOR chromosomes correlate to “satellited chromosomes”, and that the filiform region corresponds to the nucleolar organiser region (NOR). The results from Chapter 5 showed an extra probe

designed from CL132 from O978 genome represented the chromosome specific signals and rDNA associated probe.

Nevertheless, fluorescence *in situ* hybridization (FISH) on mitotic chromosomes revealed that elements from distinct repetitive DNA probes gave different signal strength, locations and pattern of genomic distributions among the three *Taraxacum* microspecies. Thus, both the graph based clustering and *in situ* hybridization showed the microspecies, although morphologically closely related, differed.

Various values of k were tested for genome size estimates to maximise the resolution and distinctness of peaks from single copy DNA. As far as we know, no previous work has compared different recent sequencing technologies nor very closely related accessions using the k-mer analysis (Chapter 6). In general terms, most k-mer lengths revealed the expected high-copy sequences in the raw reads: chloroplast and mitochondrial sequences, 5S and 45S rDNA and telomeres. Unexpectedly, the S3 reads of 300bp gave a substantially different pattern of occurrence frequency for highly abundant k-mers. These differences between the three micro species are within the top 50 most abundant k-mers at 32-mers.

A high proportion of contigs for all values of k and repeat thresholds was found to have no significant similarity to any database sequence, or were reported as matching otherwise unannotated mRNA / complete genome / genome sequence / gene / "clone" accessions. Therefore, a number of these representing different k-mers and thresholds were synthesized as labelled probes or isolated from genomic DNA by PCR and labelled before localization of chromosomes by *in situ* hybridization.

The genetic (Majeský *et al.* 2012), IRAP data (Chapter 3) and chloroplast (Chapter 4) data show that S3 and O978 are genetically the most similar accessions. These results may, or may not, reflect differences in repetitive DNA which can evolve at different rates and with different mechanisms to low-copy sequences analysed as markers, although the expectation would be that there is some correlation.

7.1.3 Chromosomal localization of abundant sequence motifs

FISH results showed all repeats had characteristic and distinct hybridization patterns on metaphases chromosomes. The results also demonstrated that there are small, but important and arguably evolutionarily significant, differences in abundant sequence localization between the microspecies (Chapter 6).

Interestingly, one of the 128mer5S.2 probes has 10 sharp bands on chromosomes, four not associated with 5S in O978 and S3 but different in A978 which does not shows these extra sites with 128mer5S.2 probes. Therefore, the genome can include something that looks like a tandem-repeat array a localized site. These are not very abundant, and could be associated with a sequence inserted in the 5S rDNA that is subsequently amplified or homogenized.

Following *in situ* hybridization, in accordance with the triploid nature of the species, chromosomes showed patterns of probe distribution that could normally be grouped into triplets, suggesting the autotriploid nature of the hybrid. This contrasts to, for example, the triploid *Crocus sativus*, also 2n=3x=24, where hybridization patterns of repetitive DNA did not form clear triplets (Alsayied *et al.* 2015).

In the RepeatExplorer outcomes we could identify the high abundant novel tandem repeated DNA in the three microspecies of *Taraxacum* genome studied here. One is represented by a cluster (CL80 in S3 genome output file) which consists of the 49 bp motifs length and repeated tandemly with a copy number of 32.4 in one contig sized 1775 bp. As shown in *in situ* results it represented a unique of double dots at centromere on just 14 chromosomes. We further assembled the sequences of this tandemly repeated DNA in *Taraxacum* to whole sequence of *Helianthus* (Staton *et al.* 2009) but no matches were found. Except this CL80 no major tandemly repeated satellite DNA motifs were identified in *Taraxacum* in the reads.

In situ hybridization is not a quantitative method, although it does show major variations in signal strength, so conclusions regarding differences between microspecies by in situ hybridization strength and counts of copy number in raw reads were in some, but mostly weak, agreement. It is most likely that 2- to 4-fold

hybridization strength differences are not seen clearly by *in situ* hybridization because exposures for micrographs are adjusted to give clear signal; forms of fluorescent image cytometry are not effective. Many authors have tried quantification, but the lack of follow-ups to positive publications suggests quantification is not reproducible. In contrast, an accuracy of 2% or better is achieved by either flow cytometry or absorbance microdensitometry (Leitch and Leitch 2013; Greilhuber *et al.* 2013). Other factors may also be involved in the weak correlation of in situ hybridization strength and read count. In the absence of high homology between probe and chromosomal target, hybridization may occur to less similar sequences. The differences may reflect family size of sequences related to the probes since the copy number counting requires a high homology between reads and sequences compared to the use of a more heterogeneous PCR product between primers and a lower homology required the PCR product to show hybridization to the chromosomes. Nevertheless, both bioinformatics and in situ methods suggest that there are differences in abundance of repetitive sequences between A978, O978 and S3 and there are differences in genomic distribution determined by *in situ* hybridization.

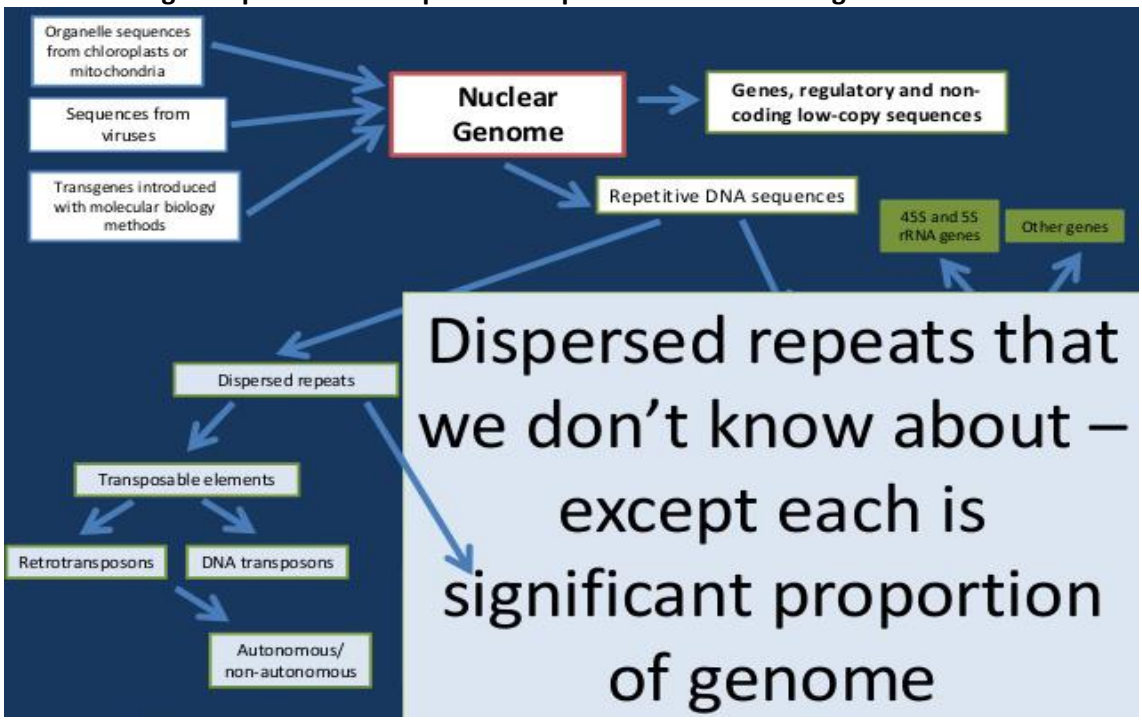
7.1.4. The nature of repetitive DNA in *Taraxacum*

Many of the probes identified from k-mer analysis of the genomic sequence reads that were used for *in situ* hybridization showed a somewhat similar pattern of chromosomal location or genomic distribution. They localized to many broad centromeric sites, some chromosomes had centromeric gaps, and some whole arms showed stronger, weaker or no signal. When two of these probes were hybridized together, they showed slightly different hybridization patterns. Therefore, the results were interpreted to be showing specific hybridization of the probes to homologous chromosome sequences, and were not related to differential denaturation of chromatin, nor to be GC-content-related, nor to high similarity between domains within probes. The sequences were not identified as having any similarity to characterized sequences in the GenBank nucleotide database. They are abundant repetitive DNA families that accumulate in potentially inactive or inert domains of the chromosomes where they can amplify and recombine but have no genetic effects. Few

showed motifs related to any transposable element domains. Most previous reports of plant repetitive DNA recognize defined transposable element derivatives, along with satellite sequences, as being the major component of repetitive DNA. We therefore suggest that the most abundant repetitive sequences in *Taraxacum*, which do show changes in abundance and chromosome distribution during evolution of the microspecies, belong to a new class we recognize as Passively Amplified DNA Sequences, PADS (Figure 7.1). The behaviour and amplification of PADS contrasts with the active amplification of retroelements; the amplification through unequal crossing-over and recombination of satellite DNA; and the genomic sweep mechanisms that lead to homogenization of repeats, seen in other species.

It can be concluded that, dispersed repetitive DNA considered as the major component in *Taraxacum* genome, and like other angiosperm plant these kind of repetitive DNAs could have role in the genomic evolution and broad range of diversity found in agamospecies. Repetitive DNA sequences along with cytological approaches showed highly differences among the closely related microspecies than chloroplast and nuclear genome, so that repetitive DNA marker can be use as comparative molecular marker to distinguish even closely related agamospecies.

Figure 7.1. Modification of figure after Heslop-Harrison and Schmidt 2012, to show *Taraxacum* agamospecies DNA sequence component of the nuclear genome.



CHAPTER 8

LITERATURE CITED

Acharya, B.R., Raina, S., Maqbool, S.B., Jagadeeswaran, G., Mosher, S.L., Appel, H.M., Schultz, J.C., Klessig, D.F. and Raina, R., 2007. Overexpression of CRK13, an Arabidopsis cysteine-rich receptor-like kinase, results in enhanced resistance to Pseudomonas syringae. Plant J. 50: 488-499.

Ahmed, I., Biggs, P.J., Matthews, P.J., Collins, L.J., Hendy, M.D. and Lockhart, P.J., 2012. Mutational dynamics of aroid chloroplast genomes. Genome Biol. Evol. 4: 1316-1323.

Akiyama, Y., Conner, J.A., Goel, S., Morishige, D.T., Mullet, J.E., Hanna, W.W. and Ozias-Akins, P., 2004. High-resolution physical mapping in *Pennisetum squamulatum* reveals extensive chromosomal heteromorphism of the genomic region associated with apomixis. Plant Physiol. 134: 1733-1741.

Alkan, C., Cardone, M.F., Catacchio, C.R., Antonacci, F., O'Brien, S.J., Ryder, O.A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E.E. and Ventura, M., 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21: 137-145.

Alsayied, N.F., Fernández, J.A., Schwarzacher, T. and Heslop-Harrison, J.S., 2015. Diversity and relationships of *Crocus sativus* and its relatives analysed by inter-retroelement amplified polymorphism (IRAP). Ann. Bot. 116: 359-368.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Arkhipova, I. and Meselson, M., 2000. Transposable elements in sexual and ancient asexual taxa. P. Natal. Acad. Sci. 97: 14473-14477.

Asker, S.E. and Jerling, L., 1992. *Apomixis in Plants*. *CRC Press*; 298.

Baarlen, P.V., Van Dijk, P.J., Hoekstra, R.F. and Jong, J.H.D., 2000. Meiotic recombination in sexual diploid and apomictic triploid dandelions (*Taraxacum officinale* L.). Genome; 43: 827-835.

Bailey, J.P. and Stace, C.A., 1992. Chromosome number, morphology, pairing, and DNA values of species and hybrids in the genus Fallopia (Polygonaceae). Plant Syst. Evol. 180: 29-52.

Bao, W., Jurka, M.G., Kapitonov, V.V. and Jurka, J., 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol. Biol. Evol. msp013.

Bartoš, J., Paux, E., Kofler, R., Havránková, M., Kopecký, D., Suchánková, P., Šafář, J., Šimková, H., Town, C.D., Lelley, T. and Feuillet, C., 2008. A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R. BMC Plant Biol. 8: 1.

Bashaw, E.C., 1980. Apomixis and its application in crop improvement. Hybridization of crop plants, (hybridizationof): 45-63.

Bell, G., 1982. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality.* CUP Archive.

Bennett, M.D. and Smith, J.B., 1976. Nuclear DNA amounts in angiosperms. Philo. Trans. R. Soci. B: Biol. Sci. 274: 227-274.

Bennett, M.D., Leitch, I.J., PRICE, H.J. and JOHNSTON, J.S., 2003. Comparisons with Caenorhabditis (~ 100 Mb) and Drosophila (~ 175 Mb) using flow cytometry show genome size in Arabidopsis to be~ 157 Mb and thus~ 25% larger than the Arabidopsis genome initiative estimate of~ 125 Mb. Ann. Bot. 91: 547-557.

Bennett, M.D., Smith, J.B. and Smith, R.L., 1982. DNA amounts of angiosperms from the Antarctic and South Georgia. Environ. Exper. Bot. 22 : 307-318.

Bennetzen, J.L., 1996. The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol. 4: 347-353.

Benson G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27: 573.

Bergman, C.M. and Quesneville, H., 2007. Discovering and detecting transposable elements in genome sequences. Bioinformatics; 8: 382-392.

Bhattacharyya, M.K., Smith, A.M., Ellis, T.N., Hedley, C. and Martin, C., 1990. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell; 60: 115-122.

Bhutkar, A., Russo, S.M., Smith, T.F. and Gelbart, W.M., 2007. Genome-scale analysis of positionally relocated genes. Genome Res. 17: 1880-1887.

Bicknell, R.A. and Koltunow, A.M., 2004. Understanding apomixis: recent advances and remaining conundrums. Plant Cell; 16: S228-245.

Bicknell, R.A., 1994. Hieracium: A model system for studying the molecular genetics of apomixis. Apomixis Newsletter; 7: 8-10.

Bicknell, R.A., Borst, N.K. and Koltunow, A.M., 2000. Monogenic inheritance of apomixis in two *Hieracium* species with distinct developmental mechanisms. Heredity; 84: 228-237.

Bicknell, R.A., Lambie, S.C. and Butler, R.C., 2003. Quantification of progeny classes in two facultatively apomictic accessions of *Hieracium*. Hereditas; 138: 11-20.

Birky, Jr., C.W., 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu. Rev. Genet. 35: 125-148.

Biscotti, M.A., Canapa, A., Forconi, M., Olmo, E. and Barucca, M., 2015. Transcription of tandemly repetitive DNA: functional roles. Chromosome Res. 23: 463-477.

Biscotti, M.A., Canapa, A., Olmo, E., Barucca, M., Teo, C.H., Schwarzacher, T., Dennerlein, S., Richter, R. and Heslop-Harrison, J.P., 2007. Repetitive DNA, molecular cytogenetics and genome organization in the King scallop (*Pecten maximus*). Gene; 406: 91-98.

Boeke, J. and Corces, V.G., 1989. Transcription and reverse transcription of retrotransposons. Annu. Rev. Microbiol. 43: 403-434.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C.S., Bliek, M., Boersma, M.R., Borghi, L., Bruggmann, R. and Bucher, M., 2016. Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. Nature Plants; 2: 16074.

Braaten, D.C., Thomas, J.R., Little, R.D., Dickson, K.R., Goldberg, I., Schlessinger, D., Ciccodicola, A. and D'Urso, M., 1988. Locations and contexts of sequences that hybridize to poly (dG-dT).(dC-dA) in mammalian ribosomal DNAs and two X-linked genes. Nucleic Acids Res. 16: 865-881.

Bringaud, F., Ghedin, E., Blandin, G., Bartholomeu, D.C., Caler, E., Levin, M.J., Baltz, T. and El-Sayed, N.M., 2006. Evolution of non-LTR retrotransposons in the trypanosomatid genomes: Leishmania major has lost the active elements. Mol. Biochem. Parasit. 145: 158-170.

Bruni I, Galimberti A, Caridi L, Scaccabarozzi D, De Mattia F, Casiraghi M, Labra M., 2015. A DNA barcoding approach to identify plant species in multiflower honey. Food chem. 170:308-15.

Bruni, I., De Mattia, F., Galimberti, A., Galasso, G., Banfi, E., Casiraghi, M., Labra, M., 2010. Identification of poisonous plants by DNA barcoding approach. Int. J. Legal. Med. 124: 595-603.

Cadle-Davidson, M.M. and Owens, C.L., 2008. Genomic amplification of the Gret1 retroelement in white-fruited accessions of wild Vitis and interspecific hybrids. Theor. Appl. Genet. 116: 1079-1094.

Calderini, O., Chang, S.B., de Jong, H., Busti, A., Paolocci, F., Arcioni, S., de Vries, S.C., Abma-Henkens, M.H., Lankhorst, R.M.K., Donnison, I.S. and Pupilli, F., 2006. Molecular cytogenetics and DNA sequence analysis of an apomixis-linked BAC in *Paspalum* simplex reveal a non pericentromere location and partial microcolinearity with rice. Theor. Appl. Genet. 112: 1179-1191.

Capy, P., Bazin, C., Higuet, D. & Langin, T. (eds), 1998. *Dynamics and evolution of transposable elements.* Library of Congress, Austin.

Carman, J.G., 1997. Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyembryony. Biol. J. Linn. Soc. 61: 51-94.

Carman, J.G., 2007. Do duplicate genes cause apomixis. Apomixis: evolution, mechanisms and perspectives. ARG Gantner Verlag KG; 169-194.

Carneiro, V.T., Dusi, D.M. and Ortiz, J.P.A., 2006. Apomixis: occurrence, applications and improvements. Floriculture, Ornamental and Plant Biotechnology: Advances and Topical; 564-571.

Castilho, A., Miller, T.E. and Heslop-Harrison, J.S., 1996. Physical mapping of translocation breakpoints in a set of wheat-*Aegilops umbellulata* recombinant lines using *in situ* hybridization. Theor. Appl. Genet. 93: 816-825.

Catanach, A.S., Erasmuson, S.K., Podivinsky, E., Jordan, B.R. and Bicknell, R., 2006. Deletion mapping of genetic regions associated with apomixis in *Hieracium*. P. Natl. Acad. Sci. 103: 18650-18655.

Chapman, H., Houliston, G.J., Robson, B. and Iline, I., 2003. A case of reversal: the evolution and maintenance of sexuals from parthenogenetic clones in *Hieracium* pilosella. Int. J. Plant Sci. 164: 719-728.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X. and Luo, K., 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PloS one; 5: e8613.

Chiguryaeva, A.A., 1976. *Palynology and apomixis*. *In Apomixis and Breeding*. Edited by S.S. Khokhlov. Amerind Publishing, *New Delhi*. 80–86.

Choi, K.S. and Park, S., 2015. The complete chloroplast genome sequence of *Aster spathulifolius* (Asteraceae); genomic features and relationship with Asteraceae. Gene; 572: 214-221.

Chrtek jun, J., Mráz, P., Zahradníčk, J., Mateo, G. and Szelag, Z., 2007. Chromosome numbers and DNA ploidy levels of selected species of *Hieracium* s. str. (Asteraceae). Folia. Geobot. 42: 411-430.

Clare, J. and Farabaugh, P., 1985. Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. P. Natl. Acad. Sci. 82: 2829-2833.

Clausen, J., Keck, D.D. and Hiesey, W.M., 1940. Experimental studies on the nature of species. I. Effect of varied environments on western North American plants. *Experimental studies on the nature of species. I. Effect of varied environments on western North American plants.*, (520).

Conant, G.C. and Wolfe, K.H., 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics, 24: 861-862.

Conger, A.D. and Fairchild, L.M., 1953. A quick-freeze method for making smear slides permanent. Stain technology, 28: 281-283.

Conner, J.A., Goel, S., Gunawan, G., Cordonnier-Pratt, M.M., Johnson, V.E., Liang, C., Wang, H., Pratt, L.H., Mullet, J.E., DeBarry, J. and Yang, L., 2008. Sequence analysis of bacterial artificial chromosome clones from the apospory-specific genomic region of *Pennisetum* and *Cenchrus*. Plant Physiol. 147: 1396-1411.

Contento, A., Heslop-Harrison, J.S. and Schwarzacher, T., 2005. Diversity of a major repetitive DNA sequence in diploid and polyploid *Triticeae*. Cytogenet. Genome Res.; 109:34-42.

Cooper, V.S., Schneider, D., Blot, M. and Lenski, R.E., 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. J. Bacteriol. 183: 2834-2841.

Cossu, R.M., Buti, M., Giordani, T., Natali, L. and Cavallini, A., 2012. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. Tree Genet. Genomes; 8: 61-75.

Curci, P.L., De Paola, D., Danzi, D., Vendramin, G.G. and Sonnante, G., 2015. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. PloS one, 10: e0120589.

D'Hont, A., Denoeud, F., Aury, J.M., Baurens, F.C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M. and Da Silva, C., 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature, 488: 213-217.

Daboussi, M.J. and Capy, P., 2003. Transposable elements in filamentous fungi. Ann. Rev. Microbiol. 57: 275-299.

Darlington, C. D. and LaCour, L. F. 1960. *The Handling of Chromosomes* (3rd ed.). Macmillan, New York.

Darlington, C.D. and Mather, K., 1950. *The elements of genetics*. *G. Allen & Unwin Ltd., London.*

de Carvalho, J.F., Oplaat, C., Pappas, N., Derks, M., de Ridder, D. and Verhoeven, K.J., 2016. Heritable gene expression differences between apomictic clone members in *Taraxacum* officinale: Insights into early stages of evolutionary divergence in asexual plants. BMC genomics; 17: 203.

Dempewolf, H., Kane, N.C., Ostevik, K.L., Geleta, M., Barker, M.S., Lai, Z., Stewart, M.L., Bekele, E., Engels, J.M., Cronk, Q.C. and Rieseberg, L.H., 2010. Establishing genomic tools and resources for *Guizotia abyssinica* (Lf) Cass.—the development of a library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome. Mol. Ecol. Res. 10: 1048-1058.

Den Nijs, A.P.M. and Van Dijk, G.E., 1993. *Apomixis. In Plant Breeding.* Springer Netherlands; 229-245.

Dickinson, T.A., 1998. Taxonomy of agamic complexes in plants: a role for metapopulation thinking. Folia Geobot. 33: 327-332.

Docking, T.R., Saade, F.E., Elliott, M.C. and Schoen, D.J., 2006. Retrotransposon sequence variation in four asexual plant species. J. Mol. Evol. 62: 375-387.

Doležel, J. and Greilhuber, J., 2010. Nuclear genome size: are we getting closer?. Cytometry A; 77: 635-642.

Dolezel, J., Bartos, J., Voglmayr, H. and Greilhuber, J., 2003. Nuclear DNA content and genome size of trout and human. ISAC; 51: 127.

Doležel, J., Doleželová, M. and Novák, F.J., 1994. Flow cytometric estimation of nuclear DNA amount in diploid bananas (*Musa acuminata* and *M. balbisiana*). Biol. Plantarum; 36: 351-357.

Dolgin, E.S. and Charlesworth, B., 2006. The fate of transposable elements in asexual populations. Genetics; 174: 817-827.

Dong, W., Liu, J., Yu, J., Wang, L. and Zhou, S., 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One; 7: e35071.

Doolittle, W.F. and Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature; 284: 601-3.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T. and Vrieling, K., 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res.: dsr002.

Dover, G.A.,2002. *Molecular drive.* Trends Genet. 18: 587-589.

Doyle, J. and Doyle, J.L., 1987. Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem Bull. 19: 11-15.

Draper, J., Mur, L.A., Jenkins, G., Ghosh-Biswas, G.C., Bablak, P., Hasterok, R. and Routledge, A.P., 2001. *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant physiol. 127: 1539-1555.

Dubcovsky, J. and Dvorák, J., 1995. Ribosomal RNA multigene loci: nomads of the *Triticeae* genomes. Genetics, 140: 1367-1377.

Dumas, C. and Mogensen, H.L., 1993. Gametes and Fertilization: Maize as a Model System for Experimental Embryogenesis in Flowering Plants. The Plant Cell; 5: 1337.

Duren, M.V., Morpurgo, R., Dolezel, J. and Afza, R., 1996. Induction and verification of autotetraploids in diploid banana (*Musa acuminata*) by in vitro techniques. Euphytica, 88(1), pp.25-34. D'Hont, A., Denoeud, F., Aury, J.M., Baurens, F.C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M. and Da Silva, C., 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature, 488: 213-217.

Edwards, R.A., Olsen, G.J. and Maloy, S.R., 2002. Comparative genomics of closely related salmonellae. Trends Microbiol. 10: 94-99.

Ellis, J.R., Bentley, K.E. and McCauley, D.E., 2008. Detection of rare paternal chloroplast inheritance in controlled crosses of the endangered sunflower *Helianthus verticillatus*. Heredity; 100: 574-580.

Ernst, A., 1918. Bastardierung als Ursache der Apogamie im Pflanzenreich.

Felsenstein J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 783-791.

Feschotte, C. and Pritham, E.J., 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41: 331.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Geneti. 5: 103-107.

Flavell, A.J., Dunbar, E., Anderson, R., Pearce, S.R., Hartley, R. and Kumar, A., 1992. Ty1–copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res. 20: 3639-3644.

Flavell, A.J., Pearce, S.R. and Kumar, A., 1994. Plant transposable elements and the genome. Curr. Opin. Genet. Dev. 4: 838-844.

Flavell, A.J., Smith, D.B. and Kumar, A., 1992. Extreme heterogeneity of Ty1-copia group retrotransposons in plants. MGG; 231: 233-242.

Ford, H. and Richards, A.J., 1985. Isozyme variation within and between *Taraxacum* agamospecies in a single locality. Heredity; 55: 289-291.

Ford, H., 1985. Life history strategies in two coexisting agamospecies of dandelion. Biolo. J. . Linn. Soc. 25: 169-186.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I., 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 32: W273-W279.

Friesen, N., Brandes, A. and Heslop-Harrison, J.S.P., 2001. Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers. Mol. Biol. Evol. 18: 1176-1188.

Fuchs, J., Brandes, A. and Schubert, I., 1995. Telomere sequence localization and karyotype evolution in higher plants. Plant Syst. Evol. 196: 227-241.

Gadella, T.W.J., 1991. Variation, hybridization and reproductive biology of *Hieracium pilosella* L. P. K. Ned. Akad. Wetensc. 94: 455-488.

Gerlach, W.L. and Bedbrook, J.R., 1979. Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res. 7: 1869-1885.

Gerlach, W.L. and Dyer, T.A., 1980. Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. Nucleic Acids Res. 8: 4851-4865.

Glunčić, M., Paar, V., Basar, I., Vlahović, I., Rosandić, M., Dekanić, K., Citković, M., Jelovina, D., Paar, P., Kelić, A. and Batista, J., 2013, January. Direct mapping of symbolic DNA sequence into frequency domain and identification of higher order repeats. Bioinfo. Biol. Physics: P. Scient. Meeting.

Goddard, M.R., Greig, D. and Burt, A., 2001. Outcrossed sex allows a selfish gene to invade yeast populations. P. Roy. Soc. Lond. B. Bio. 268: 2537-2542.

González, L.G. and Deyholos, M.K., 2012. Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC genomics; 13: 644.

Gortner, G., Pfenninger, M., Kahl, G. and Weising, K., 1996. Northern blot analysis of simple repetitive sequence transcription in plants. Electrophoresis' 17: 1183-1189.

Grandbastien, M.A., Spielmann, A. and Caboche, M., 1989. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature; 337: 376-380.

Grant, V., 1981. *Plant speciation*. 2nd edition. *New York: Columbia University Press xii*, 563p.

Greenblatt, I.M. and Brink, R.A., 1962. Twin mutations in medium variegated pericarp maize. Genetics; 47: 489.

Greilhuber, J. and Leitch, I.J., 2013. Genome size and the phenotype. In *Plant Genome Diversity Volume 2* (pp. 323-344). Springer Vienna.

Grimanelli, D., Leblanc, O., Perotti, E. and Grossniklaus, U., 2001a. Developmental genetics of gametophytic apomixis. Trends Genet. 17: 597-604.

Grimanelli, D., Tohme, J.M. and González de León, D., 2001b. Applications of molecular genetics in apomixis research.

Grossniklaus, U., Nogler, G.A. and van Dijk, P.J., 2001. How to avoid sex the genetic control of gametophytic apomixis. Plant Cell; 13: 1491-1498.

Grossniklaus, U., Spillane, C., Page, D.R. and Köhler, C., 2001. Genomic imprinting and seed development: endosperm formation with and without sex. Curr. Opin. Plant Biol. 4: 21-27.

Group, C.P.W., Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L. and Fazekas, A.J., 2009. A DNA barcode for land plants. P. Natl. Acad. Sci. USA. 106: 12794-12797.

Gustafsson, A., 1935. Studies on the mechanism of parthenogenesis. Hereditas; 21: 1-112.

Hamada, H., Petrino, M.G. and Kakunaga, T., 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. P. Natl. Aca. Sci. 79: 6465-6469.

Hand, M.L. and Koltunow, A.M., 2014. The genetic control of apomixis: asexual seed formation. Genetics, 197: 441-450.

Hanna, W. and Bashaw, E.C., 1987. Apomixis: its identification and use in plant breeding. Crop Sci. 27: 1136-1139.

Hanna, W.W., 1995. Use of apomixis in cultivar development. Adv. Agron. 54: 333–350.

Hansen, C. and Heslop-Harrison, J.S., 2004. Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. Adv. Bot. Res. 41: 165-193.

Hansen, L.J., Chalker, D.L. and Sandmeyer, S.B., 1988. Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. Mol. Cell. Biol. 8: 5245-5256.

Hearne, C.M., Ghosh, S. and Todd, J.A., 1992. Microsatellites for linkage analysis of genetic traits. Trends Genet; 8: 288-294.

Hebert, P.D., Ratnasingham, S. and de Waard, J.R., 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. P. Biol. Sci. 270: S96-S99.

Heenan, P.B., Dawson, M.I. and Bicknell, R.A., 2002. Evidence for apomictic seed formation in *Coprosma waima* (Rubiaceae). New Zealand Journal of Botany, 40(3), pp.347-355.

Heslop-Harrison, J., 1959. *Apomixis, environment and adaptation*. publisher not identified.

Heslop-Harrison, J., 1972. Sexuality of angiosperms. *Plant physiology—a treatise (ed. FC Steward)*; 106: 133-271.

Heslop-Harrison, J.P., Brandes, A., Taketa, S., Schmidt, T., Vershinin, A.V., Alkhimova, E.G., Kamm, A., Doudrick, R.L., Schwarzacher, T., Katsiotis, A. and Kubis, S., 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetica; 100: 197-204.

Heslop-Harrison, J.P., Brandes, A., Taketa, S., Schmidt, T., Vershinin, A.V., Alkhimova, E.G., Kamm, A., Doudrick, R.L., Schwarzacher, T., Katsiotis, A. and Kubis, S., 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetica; 100: 197-204.

Heslop-Harrison, J.S. and Schwarzacher, T., 1993. Molecular cytogenetics—biology and applications in plant breeding. In *Chromosomes today* (pp. 191-198). Springer Netherlands.

Heslop-Harrison, J.S. and Schwarzacher, T., 1996. Genomic southern and *in situ* hybridization for plant genome analysis. Methods of genome analysis in plants; 163-179.

Heslop-Harrison, J.S. and Schwarzacher, T., 2011. Organisation of the plant genome in chromosomes. Plant J. 66: 18-33.

Hickey, D.A., 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics; 101: 519-531.

Hilbricht, T., Varotto, S., Sgaramella, V., Bartels, D., Salamini, F. and Furini, A., 2008. Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the *plant Craterostigma plantagineum*. New Phytol.; 179: 877-887.

Hill, P., Burford, D., Martin, D.M. and Flavell, A.J., 2005. Retrotransposon populations of *Vicia* species with varying genome size. Mol. Genet. Genom. 273: 371-381.

Hirochika, H. and Hirochika, R., 1993. Ty1-copia group retrotransposons as ubiquitous components of plant genomes. Jpn. J. Genet. 68: 35-46.

Hollingsworth, P.M., Li, D.Z., van der Bank, M. and Twyford, A.D., 2016. Telling plant species apart with DNA: from barcodes to genomes. Phil. Trans. R. Soc. B. 371: 20150338.

Holmquist, G.P., 1992. Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. 51: 17.

Hörandl, E. and Hojsgaard, D., 2012. The evolution of apomixis in angiosperms: a reappraisal. Plant Biosystems; 146: 681-693.

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

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H. and Haberer, G., 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature Genet. 43: 476-481.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P. and Ren, Y., 2009. The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. 41: 1275-1281.

Hua-Van, A., Le Rouzic, A., Boutin, T.S., Filée, J. and Capy, P., 2011. The struggle for life of the genome's selfish architects. Biolo. Direct; 6: 1.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C. and Capy, P., 2005. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. Cytogenet. Genome Res. 110: 426-440.

Hughes, J. and Richards, A.J., 1988. The genetic structure of populations of sexual and asexual *Taraxacum* (dandelions). Heredity; 60: 161-171.

Idänheimo, N., Gauthier, A., Salojärvi, J., Siligato, R., Brosché, M., Kollist, H., Mähönen, A.P., Kangasjärvi, J. and Wrzaczek, M., 2014. The Arabidopsis thaliana cysteine-rich receptor-like kinases CRK6 and CRK7 protect against apoplastic oxidative stress. Biochem. Bioph. Res. Co. 445: 457-462.

II, A., 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141: 399-436.

Iltis, H., 1932. *Life of Mendel*. Life of Mendel.

International, R.G.S.P., 2005. The map-based sequence of the rice genome. Nature; 436: 793.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. and Nicaud, S., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature; 431: 946-957.

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

Jamilena M. C. Ruiz Rejon M. Ruiz Rejon 1993. Repetitive DNA sequence families in *Crepis capillaris*. Chromosoma 102: 272-278.

Jamilena, M., Garrido-Ramos, M., Rejon, M.R., Rejon, C.R. and Parker, J.S., 1995. Characterisation of repeated sequences from microdissected B chromosomes of *Crepis capillaris*. Chromosoma; 104: 113-120.

Jansen, R.K. and Palmer, J.D., 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). P.Natl. Acad. Sci. 84: 5818-5822.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W. and Lee, S.B., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. P.Natl. Acad. Sci. 104: 19369-19374.

Jansen, R.K., Raubeson, L.A., Boore, J.L., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., Fourcade, H.M. and Kuehl, J.V., 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 395: 348-384.

Jurka, J. and Kapitonov, V.V., 2001. PIFs meet Tourists and Harbingers: a superfamily reunion. P. Natl. Acad. Sci. 98: 12315-12316.

Jurka, J., Kapitonov, V.V., Kohany, O. and Jurka, M.V., 2007. Repetitive sequences in complex genomes: structure and evolution. Annu. Rev. Genomics Hum. Genet. 8: 241-259.

Jurka, J., Walichiewicz, J. and Milosavljevic, A., 1992. Prototypic sequences for human repetitive DNA. J. Mol. Evol. 35: 286-291.

Kakutani, T., Munakata, K., Richards, E.J. and Hirochika, H., 1999. Meiotically and mitotically stable inheritance of DNA hypomethylation induced by ddm1 mutation of *Arabidopsis thaliana*. Genetics; 151: 831-838.

Kalendar, R. and Schulman, A.H., 2006. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nature Prot. 1: 2478-2484.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. P.Natl. Acad. Sci. 97: 6603-6607.

Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A. and Schulman, A.H., 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics; 166: 1437-1450.

Kanamoto, H., Yamashita, A., Okumura, S., Hattori, M. and Tomizawa, K.I., 2004, March. The complete genome sequence of the *Lactuca sativa* (lettuce) chloroplast. In *Plant and Cell Physiology Supplement Supplement to Plant and Cell Physiology Vol. 45* (pp. 045-045). The Japanese Society of Plant Physiologists.

Kane, N.C., Gill, N., King, M.G., Bowers, J.E., Berges, H., Gouzy, J., Bachlava, E., Langlade, N.B., Lai, Z., Stewart, M. and Burke, J.M., 2011. Progress towards a reference genome for sunflower. Botany; 89: 429-437.

Kapitonov, V.V. and Jurka, J., 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J. Mol. Eevol. 48: 248-251.

Kapitonov, V.V. and Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9: 411-412.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. and Thierer, T., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics; 28: 1647-1649.

Kelley, D.R., Schatz, M.C. and Salzberg, S.L., 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11: R116.

Kelley, D.R., Schatz, M.C. and Salzberg, S.L., 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11: 1.

Khokhlov, S.S., 1976. *"Evolutionary-genetic problems of Apomixis in angiosperms, in Apomixis and Breeding".* S.S. Khokhlov, Ed. Amerind Publ. Co. Pvt. Ltd., New Delhi, 1–102. (Translated from Russian.).

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica; 115: 49-63.

Kidwell, M.G., 2005. Transposable elements. *The Evolution of the genome;* 165-221.

Kilian, N., Gemeinholzer, B. and Lack, H.W., 2009. Cichorieae. *Systematics, evolution, and biogeography of Compositae*, pp.343-383.

Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P. and Yim, S.H., 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature; 479: 223-227.

Kim, J.H., Roh, J.Y., Kwon, D.H., Kim, Y.H., Yoon, K.A., Yoo, S., Noh, S.J., Park, J., Shin, E.H., Park, M.Y. and Lee, S.H., 2014. Estimation of the genome sizes of the chigger mites *Leptotrombidium pallidum* and *Leptotrombidium scutellare* based on quantitative PCR and k-mer analysis. Parasit Vectors; 7: 279.

Kim, K.J., Choi, K.S. and Jansen, R.K., 2005. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol. Biol. Evol. 22: 1783-1792.

King LM, Schaal BA., 1990. Genotypic variation within asexual lineages of *Taraxacum officinale*. P. Natl. Acad. Sci. USA. 87: 998-1002.

King, L.M., 1993. Origins of genotypic variation in North American dandelions inferred from ribosomal DNA and chloroplast DNA restriction enzyme analysis. Evolution; 136-151.

Kirschner, J., Drábková, L.Z., Štěpánek, J. and Uhlemann, I., 2015. Towards a better understanding of the *Taraxacum* evolution (Compositae–Cichorieae) on the basis of nrDNA of sexually reproducing species. Plant Syst. Evol. 301: 1135-1156.

Kirschner, J., Oplaat, C., Verhoeven, K.J., Zeisek, V., Uhlemann, I., Trávníček, B., Räsänen, J., Wilschut, R.A. and Štěpánek, J., 2016. Identification of oligoclonal agamospermous microspecies: taxonomic specialists versus microsatellites. Preslia, 88: 1-17.

Kirschner, J., Štepánek, J., Mes, T.H.M., Nijs, J.D., Oosterveld, P., Štorchová, H. and Kuperus, P., 2003. Principal features of the cpDNA evolution in *Taraxacum* (Asteraceae, Lactuceae): a conflict with taxonomy. Plant Syst. Evol. 239: 231-255.

Kobayashi, S., Goto-Yamamoto, N. and Hirochika, H., 2004. Retrotransposon-induced mutations in grape skin color. Science; 304: 982-982.

Koltunow, A., 2012. Apomixis. eLS.

Koltunow, A.M. and Grossniklaus, U., 2003. Apomixis: a developmental perspective. Annu. Rev. Plant Biol. 54: 547-574.

Koltunow, A.M. and Tucker, M.R., 2008. Functional embryo sac formation in Arabidopsis without meiosis—one step towards asexual seed formation (apomixis) in crops?. J. Biosciences; 33: 309-311.

Koltunow, A.M., 1993. Apomixis: embryo sacs and embryos formed without meiosis or fertilization in ovules. Plant Cell; 5: 1425.

Koltunow, A.M., Bicknell, R.A. and Chaudhury, A.M., 1995. Apomixis: Molecular strategies for the generation of genetically identical seeds without fertilization. Plant Physiol. 108: 1345.

Koltunow, A.M., Johnson, S.D. and Bicknell, R.A., 1998. Sexual and apomictic development in *Hieracium*. Sex. Plant Repro. 11: 213-230.

Koltunow, A.M., Johnson, S.D. and Okada, T., 2011. Apomixis in hawkweed: Mendel's experimental nemesis. J. Exp. Bot. 62: 1699-1707.

Koltunow, A.M., Johnson, S.D., Lynch, M., Yoshihara, T. and Costantino, P., 2001. Expression of rolB in apomictic *Hieracium piloselloides* Vill. causes ectopic meristems in planta and

changes in ovule formation, where apomixis initiates at higher frequency. Planta; 214: 196-205.

Koltunow, A.M., Soltys, K., Nito, N. and McClure, S., 1995. Anther, ovule, seed, and nucellar embryo development in *Citrus sinensis* cv. *Valencia*. Can.J. Bot. 73: 1567-1582.

Konieczny, A., Voytas, D.F., Cummings, M.P. and Ausubel, F.M., 1991. A superfamily of *Arabidopsis thaliana* retrotransposons. Genetics; 127: 801-809.

Krassovsky, K. and Henikoff, S., 2014. Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster. BMC genomics; 15: 1.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. and Janzen, D.H., 2005. Use of DNA barcodes to identify flowering plants. P. Natl. Acad. Sci. USA. 102: 8369-8374.

Kubis, S., Schmidt, T. and Heslop-Harrison, J.S.P., 1998. Repetitive DNA elements as a major component of plant genomes. Ann. Bot. 82: 45-55.

Kubis, S.E., Castilho, A.M., Vershinin, A.V. and Heslop-Harrison, J.S.P., 2003. Retroelements, transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somaclonal variation. Plant Mol. Biol. 52: 69-79.

Kuhn, G.C., Sene, F.M., Moreira-Filho, O., Schwarzacher, T. and Heslop-Harrison, J.S., 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Res. 16: 307-324.

Kuhrová, V., Bezděk, M., Vyskot, B., Koukalova, B. and Fajkus, J., 1991. Isolation and characterization of two middle repetitive DNA sequences of nuclear tobacco genome. Theor. App. Genet. 81: 740-744.

Kumar, A. and Bennetzen, J.L., 1999. Plant retrotransposons. Annu. Rev. Genet. 33: 479-532.

Kumar, A. and Bennetzen, J.L., 2000. Retrotransposons: central players in the structure, evolution and function of plant genomes. Trends Plant Sci. 5: 509-510.

Kumar, S., Hahn, F.M., McMahan, C.M., Cornish, K. and Whalen, M.C., 2009. Comparative analysis of the complete sequence of the plastid genome of *Parthenium argentatum*

and identification of DNA barcodes to differentiate Parthenium species and lines. BMC Plant Biol. 9: 131.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29: 4633-4642.

Kurtz, S., Narechania, A., Stein, J.C. and Ware, D., 2008. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics; 9: 517.

Lapitanz, N.L., 1992. Organization and evolution of higher plant nuclear genomes. Genome; 35: 171-181.

Leigh, F., Kalendar, R., Lea, V., Lee, D., Donini, P. and Schulman, A.H., 2003. Comparison of the utility of barley retrotransposon families for genetic analysis by molecular marker techniques. Mol. Genet. Genom. 269: 464-474.

Leitch, I.J. and Heslop-Harrison, J.S., 1993. Physical mapping of four sites of 5S rDNA sequences and one site of the α-amylase-2 gene in barley (*Hordeum vulgare*). Genome; 36: 517-523.

Leitch, I.J. and Leitch, A.R., 2013. Genome size diversity and evolution in land plants. In *Plant Genome Diversity Volume 2* (pp. 307-322). Springer Vienna.

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.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y. and Zhang, Z., 2010. The sequence and *de novo* assembly of the giant panda genome. Nature; 463: 311-317.

Lisch, D., 2013. How important are transposons for plant evolution?. Nat. Rev. Genetics; 14: 49-61.

Liu, P.L., Wan, Q., Guo, Y.P., Yang, J. and Rao, G.Y., 2012. Phylogeny of the genus *Chrysanthemum* L.: evidence from single-copy nuclear gene and chloroplast DNA sequences. PloS One; 7: e48970.

Liu, Y., Huo, N., Dong, L., Wang, Y., Zhang, S., Young, H.A., Feng, X. and Gu, Y.Q., 2013. Complete chloroplast genome sequences of Mongolia medicine *Artemisia frigida* and phylogenetic relationships with other plants. PLoS One; 8: e57533.

Long, E.O. and Dawid, I.B., 1980. Repeated genes in eukaryotes. Ann. Rev. Biochemistry; 49: 727-764.

López-Flores, I. and Garrido-Ramos, M.A., 2012. The repetitive DNA content of eukaryotic genomes. In *Repetitive DNA* (Vol. 7, pp. 1-28). Karger Publishers.

Lowe, T.M. and Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(: 955-964.

Lu, C., Shen, Q., Yang, J., Wang, B. and Song, C., 2016. The complete chloroplast genome sequence of Safflower (*Carthamus tinctorius* L.). Mitochondrial DNA A; 27: 3351-3353.

Luo, J., Hou, B.W., Niu, Z.T., Liu, W., Xue, Q.Y. and Ding, X.Y., 2014. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PloS One; 9: e99016.

Lutts, S., Ndikumana, J. and Louant, B.P., 1994. Male and female sporogenesis and gametogenesis in apomictic *Brachiaria brizantha*, *Brachiaria decumbens* and F1 hybrids with sexual colchicine induced tetraploid Brachiaria ruziziensis. Euphytica; 78: 19-25.

Lyman, J.C. and Ellstrand, N.C., 1984. Clonal diversity in *Taraxacum officinale* (Compositae), an apomict. Heredity; 53: 1-10.

Lysak, M.A., Dolez, M., Horry, J.P., Swennen, R. and Dolez, J., 1999. Flow cytometric analysis of nuclear DNA content in Musa. Theoretical App. Genet. 98: 1344-1350.

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A. and Vyskot, B., 2011. Next generation sequencing-based analysis of repetitive DNA in the model dioceous plant *Silene latifolia*. PLoS One; 6: e27335.

Macas, J., Neumann, P. and 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: 1.

Maheshwari, P., 1950. Embryology in relation to taxonomy. *Proc. Brit. Ass. Adv. Sci.* General article Review article, Ovule Embryology (PMBD, 185204142).

Majeský Ľ, Vašut RJ, Kitner M, Trávníček B., 2012. The pattern of genetic variability in apomictic clones of *Taraxacum officinale* indicates the alternation of asexual and sexual histories of apomicts. PLoS One. 7: e41868.

Majeský, Ľ. 2013. Microevolutionary processes in apomictic genus *Taraxacum*. Dissertation, Palacký University, Olomouc.

Majeský, Ľ., Vašut, R.J. and Kitner, M., 2015. Genotypic diversity of apomictic microspecies of the *Taraxacum scanicum* group (*Taraxacum* sect. Erythrosperma). *Plant Systematics and Evolution*, *301*(8), pp.2105-2124.

Makałowski, W., Pande, A., Gotea, V. and Makałowska, I., 2012. Transposable elements and their identification. Evol. Genomics: Stat. Comput. Methods; Volume 1; 337-359.

Manninen, I. and Schulman, A.H., 1993. BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol. Biol. 22: 829-846.

Manninen, O., Kalendar, R., Robinson, J. and Schulman, A.H., 2000. Application of BARE-1 retrotransposon markers to the mapping of a major resistance gene for net blotch in barley. MGG; 264: 325-334.

Mao, L., Wood, T.C., Yu, Y., Budiman, M.A., Tomkins, J., Woo, S.S., Sasinowski, M., Presting, G., Frisch, D., Goff, S. and Dean, R.A., 2000. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. 10: 982-990.

Marçais, G. and Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics; 27: 764-770.

Marlor, R.L., Parkhurst, S.M. and Corces, V.G., 1986. The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cel. Biol. 6: 1129-1134.

Matsuoka, Y. and Tsunewaki, K., 1996. Wheat retrotransposon families identified by reverse transcriptase domain analysis. Mol. Biol. Evol. 13: 1384-1392.

Matsuoka, Y. and Tsunewaki, K., 1999. Evolutionary dynamics of Ty1-copia group retrotransposons in grass shown by reverse transcriptase domain analysis. Mol. Biol. Evol. 16: 208-217.

McCauley, D.E., Sundby, A.K., Bailey, M.F. and Welch, M.E., 2007. Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations. Am. J. Bot. 94: 1333-1337.

McClintock, Barbara. 1952. Mutable loci in maize: Origins of instability at the A1 and A2 loci. Instability of Sh1 action induced by Ds. Summary. Carnegie Institution of Washington Year Book No. 51 [Issued 12 December 1952, submitted June 1952]: 212-219.

Melsted, P. and Pritchard, J.K., 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics; 12: 1.

Mendel G., 1866. Versuche über Pflanzenhybriden, Verhandlungen des naturforschenden Vereines in Brunn. 4: 44.

Menken, S.B., Smit, E. and Hans (J.) CM Den Nijs, 1995. Genetical population structure in plants: gene flow between diploid sexual and triploid asexual dandelions (*Taraxacum* section Ruderalia). Evolution; 1108-1118.

Menzel, G., Heitkam, T., Seibt, K.M., Nouroz, F., Müller-Stoermer, M., Heslop-Harrison, J.S. and Schmidt, T., 2014. The diversification and activity of hAT transposons in Musa genomes. Chromosome Res. 22: 559-571.

Mes, T.H., Kuperus, P., Kirschner, J., Stepanek, J., Oosterveld, P., Storchova, H. and den Nijs, J.C., 2000. Hairpins involving both inverted and direct repeats are associated with homoplasious indels in non-coding chloroplast DNA of *Taraxacum* (Lactuceae: Asteraceae). Genome; 43: 634-641.

Mes, T.H., Kuperus, P., Kirschner, J., Štepánek, J., Štorchová, H., Oosterveld, P. and Den Nijs, J.C.M., 2002. Detection of genetically divergent clone mates in apomictic dandelions. Mol. Ecol. 11: 253-265.

Meyers, B.C., Tingey, S.V. and Morgante, M., 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 11: 1660-1676.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L. and Salzberg, S.L., 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature; 452: 991-996.

Mogie M, Ford H., 1988. Sexual and asexual *Taraxacum* species. Bot J Linn Soc. 35: 155-168.

Mogie, M. and Richards, A.J., 1983. Satellited chromosomes, systematics and phylogeny *in Taraxacum* (Asteraceae). Plant Syst. Evol. 141: 219-229.

Mogie, M., 1985. Morphological, developmental and electrophoretic variation within and between obligately apomictic *Taraxacum* species. Biol. J. ournal of the Linn. Soc. 24: 207-216.

Mogie, M., 1992. *The evolution of asexual reproduction in plants* (Vol. 412442205). London: Chapman and Hall 276p.-. ISBN.

Moore, M.J., Bell, C.D., Soltis, P.S. and Soltis, D.E., 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. P. Natl. Acad. Sci. 104: 19363-19368.

Moore, M.J., Soltis, P.S., Bell, C.D., Burleigh, J.G. and Soltis, D.E., 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. P. Natl. Acad. Sci. 107: 4623-4628.

Moran, N.A., 1992. The evolution of aphid life cycles. Annu. Rev. Entomol. 37: 321-348.

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

Mount, S.M., and Rubin, G.M., 1985. Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. Mol. Cel. Biol. 5: 1630-1638.

Mráz, P., Chrtek, J. and Fehrer, J., 2011. Interspecific hybridization in the genus *Hieracium* s. str.: evidence for bidirectional gene flow and spontaneous allopolyploidization. Plant Syst. Evol. 293: 237-245.

Mur, L.A., Allainguillaume, J., Catalán, P., Hasterok, R., Jenkins, G., Lesniewska, K., Thomas, I. and Vogel, J., 2011. Exploiting the Brachypodium Tool Box in cereal and grass research. New Phytol. 191: 334-347.

Musiał, K., Płachno, B.J., Świątek, P. and Marciniuk, J., 2013. Anatomy of ovary and ovule in dandelions (*Taraxacum*, Asteraceae). Protoplasma, 250: 715-722.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A. and Anson, E.L., 2000. A whole-genome assembly of Drosophila. Science; 287: 2196-2204.

Nagl, W., 1976. DNA endoreduplication and polyteny understood as evolutionary strategies. Nature; 261: 614-615.

Nair, A.S., Teo, C.H., Schwarzacher, T. and Harrison, P.H., 2005. Genome classification of banana cultivars from South India using IRAP markers. Euphytica; 144: 285-290.

Nassif, N., Penney, J., Pal, S., Engels, W.R. and Gloor, G.B., 1994. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. Mol. Cel. Biol. 14: 1613-1625.

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg, L. and Cavallini, A., 2013. The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. BMC Genomics; 14: 1.

Naumova, T.N., 1993. Apomixis in Angiosperms. *Nucellar and Integumentary Embryony*. CRC Press LLC, Boca Raton, FL.

Neumann, P., Požárková, D. and Macas, J., 2003. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plant Mol. Biol. 53: 399-410.

Nie, X., Lv, S., Zhang, Y., Du, X., Wang, L., Biradar, S.S., Tan, X., Wan, F. and Weining, S., 2012. Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). PloS One; 7: e36869.

Nogler, G.A., 1984. Gametophytic apomixis. In *Embryology of angiosperms* (pp. 475-518). Springer Berlin Heidelberg.

Nogler, G.A., 2006. The lesser-known Mendel: his experiments on Hieracium. Genetics, 172: 1-6.

Nouroz, F., 2012. The structures and abundance of transposable elements contributing to genome diversity in the diploid and polyploid Brassica and Musa crops (Doctoral dissertation, University of Leicester).

Nouroz, F., Noreen, S. and Heslop-Harrison, J.S., 2015. Evolutionary genomics of miniature inverted-repeat transposable elements (MITEs) in Brassica. Mol. Genet. Genomics; 290: 2297-2312.

Novak, J.D., 2010. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Routledge.

Novák, P., Hřibová, E., Neumann, P., Koblížková, A., Doležel, J. and Macas, J., 2014. Genome-wide analysis of repeat diversity across the family Musaceae. PloS One; 9: e98918.

Novák, P., Neumann, P. and Macas, J., 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics; 11: 1.

Noyes, R.D. and Rieseberg, L.H., 2000. Two independent loci control agamospermy (apomixis) in the triploid flowering plant *Erigeron annuus*. Genetics; 155: 379-390.

Noyes, R.D., 2007. Apomixis in the Asteraceae: diamonds in the rough. Funct. Plant Sci. Biotech. 1: 207-222.

Noyes, R.D., Baker, R. and Mai, B., 2007. Mendelian segregation for two-factor apomixis in *Erigeron annuus* (Asteraceae). Heredity; 98: 92-98.

Nygren, A., 1954. *Apomixis in the angiosperms. II.* Bot. Rev. 20: 577-649.

Nygren, A., 1967. Apomixis in the angiosperms. In *Sexualität· Fortpflanzung Generationswechsel/Sexuality·* Reproduction Alternation of Generations (pp. 551-596). Springer Berlin Heidelberg.

O'Connell, L.M. and Eckert, C.G., 1999. Differentiation in sexuality among populations of *Antennaria parlinii* (Asteraceae). Int. j. Plant Sci. 160: 567-575.

Okada, T., Ito, K., Johnson, S.D., Oelkers, K., Suzuki, G., Houben, A., Mukai, Y. and Koltunow, A.M., 2011. Chromosomes carrying meiotic avoidance loci in three apomictic eudicot *Hieracium* subgenus Pilosella species share structural features with two monocot apomicts. Plant Physiol. 157: 1327-1341.

Orgel, L.E. and Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature; 284: 604.

Ozias-Akins, P. and van Dijk, P.J., 2007. Mendelian genetics of apomixis in plants. Annu. Rev. Genet. 41: 509-537.

Ozias-Akins, Peggy. 2006. Apomixis: Developmental Characteristics and Genetics. Critical Rev. Plant Sci. 25: 199-214.

Ozias-Akins, Peggy. 2007. Apomixis: Developmental characteristics and genetics, *critical reviews in plant sciences, 25:2*, 199-214.

Pagán, H.J., Macas, J., Novák, P., McCulloch, E.S., Stevens, R.D. and Ray, D.A., 2012. Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. Genome Biol. Evol. 4: 575-585.

Palmer, J.D. and Thompson, W.F., 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell; 29: 537-550.

Parks, M., Cronn, R. and Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology; 7: 84.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. and Schmutz, J., 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature; 457: 551-556.

Paun, O. and Hörandl, E., 2006. Evolution of hypervariable microsatellites in apomictic polyploid lineages of *Ranunculus carpaticola*: directional bias at dinucleotide loci. Genetics, 174: 387-398.

Pearce, S.R., Li, D., Flavell, A.J., Harrison, G., Heslop-Harrison, J.S. and Kumar, A., 1996. TheTy1-copia group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. MGG; 250: 305-315.

Pearce, S.R., Pich, U., Harrison, G., Flavell, A.J., Heslop-Harrison, J.P., Schubert, I. and Kumar, A., 1996. TheTy1-copia group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosome Res. 4: 357-364.

Pennisi, E., 2007. Human genetic variation. Science, 318: 1842-1843.

Pevzner, P.A., Tang, H. and Tesler, G., 2004. De novo repeat classification and fragment assembly. Genome Res. 14: 1786-1796.

Philipson, M.N., 1978. Apomixis in *Cortaderia jubata* (Gramineae). New Zeal. J. Bot.; 16: 45-59.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A. and Panaud, O., 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 16: 1262-1269.

Pires, J.C., Lim, K.Y., Kovarík, A., Matyásek, R., Boyd, A., Leitch, A.R., Leitch, I.J., Bennett, M.D., Soltis, P.S. and Soltis, D.E., 2004. Molecular cytogenetic analysis of recently evolved *Tragopogon* (Asteraceae) allopolyploids reveal a karyotype that is additive of the diploid progenitors. Am. J. Bot. 91: 1022-1035.

Potato Genome Sequencing Consortium, 2011. Genome sequence and analysis of the tuber crop potato. Nature; 475: 189-195.

Poulsen, G.B., Kahl, G. and Weising, K., 1993. Abundance and polymorphism of simple repetitive DNA sequences in *Bmssica napus* L. Theor. Appl. Genet. 85: 994-1000.

Prokopowich, C.D., Gregory, T.R. and Crease, T.J., 2003. The correlation between rDNA copy number and genome size in eukaryotes. Genome; 46: 48-50.

Ramachandra, C. and Raghavan, V., 1992. Apomixis in distant hybridization: in *Distant hybridization of crop plants* (Vol. 16). Springer Science & Business Media.

Ramulu, K.S., Sharma, V.K., Naumova, T.N., Dijkhuis, P. and van Lookeren Campagne, M.M., 1999. Apomixis for crop improvement. Protoplasma; 208: 196-205.

Raubeson, L.A. and Jansen, R.K., 1992. A rare chloroplast-DNA structural mutation is shared by all conifers. Biochem. Syst. Ecol. 20: 17-24.

Rauscher, G. and Simko, I., 2013. Development of genomic SSR markers for fingerprinting lettuce (*Lactuca sativa* L.) cultivars and mapping genes. BMC Plant Biol. 13: 1.

Reamon-Büttner, S.M., Schmidt, T. and Jung, C., 1999. AFLPs represent highly repetitive sequences in Asparagus officinalis L. Chromosome Res. 7: 297-304.

Reisch, C., 2004. Molecular differentiation between coexisting species of *Taraxacum* sect. Erythrosperma (Asteraceae) from populations in south-east and west Germany. Bot. J. Linn. Soc. 145: 109-117.

Richards, A. J., 1970. Eutriploid facultative agamospermy in *Taraxacum*. New Phytologist; 69: 761–774. JSTOR 2430530.

Richards, A. J., 1973. The origin of *Taraxacum* agamospecies. Bot. J. Linn. Soc. 66: 189-211.

Richards, A. J., 1986. *Plant breeding systems*. George Allen & Unwin.

Richards, A. J., 1989. A comparison of within-plant karyological heterogeneity between agamospermous and sexual *Taraxacum* (Compositae) as assessed by the nucleolar organiser chromosome. Plant Syst. Evol. 163: 177-185.

Richards, A. J., 1996. Why is gametophytic apomixis almost restricted to polyploids. The gametophyte-expressed lethal model. *Apomixis Newslett*, *9*, pp.1-3.

Richards, A. J., 1997. Dandelions of Great Britain and Ireland (Handbooks for Field dentification). *BSBI Publications.* p. 330. ISBN 978-0-901158-25-3.

Richards, A. J., 1997. *Plant breeding systems*. Garland Science.

Richards, A. J., 2003. Apomixis in flowering plants: an overview. Philos. Trans. R. Soc. B. 358:1085–1093.

Roche, D., Cong, P., Chen, Z., Hanna, W.W., Gustine, D.L., Sherwood, R.T. and Ozias-Akins, P., 1999. An apospory-specific genomic region is conserved between Buffelgrass (*Cenchrus ciliaris* L.) and *Pennisetum squamulatum* Fresen. Plant J. 19: 203-208.

Rosenberg, O., 1906. Über die Embryobildung in der Gattung *Hieracium*. Ber. Dtsch. Bot. Ges. 24, 157–161.

Rosenberg, O., 1907. Experimental and cytological studies in the *Hieracia*. II. Cytological studies on the apogamy in *Hieracium*. Bot. Tidsskr. 28, 143–170.

Rozen, S. and Skaletsky, H., 1999. Primer3 on the WWW for general users and for biologist programmers. Bioinformatics; 365-386.

Sabot, F. and Schulman, A.H., 2006. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity, 97: 381-388.

Sabot, F., Kalendar, R., Jääskeläinen, M., Wei, C., Tanskanen, J. and Schulman, A.H., 2006. Retrotransposons: metaparasites and agents of genome evolution. Isr. J. Ecol. Evol. 52: 319-330.

Saeidi, H., Rahiminejad, M.R. and Heslop-Harrison, J.S., 2008. Retroelement insertional polymorphisms, diversity and phylogeography within diploid, D-genome *Aegilops tauschii* (Triticeae, Poaceae) sub-taxa in Iran. Ann. Bot. 101: 855-861.

Saeidi, H., Rahiminejad, M.R., Vallian, S., Heslop-Harrison J. S., 2006. Biodiversity of diploid D-genome *Aegilops tauschii* Coss. In *Iran measured using microsatellites*. Genet Resour Crop Evol. 53: 1477-1484.

Saha, S., Bridges, S., Magbanua, Z.V. and Peterson, D.G., 2008. Computational approaches and tools used in identification of dispersed repetitive DNA sequences. Tropical Plant Biol. 1: 85-96.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

Salih, R.H.M., Majeský, Ľ., Schwarzacher, T., Gornall, R. and Heslop-Harrison, P., 2017. Complete chloroplast genomes from apomictic *Taraxacum* (Asteraceae): Identity and variation between three microspecies. PloS One; 12: e0168008.

Sambrook, J. and Russell, D.W., 2001. Southern hybridization, p. 6.33–6.64. Molecular cloning: a laboratory manual, 1.

Sang, T., Crawford, D. and Stuessy, T., 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. 84: 1120-1120.

Sanmiguel, P. and Bennetzen, J.L., 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. 82: 37-44.

SanMiguel, P., Tikhonov, A., Jin, Y.K. and Motchoulskaia, N., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science; 274: 765.

Santini, S., Cavallini, A., Natali, L., Minelli, S., Maggini, F. and Cionini, P., 2002. Ty1/copia-and Ty3/gypsy-like DNA sequences in Helianthus species. Chromosoma; 111: 192-200.

Sardos, J., Perrier, X., Doležel, J., Hřibová, E., Christelova, P., Kilian, A., Roux, N., 2016. DArT whole genome profiling provides insights on the evolution and taxonomy of edible Banana (*Musa* spp.). Ann. Bot. 118: 1269-78.

Särkinen, T. and George, M., 2013. Predicting plastid marker variation: can complete plastid genomes from closely related species help?. PLoS One; 8: e82266.

Savidan, Y., 1983. Genetics and utilization of apomixis for the improvement of guineograss (*Panicum maximum* Jocq.). In J.A. Smith and V.W. Hays (eds.), Proc. XIV International Grassland Congress, Lexington, Kentucky, 1981. *Boulder, Colcorodo: Westview Press*. Pp. 182-84.

Savidan, Y., 2000. Apomixis: genetics and breeding. *Plant Breeding Reviews 18*, 13-86.

Schmidt, T. and Heslop-Harrison, J.S., 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci. 3: 195-199.

Schmidt, T., 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol. Biol. 40: 903-910.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. and Minx, P., 2009. The B73 maize genome: complexity, diversity, and dynamics. Science; 326: 1112-1115.

Schulman, A.H. and Kalendar, R., 2005. A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cyto. Genome Res. 110: 598-605.

Schwarzacher, T. and Heslop-Harrison, J.S., 1991. *In situ* hybridization to plant telomeres using synthetic oligomers. Genome; 34: 317-323.

Schwarzacher, T. and Heslop-Harrison, P., 2000. *Practical in situ Hybridization*. BIOS Scientific Publishers Ltd.

Schwarzacher, T., Heslop-Harrison, J.S. and Richert-Pöggeler, K.R., 2015. Analysis of *Petunia vein* clearing virus (PVCV) sequences, retroelements and tandem repeats in *Petunia axillaris* N and *P. inflata* S6. Sherwood, R.T., Berg, C.C. and Young, B.A., 1994. Inheritance of apospory in buffelgrass. Crop Sci. 34: 1490-1494.

Shaw, J., Shafer, H.L., Leonard, O.R., Kovach, M.J., Schorr, M. and Morris, A.B., 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. Am. J. Bot. 101: 1987-2004.

Sherwood, R.T., Berg, C.C. and Young, B.A., 1994. Inheritance of apospory in buffelgrass. Crop Sci. 34: 1490-1494.

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I. and Okada, N., 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature; 388: 666-670.

Shimazaki, M., Fujita, K., Kobayashi, H. and Suzuki, S., 2011. Pink-colored grape berry is the result of short insertion in intron of colour regulatory gene. PLoS One; 6: e21308.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K. and Ohto, C., 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J., 5: 2043.

Singh, R., Ong-Abdullah, M., Low, E.T.L., Manaf, M.A.A., Rosli, R., Nookiah, R., Ooi, L.C.L., Ooi, S.E., Chan, K.L., Halim, M.A. and Azizi, N., 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature; 500: 335-339.

Singh, R.J., 2003. *Plant cytogenetics*. CRC press.

Skipper, K.A., Andersen, P.R., Sharma, N. and Mikkelsen, J.G., 2013. DNA transposon-based gene vehicles-scenes from an evolutionary drive. J. Biomed. Sci. 20: 1.

Slotkin, R.K. and Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nature Rev. Genet. 8: 272-285.

Šmarda, P., Bureš, P., Horová, L., Leitch, I.J., Mucina, L., Pacini, E., Tichý, L., Grulich, V. and Rotreklová, O., 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. P. Nat. Acad. Sci. 111: e4096-4102.

Smith, J., 1841. Notice of a plant which produces perfect seeds without any apparent action of pollen. Trans Linn. Soc. Lond. 18:509–512.

Solntseva, M.P., 1976. Basis of embryological classification of apomixis in angiosperms, in Apomixis and Breeding. *S.S. Khokhlov, Ed. Amerind Publ., New Delhi*, pp. 89–101. (Translated from Russian.)

Sørensen, T., 1958. Sexual chromosome-aberrants in triploid apomictic *Taraxaca.* Bot. Tidskr. 54: 1–22.

Sørensen, T., Gudjonsson, G., 1946. Spontaneous chromosome-aberrants in apomictic *Taraxaca*. Kon Dansk Vidensk Selsk Biol Skrift 4: 1–48.

Springer, N.M., Lisch, D. and Li, Q., 2016. Creating order from chaos: epigenome dynamics in plants with complex genomes. Plant Cell; 28: 314-325.

Stacey, M.W., Neumann, S.A., Dooley, A., Segna, K., Kelly, R.E., Nuss, D., Kuhn, A.M., Goretsky, M.J., Fecteau, A.H., Pastor, A. and Proud, V.K., 2010. Variable number of tandem repeat polymorphisms (VNTRs) in the ACAN gene associated with pectus excavatum. Clin. Genet. 78: 502-504.

Staton, S.E., Bakken, B.H., Blackman, B.K., Chapman, M.A., Kane, N.C., Tang, S., Ungerer, M.C., Knapp, S.J., Rieseberg, L.H. and Burke, J.M., 2012. The sunflower (Helianthus annuus L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72: 142-153.

Stebbins Jr, C.L., 1950. Variation and evolution in plants. *Variation and evolution in plants.* Columbia University Press, New York.

Stegemann S, Keuthe M, Greiner S, Bock R, 2012. Horizontal transfer of chloroplast genomes between plant species. P.. Natl. Acad. Sci. USA. 109: 2434-2438.

Stelzer, C.P., Schmidt, J., Wiedlroither, A. and Riss, S., 2010. Loss of sexual reproduction and dwarfing in a small metazoan. PLoS One; 5: e12854.

Sterk, A.A., Hommels, C.H., Jenniskens, M.J.P.J., Neuteboom, J.H., den Nijs, J.C.M., Oosterveld, P. and Segal, S., 1987. *Paardebloemen: planten zonder vader: variatie, evolutie en toepassingen van het geslacht paardebloem ('Taraxacum')*. Stichting Uitgeverij Koninklijke Nederlandse Natuurhistorische Vereniging.

Suda, J., Krahulcová, A., Trávníček, P., Rosenbaumová, R., Peckert, T. and Krahulec, F., 2007. Genome size variation and species relationships in *Hieracium* sub-genus *Pilosella* (Asteraceae) as inferred by flow cytometry. Ann. Bot. 100: 1323-1335.

Suomalainen, E., Saura, A. and Lokki, J., 1987. *Cytology and evolution in parthenogenesis*. CRC Press.

Suoniemi, A., Tanskanen, J. and Schulman, A.H., 1998. Gypsy-like retrotransposons are widespread in the plant kingdom. Plant J. 13: 699-705.

Sveinsson, S., Gill, N., Kane, N.C. and Cronk, Q., 2013. Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (*Theobroma cacao* L.) and related species. BMC genomics; 14: 502.

Sveinsson, S., Gill, N., Kane, N.C. and Cronk, Q., 2013. Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (*Theobroma cacao* L.) and related species. BMC genomics; 14: 1.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30: 2725-2729.

Tas, I.C. and Van Dijk, P.J., 1999. Crosses between sexual and apomictic dandelions (*Taraxacum*). I. The inheritance of apomixis. Heredity, 83: 707-714.

Tautz, D. and Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12: 4127-4138.

Temsch, E.M., Temsch, W., Ehrendorfer-Schratt, L. and Greilhuber, J., 2010. Heavy metal pollution, selection, and genome size: the species of the Žerjav study revisited with flow cytometry. J. Bot. 596542: 11 .

Tenaillon, M.I., Hufford, M.B., Gaut, B.S. and Ross-Ibarra, J., 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol. Evol. 3: 219-229.

Teo, C.H., Tan, S.H., Ho, C.L., Faridah, Q.Z., Othman, Y.R., Heslop-Harrison, J.S., Kalendar, R. and Schulman, A.H., 2005. Genome constitution and classification using retrotransposon-based markers in the orphan crop banana. J. Plant Biol. 48: 96-105.

Teo, C.H., Tan, S.H., Othman, Y.R. and Schwarzacher, T., 2002. The Cloning of Ty 1- copia-like Retrotransposons from 10 Varieties of Banana (*Musa* Sp.). J. Biochem. Mol. Biol. Biophysics; 6: 193-201.

Thomas, S., Bailey, J.P. and Rich, T.C.G., 2011. Pollen and chromosome studies in Hieracium sect. Alpestria (Asteraceae). Nordic J. Bot. 29: 244-248.

Timberlake, W.E., 1978. Low repetitive DNA content in Aspergillus nidulans. Science; 202: 973-975.

Timme, R.E., Kuehl, J.V., Boore, J.L. and Jansen, R.K., 2007. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am. J. Bot. 94: 302-312.

Toll-Riera, M., Radó-Trilla, N., Martys, F. and Albà, M.M., 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. 29: 883-886.

Tör, M., Lotze, M.T. and Holton, N., 2009. Receptor-mediated signalling in plants: molecular patterns and programmes. J. Exp. Bot. 60: 3645-3654.

Tóth, G., Gáspári, Z. and Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10: 967-981.

Traut W., 1991. Chromosomen. *Klassische und molekulare Zytogenetik*. Berlin, Heidelberg: Springer-Verlag.

Tremetsberger, K., Gemeinholzer, B., Zetzsche, H., Blackmore, S., Kilian, N. and Talavera, S., 2013. Divergence time estimation in Cichorieae (Asteraceae) using a fossil-calibrated relaxed molecular clock. Org. Divers. Evol. 13: 1-13.

Tucker, M.R. and Koltunow, A.M., 2009. Sexual and asexual (apomictic) seed development in flowering plants: molecular, morphological and evolutionary relationships. Funct. Plant Biol. 36: 490-504.

Turner, K.G. and Grassa, C.J., 2014. Complete plastid genome assembly of invasive plant, Centaurea diffusa. BioRxiv; p.005900.

Valárik, M., Šimková, H., Hřibová, E., Šafář, J., Doleželová, M. and Doležel, J., 2002. Isolation, characterization and chromosome localization of repetitive DNA sequences in bananas (Musa spp.). Chromosome Res. 10: 89-100.

Valle, C.D., Glienke, C. and Leguizamon, G.O.C., 1994. Inheritance of apomixis in *Brachiaria*, a tropical forage grass. Apomixis Newsl, 7: 42-43.

Van der Hulst RG, Mes TH, Den Nijs JC, Bachmann K., 2000. Amplified fragment length polymorphism (AFLP) markers reveal that population structure of triploid dandelions (*Taraxacum officinale*) exhibits both clonality and recombination. Mol Ecol. 9: 1-8.

Van der Hulst RG, Mes TH, Falque M, Stam P, Den Nijs JC, Bachmann K., 2003. Genetic structure of a population sample of apomictic dandelions. Heredity. 90: 326-335.

Van Dijk, P., de Jong, H., Vijverberg, K. and Biere, A., 2009. An apomixis-gene's view on dandelions. In Lost Sex (Chapter 22-pp. 475-493). Springer Netherlands.

Van Dijk, P., Hartog, M. and Delden, W.V., 1992. Single cytotype areas in autopolyploid Plantago media L. Biol. J. Linnean Soc. 46: 315-331.

Van Dijk, P.J. and Bakx-Schotman, J.T., 2004. Formation of unreduced megaspores (diplospory) in apomictic dandelions (*Taraxacum officinale*, sl) is controlled by a sex-specific dominant locus. Genetics; 166: 483-492.

Van Dijk, P.J. and Vijverberg, K., 2005. The significance of apomixis in the evolution of the angiosperms: a reappraisal. Regnum Vegetabile; 143: 101.

Van Dijk, P.J., 2003. Ecological and evolutionary opportunities of apomixis: insights from *Taraxacum* and Chondrilla. Philos. T. Roy. Soc. London B; 358: 1113-1121.

Van Dijk, P.J., Tas, I.C., Falque, M. and Bakx-Schotman, T., 1999. Crosses between sexual and apomictic dandelions (*Taraxacum*). II. The breakdown of apomixis. Heredity; 83: 715-721.

Van Oostrum, H., Sterk, A.A. and Wijsman, H.J.W., 1985. Genetic variation in agamospermous microspecies of *Taraxacum* sect. Erythrosperma and sect. Obliqua. Heredity; 55: 223-228.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J. and Malacarne, G., 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PloS One; 2: e1326.

Verduijn, M.H., Van Dijk, P.J. and Van Damme, J.M.M., 2004. The role of tetraploids in the sexual–asexual cycle in dandelions (*Taraxacum*). Heredity; 93: 390-398.

Vergnaud, G. and Denoeud, F., 2000. Minisatellites: mutability and genome architecture. Genome Res. 10: 899-907.

Vergnaud, G., 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. Nucleic Acids Res. 17: 7623-7630.

Verhoeven, K.J., Van Dijk, P.J. and Biere, A., 2010. Changes in genomic methylation patterns during the formation of triploid asexual dandelion lineages. Mol. Ecol. 19: 315-324.

Vicient, C.M., Jääskeläinen, M.J., Kalendar, R. and Schulman, A.H., 2001. Active retrotransposons are a common feature of grass genomes. Plant Physiol. 125: 1283-1292.

Vidic, T., Greilhuber, J., Vilhar, B. and Dermastia, M., 2009. Selective significance of genome size in a plant community with heavy metal pollution. Ecol. Appl. 19: 1515-1521.

Vijverberg, K., Milanovic-Ivanovic, S., Bakx-Schotman, T. and van Dijk, P.J., 2010. Genetic fine-mapping of DIPLOSPOROUS in *Taraxacum* (dandelion; Asteraceae) indicates a duplicated DIP-gene. BMC Plant Biol. 10: 1.

Vitte, C. and Panaud, O., 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet. Genome Res. 110: 91-107.

Voytas, D.F., Cummings, M.P., Koniczny, A., Ausubel, F.M. and Rodermel, S.R., 1992. Copia-like retrotransposons are ubiquitous among plants. P. Natl. Acad. Sci. 89: 7124-7128.

Vukich, M., Schulman, A.H., Giordani, T., Natali, L., Kalendar, R. and Cavallini, A., 2009. Genetic variability in sunflower (*Helianthus annuus* L.) and in the Helianthus genus as assessed by retrotransposon-based molecular markers. Theoret. Appl. Genet. 119: 1027-1038.

Walker, J.F., Zanis, M.J. and Emery, N.C., 2014. Comparative analysis of complete chloroplast genome sequence and inversion variation in *Lasthenia burkei* (Madieae, Asteraceae). Am. J. Bot. 101: 722-729.

Wang, M., Cui, L., Feng, K., Deng, P., Du, X., Wan, F., Weining, S. and Nie, X., 2015. Comparative analysis of Asteraceae chloroplast genomes: structural organization, RNA editing and evolution. Plant Mol. Biol. Reporter; 33: 1526-1538.

Wang, Z. H., H. Peng, and N. Kilian, 2013. Molecular phylogeny of the *Lactuca alliance* (Cichorieae subtribe Lactucinae, Asteraceae) with focus on their Chinese centre of diversity detects potential events of reticulation and chloroplast capture. PloS One; 8: e82692.

Warmke, H.E., 1954. Apomixis in *Panicum maximum*. Am. J. Bot. 41: 5-11.

White, S.E., Habera, L.F. and Wessler, S.R., 1994. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. P. Natl. Acad. Sci. 91: 11792-11796.

White, T.J., Bruns, T. and Lee, S., 1990. TaylorJ. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications*, *18*, pp.315-322.

Whitton, J., Sears, C.J., Baack, E.J. and Otto, S.P., 2008. The dynamic nature of apomixis in the angiosperms. IJPS, 169: 169-182.

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T., Graner, A., Ware, D. and Stein, N., 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics; 9: 1.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. and Paux, E., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973-982.

Williams, D., Trimble, W.L., Shilts, M., Meyer, F. and Ochman, H., 2013. R apid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genomics, 14: 537.

Witte, C.P., Le, Q.H., Bureau, T. and Kumar, A., 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. P. Natl. Acad. Sci. 98: 13778-13783.

Wittzell, H., 1999. Chloroplast DNA variation and reticulate evolution in sexual and apomictic sections of dandelions. Mol. Ecol. 8: 2023-2035.

Wojciechowski MF. IRLC (Inverted Repeat Lacking Clade). Version 11 July 2006: http://tolweb. org/IRLC_% 28Inverted_Repeat-lacking_ clade% 29/60358/2006.07. 11. The tree of life web project, http:// tolweb. org.

Wolfe KH, Li WH, Sharp PM., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. P. Natl. Acad. Sci. USA. 84: 9054-9058.

Woodhouse, M.R., Pedersen, B. and Freeling, M., 2010. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet; 6: e1000949.

Wyman, S.K., Jansen, R.K. and Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics; 20: 3252-3255.

Yang, S., Arguello, J.R., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., Zhao, R., Brunet, F., Peng, L. and Long, M., 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. PLoS Genet; 4: e3.

Yao, J.L., Dong, Y.H. and Morris, B.A., 2001. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. P. Natl. Acad. Sci. 98: 1306-1311.

Zaki, N. M., Madon, M., Schwarzacher. T., Heslop-Harrison, J. S., 2017. Chromosomes, cytology and molecular cytogenetics of oil palm. In *Oil Palm Breeding, Genetics and Genomics*. Eds Aik Chin Soh, Jerry Roberts and Sean Mayes. Taylor and Francis (In press).

Záveský, L., Jarolímová, V. and Štěpánek, J., 2005. Nuclear DNA content variation within the genus *Taraxacum* (Asteraceae). Folia Geobot. 40: 91-104.

Zehdi-Azouzi, S., Cherif, E., Moussouni, S., Gros-Balthazard, M., Naqvi, S.A., Ludeña, B., Castillo, K., Chabrillange, N., Bouguedoura, N., Bennaceur, M. and Si-Dehbi, F., 2015. Genetic structure of the date palm (*Phoenix dactylifera*) in the Old World reveals a strong differentiation between eastern and western populations. Ann. Bot. 116: 101-112.

Zeyl, C., Bell, G. and Green, D.M., 1996. Sex and the spread of retrotransposon Ty3 in experimental populations of Saccharomyces cerevisiae. Genetics; 143: 1567-1577.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. and Shen, B., 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PloS One; 6: e17915.

Zhang, Y., Li, L., Yan, T.L. and Liu, Q., 2014. Complete chloroplast genome sequences of Praxelis (*Eupatorium catarium* Veldkamp), an important invasive species. Gene; 549: 58-69.

CHAPTER 9

APPENDICES

9.1 Appendices from Chapter 2_ Solutions and media

9.1.1. Genomic DNA isolation

- CTAB buffer: 2% (w/v) cetyltrimethylammonium bromide, 100mM Tris-HCl, 1.4M NaCl, 20mM EDTA. (pH 7.5 - 8.0)
- DNA wash buffer: 76 % ethanol, 10mM ammonium acetate. No autoclaving.
- 10x TE buffer: 100mM Tris (tris-hydroxymethylamino-methane)-HCl, 10mM EDTA (ethylene-diamine-tetra-acetic acid). (pH 8.0)

9.1.2. Gel electrophoresis

- 6x gel loading buffer: 0.25% bromophenol blue, 0.25% xylene cyanol FF, 60% glycerol. No autoclaving and stored at 4°C.
- 50x TAE: 242g of Tris-base, 57.1ml of glacial acetic acid, 100ml of 0.5M EDTA. Final volume 1000ml with sterile distilled water. (pH 8.0)
- ethidium bromide (10 mg/ml): 1g ethidium bromide, 100ml of sterile distilled water. No autoclaving and stored at 4°C.

9.1.3 Cloning

- ampicillin (10mg/ml): 10 mg of ampicillin powder dissolved in 1 ml distilled water). No autoclaving and stored at -20°C.

- SOB medium: super optimal broth. 2% tryptone (Oxoid), 0.5% yeast extract (Oxoid), 8.5 mM NaCl (Fisher Scientific), 2.5 mM KCl (Fisher Scientific), 100 mM $MgCl_2$ (Fisher Scientific), pH 7

- LB (Lysogeny Broth) agar plates: 2.5% LB broth (Melford), 1.5% agar (ForMedium), 100 µg/ml ampicillin (Sigma-Aldrich), 80 µg/ml x-gal (Sigma-Aldrich), 0.5 mM IPTG (**i**so**p**ropyl β-D-1-**t**hio**g**alactopyranoside; Sigma-Aldrich), pH 7.2

- LB medium: Luria-Bertani. 10g Tryptone, 5g Yeast extract, 10g NaCl. Final volume 1000ml with sterile distilled water and autoclaved. pH 7.0).

- LB medium Agar: 10g tryptone, 5g yeast extract, 10g NaCl. Final volume 1000ml with sterile distilled water and 1.5% Agar.

- IPTG (200 mM): 476mg/ml isopropyl-B-D-thiogalacto-pyronoside (dissolved in 10ml distilled water). Filter sterilized and stored at -20°C.

- Xgal (40mg/ml): 1g of 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside with 25ml of dimethylformamide. Filter sterilized, stored at -20°C.

9.1.4. Chromosome preparations

- 10x Enzyme buffer: 100 mM citric acid (Fisher Scientific), 100 mM tri-sodium-citrate (Fisher Scientific). No autoclaving, stored at 4°C. (pH 4.6). Diluted to 1 x in deionised H_2O.

- 1x Enzyme solution: 3% (v/v) pectinase (13.5 U/ml; P4716; Sigma-Aldrich), 0.2% (w/v) cellulose (10 U/ml; Onzuka RS), 1.8% (w/v) cellulose (72 U/ml; 21947; Calbiochem) in 1x enzyme buffer. No autoclaving and stored at -20°C.

- 8-hydroxyquinoline (0.002M): Dissolving 0.29 g of S-*hydroxyquinoline* in 1000 ml of ddH2O. Store in the dark at 4°C.

9.1.5. Colourimetric dot-blot

- Buffer 1: 100 mM Tris-HCl [tris (hydroxymethyl) aminomethane], pH 7.5 (Sigma-Aldrich), 15 mM NaCl (Fisher Scientific)

- Buffer 2: 0.5% (w/v) blocking reagent (Roche), Prepared in buffer 1.

- Buffer 3: 100 mM Tris-HCl (Fisher Scientific), 100 mM NaCl (Fisher Scientific), 50 mM $MgCl_2$ (Fisher Scientific

9.1.6. Fluorescent in situ hybridization (FISH)

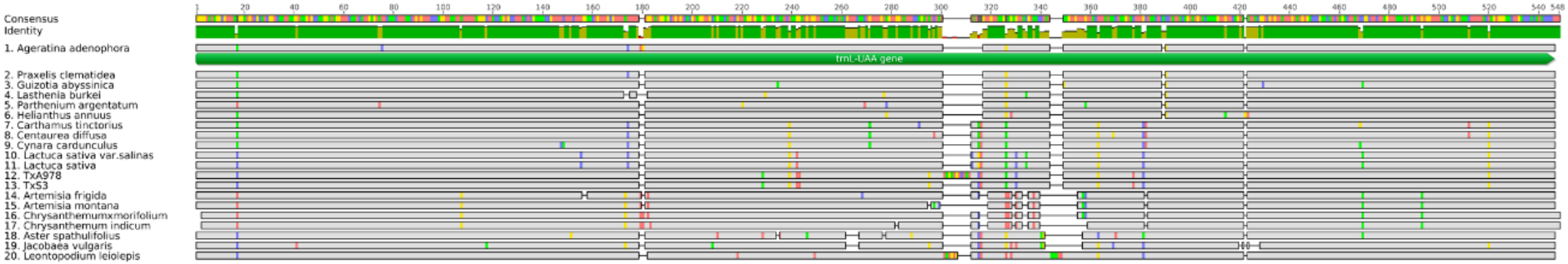
- SSC (20X): (saline sodium citrate, pH 7.0)/ 0.3M NaCl, 0.03M sodium citrate.
- EDTA (0.5 M): 186.1g disodium ethylenediamine tetraacetate.$2H_2O$ into 800ml of distilled water. Adjust pH to 8.0 with NaOH. Final volume 1 liter. (pH 8.0)
- Detection buffer: 4x SSC, 0.2% (v/v) Tween 20.
- SDS (20%): 2g Sodium dodecyl sulfate (SDS) with 8ml water Not autoclaved
- Blocking DNA: Autoclave genomic DNA at 114°C for 5min
- DAPI (100µg/ml): 5g of DAPI (4',6-diamidino-2-phenylindole) dissolved in Sigma water. Final volume 50ml. No autoclaving and stored at -20°C.
- DAPI was diluted in water for stock of 100µg/ml and then diluted with McIlvaine's buffer to the final concentration of 4µg/ml.
- McIlvaine's buffer: 0.1M citric acid, 0.2M di-sodium hydrogen phosphate. pH 7.0.
- Dextran sulfate (50%): 50 gm dextran sulfate with 100 ml distilled water, Filter sterilized and stored at -20°C.

9.1.7. Southern hybridization

- Southern denaturing solution: 0.25M NaOH, 1M HCl.
- Southern depurinating solution: 0.25M HCl.
- Southern neutralizing solution: 0.5M Tris-HCl, 3M NaCl. (pH 7.5)
- Southern transfer buffer: 0.4M NaOH.
- Salmon sperm DNA: 1mg/ml of sheared salmon sperm DNA.
- Wash buffer 1: 0.1M maleic acid, 0.15 M NaCl, 0.3% (v/v) Tween 20. pH 7.5.
- Buffer 1: 0.1M maleic acid, 0.15 M NaCl. pH 7.5.
- Buffer 2: 1% (w/v) blocking reagent (Roche Diagnostics) in buffer 1.
- Buffer 3: 0.1M Tris-HCl, 0.1 M NaCl. pH 9.5

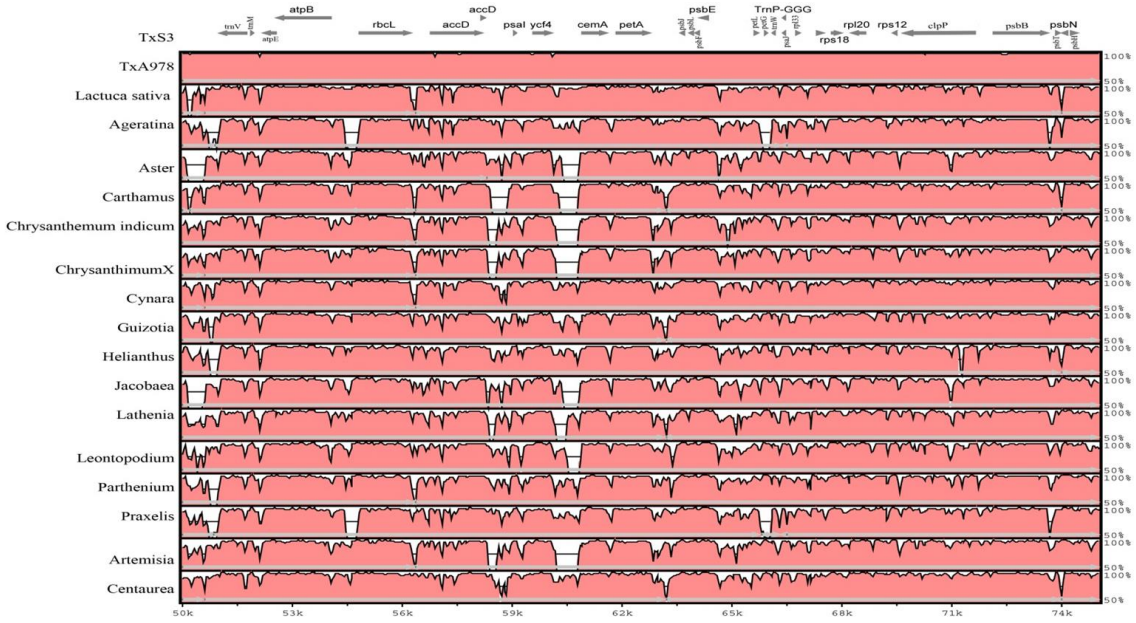
9.2 Appendices from Chapter 4

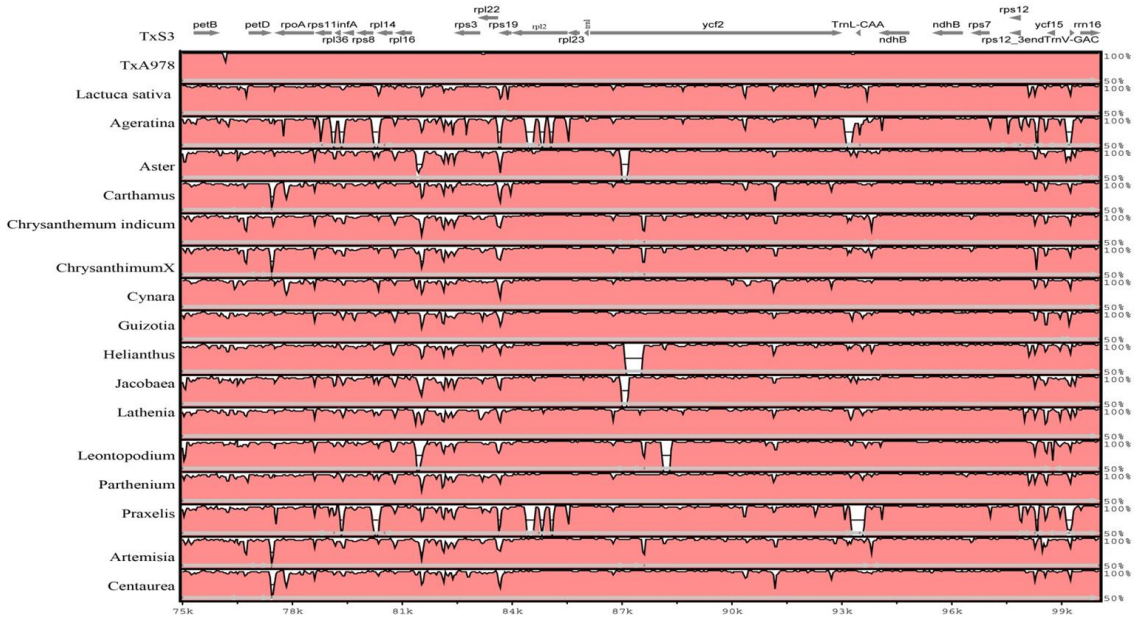
Supplementary Figure 4.1. Alignment of *trn*L-UAA sequence from 19 Asteraceae species including the two *Taraxacum* (A978 and O978) species sequenced in the present study. Arrowhead indicates a 22bp insertion in A978 with respect to O978 and other species.



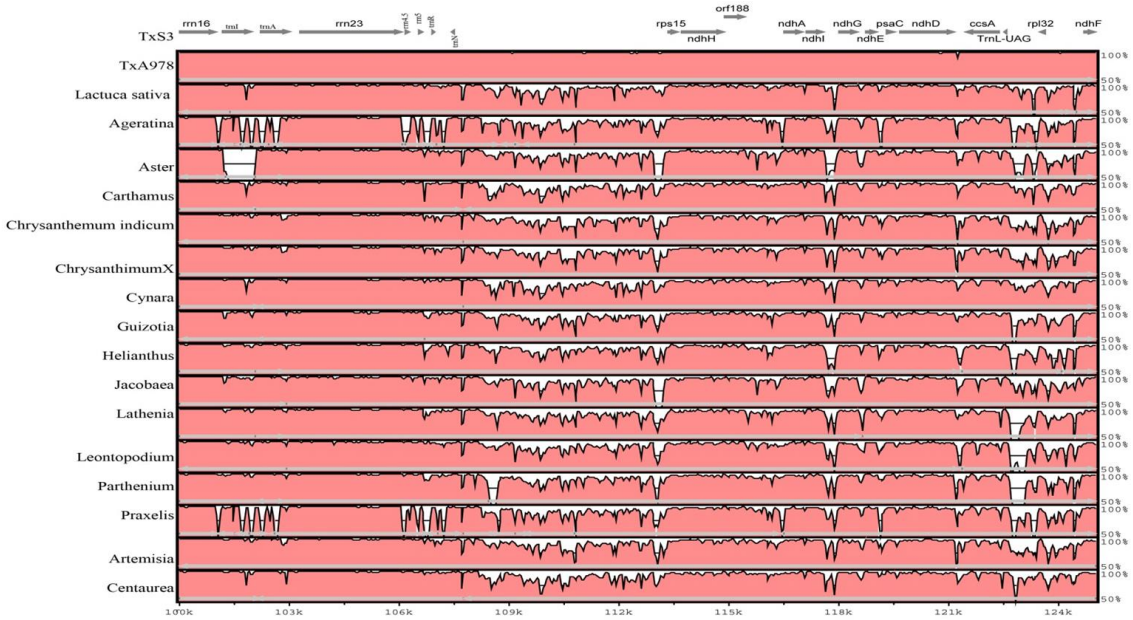
Supplementary Figure 4.2. Comparison of plastome sequences of 18 Asteraceae accessions, two *Taraxacum* plastomes generated in this study and 16 previously reported plastomes using mVISTA program. The Y-scale represents the percent of identity ranging from 50 to 100%. Arrows above the graphs indicate the direction of transcription.

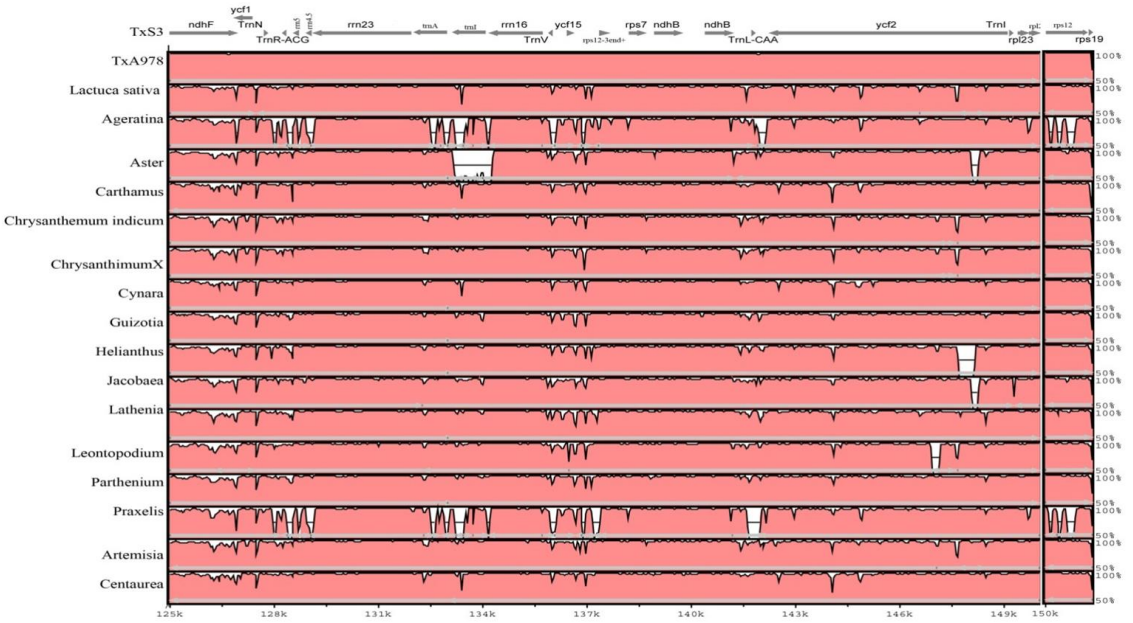
Supplementary Figure 4.2 continue…



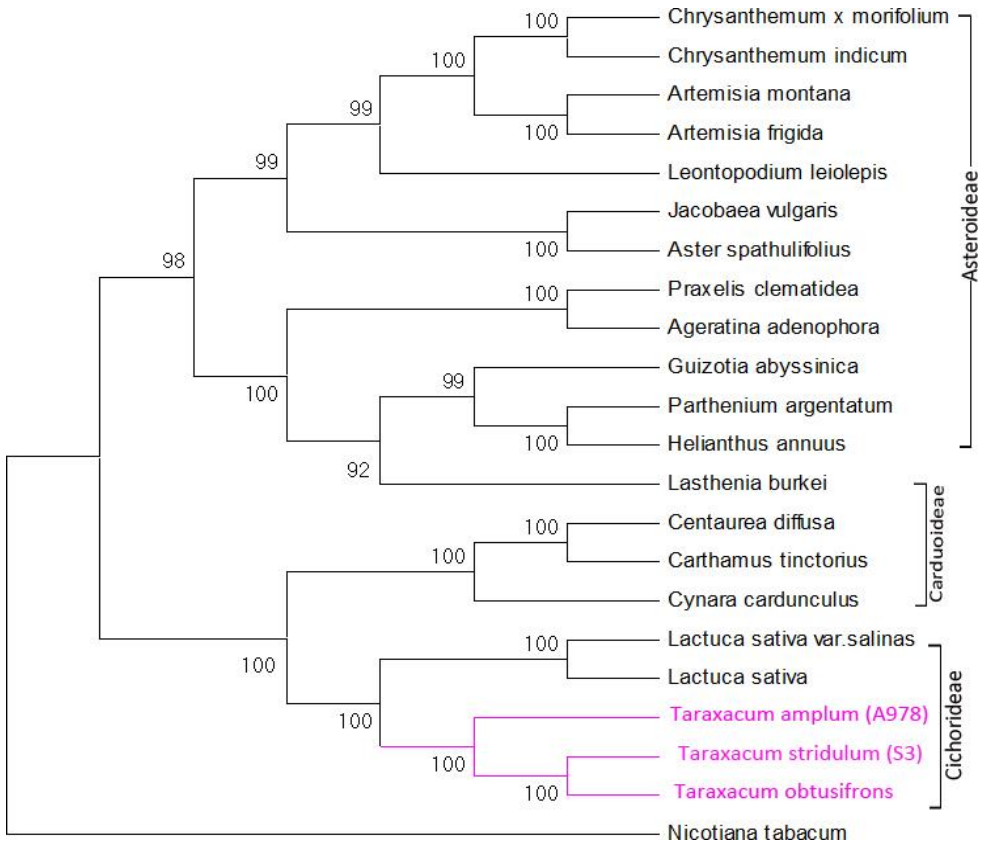


Supplementary Figure 4.2 continue…

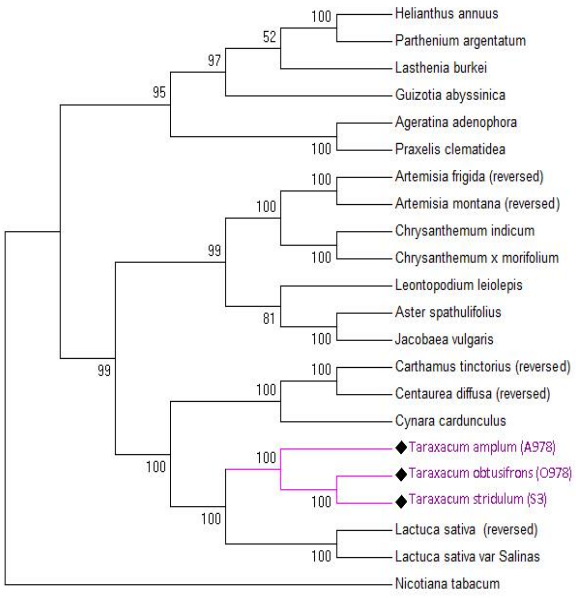




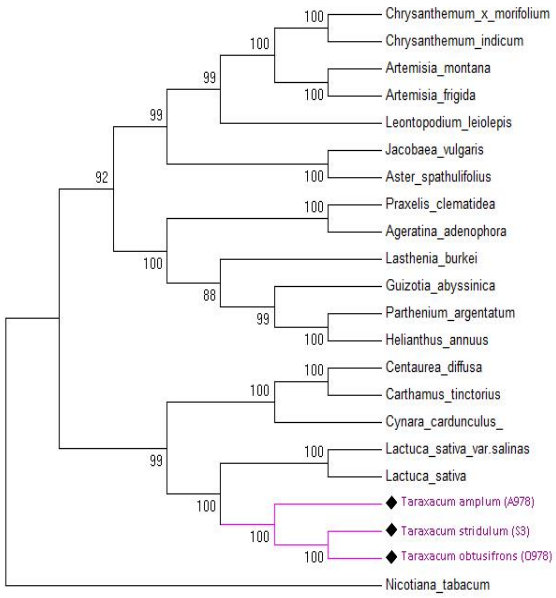
Supplementary Figure 4.3. Phylogenetic trees derived from maximum likelihood analysis of alignments of DNA sequences of 21 different Asteraceae species of a total of whole plastome and 40 different chloroplast regions indicated below the trees. Numbers above node are bootstrap support values.



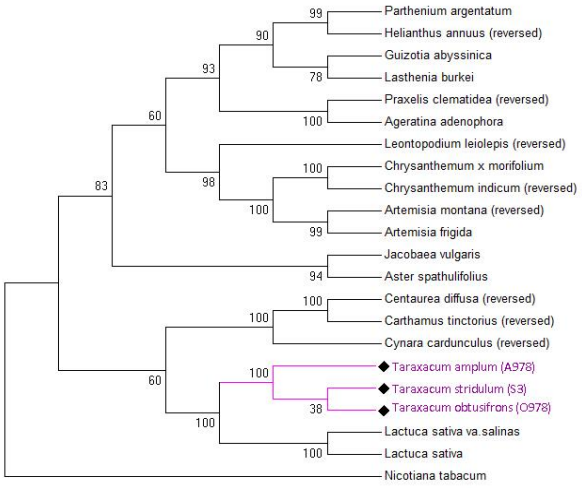
Whole chloroplast genome



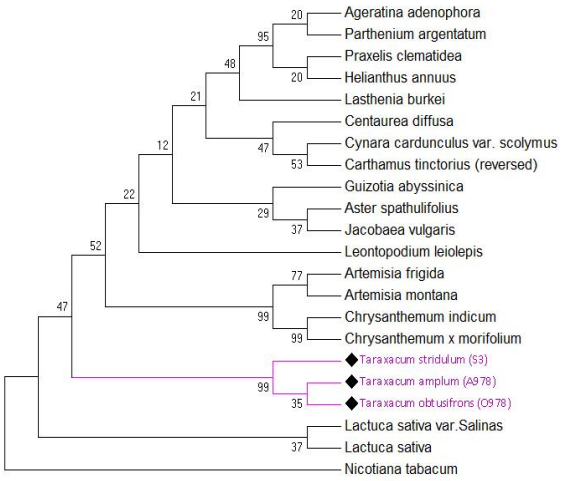
SSC region



LSC region

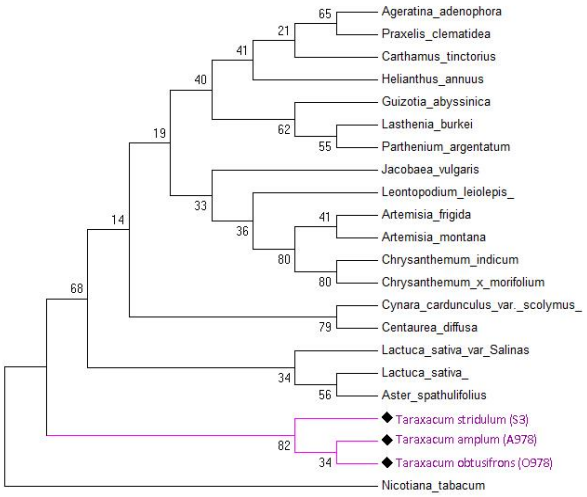


IR region

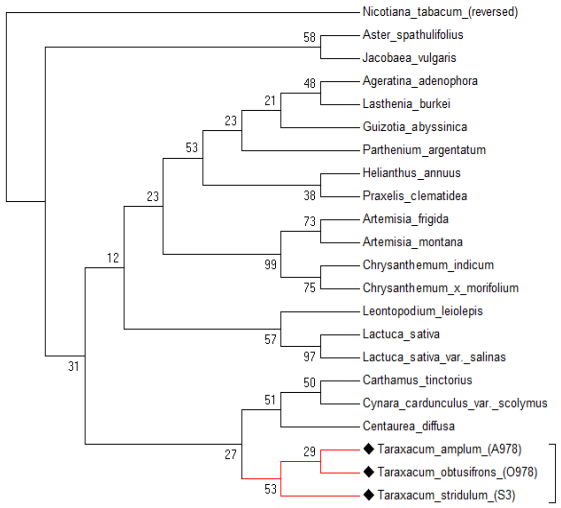


rRNA regions

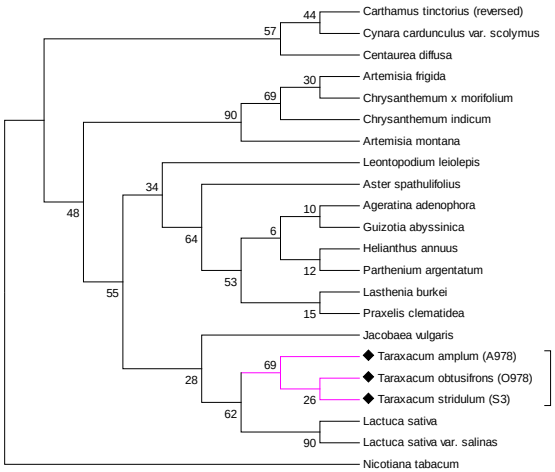
Non-coding region



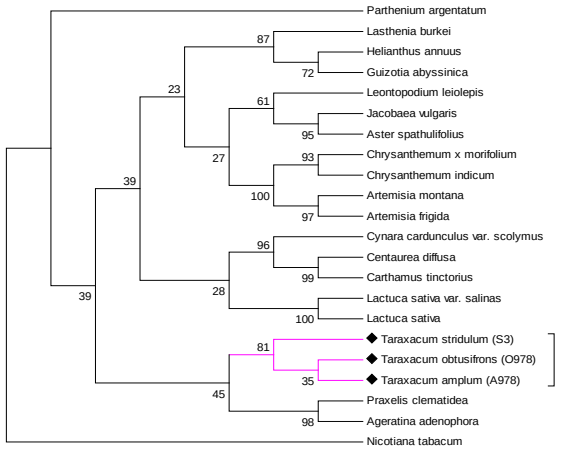
tRNA



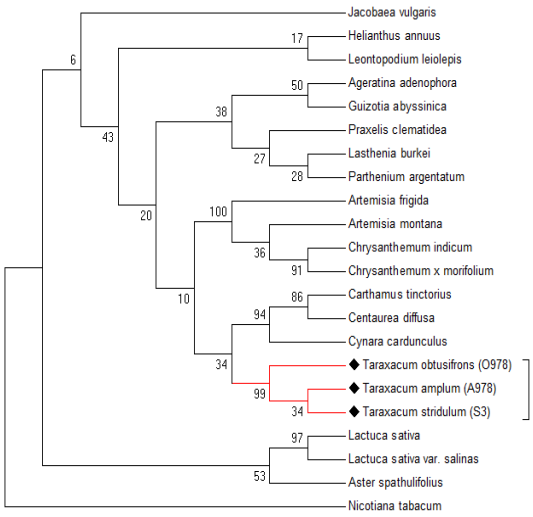
*rpo*C2-*rps*2



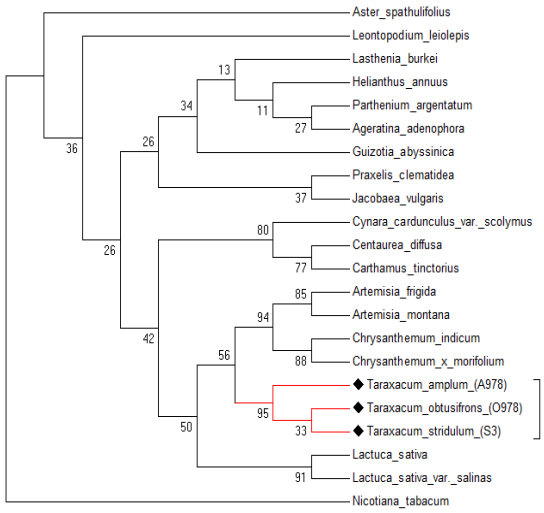
*trn*N-*ycf*1



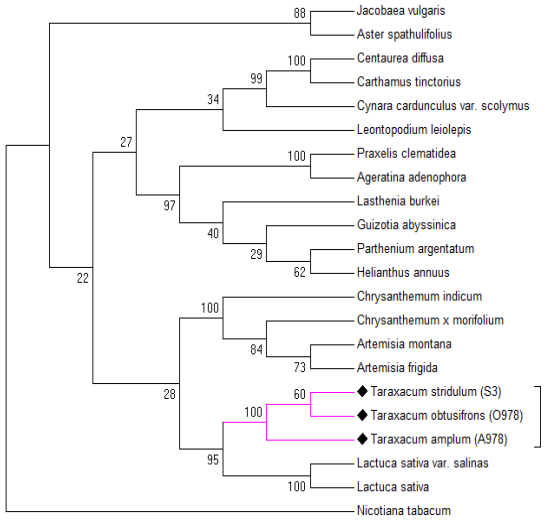
*trn*S-*trn*C



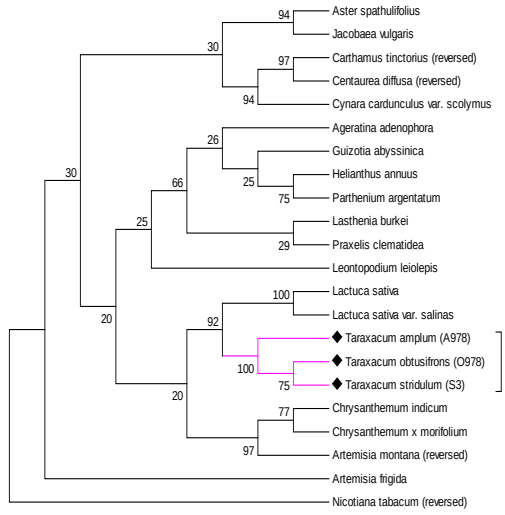
*trn*H-*psb*A



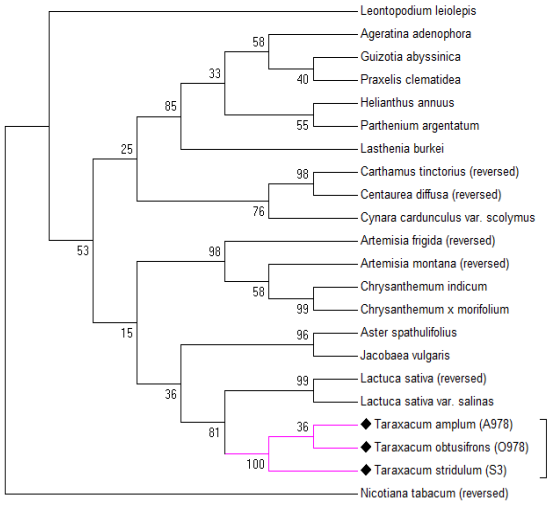
*trn*G-*trnf*M



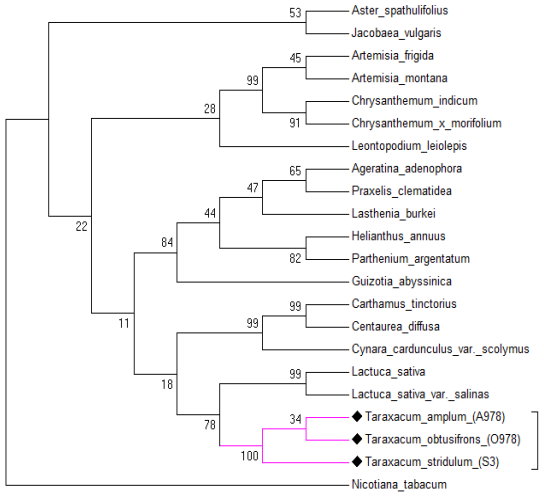
*ycf*3-*trn*S



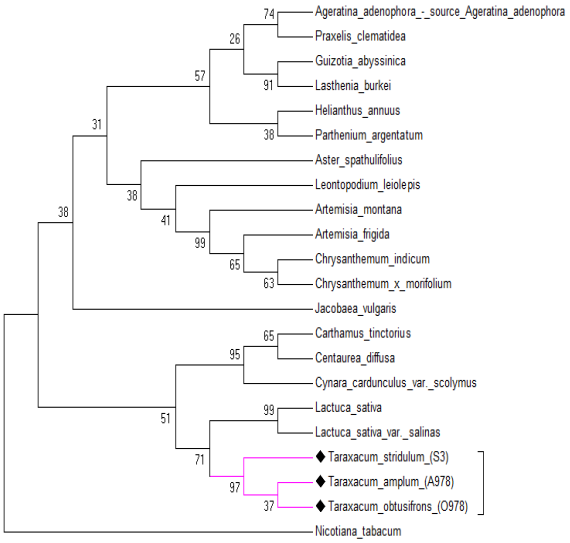
*ycf*1-*rps*15



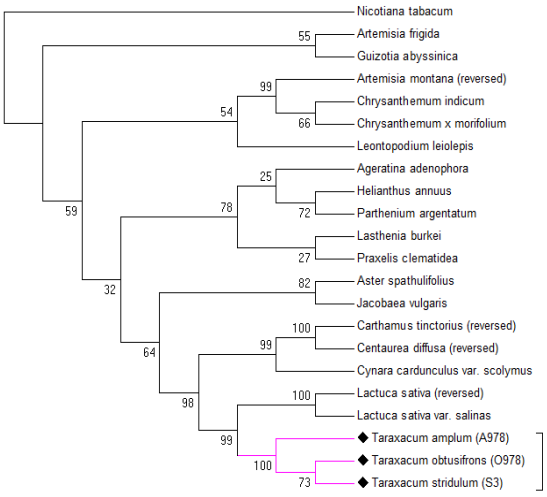
*ndh*I-*ndh*G



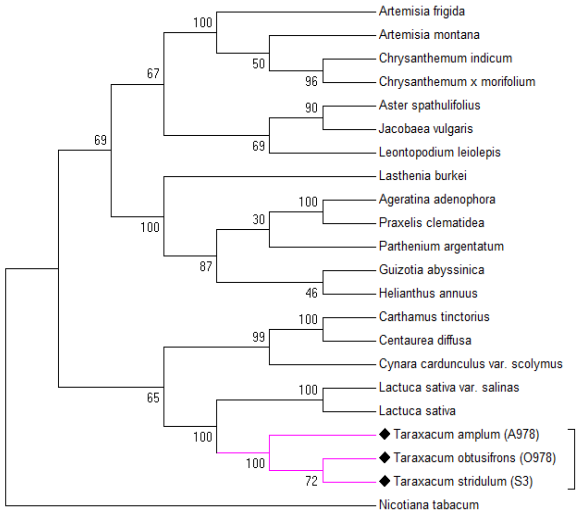
*trn*T-*trn*L



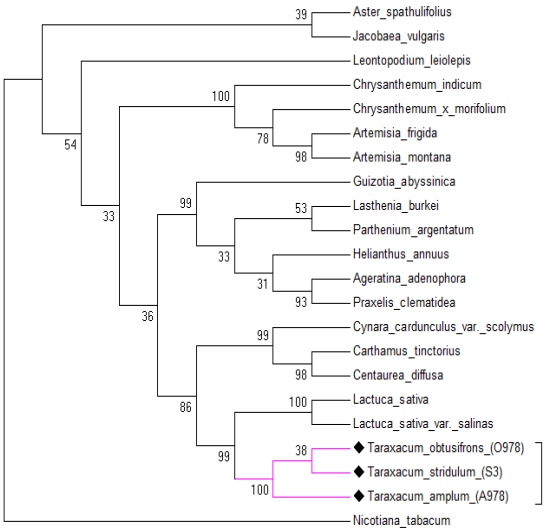
*psb*I-*trn*S



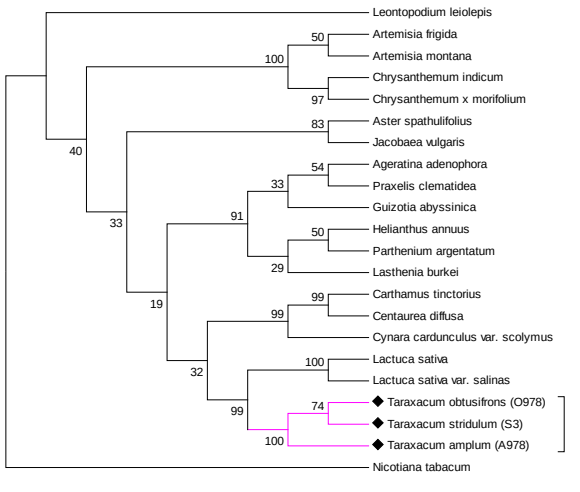
*rpl*32-*ndh*F



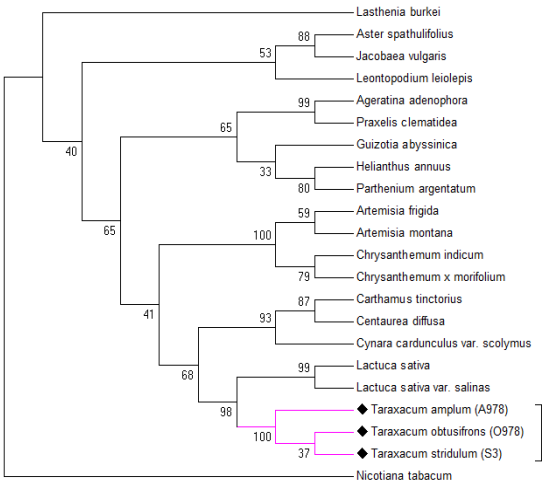
*ndh*C-*trn*V



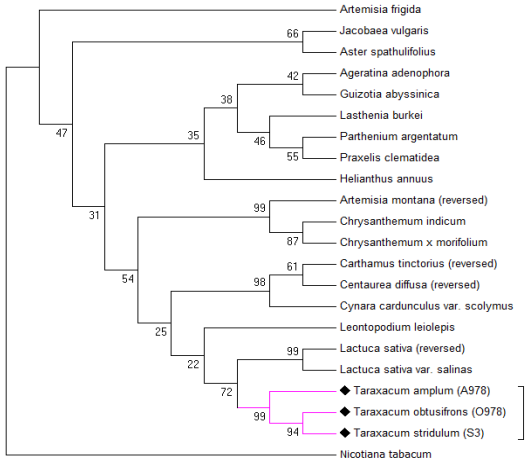
*rps*16-*trn*Q



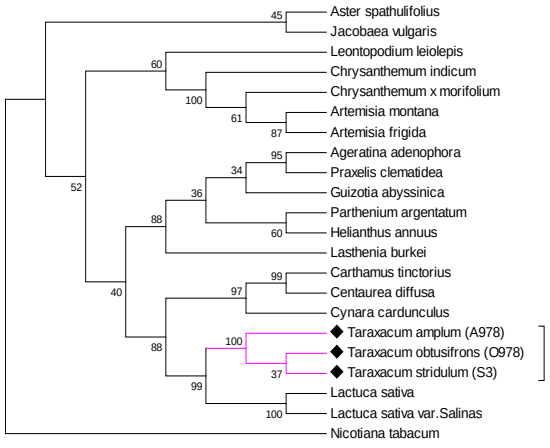
*trn*K-*rps*16



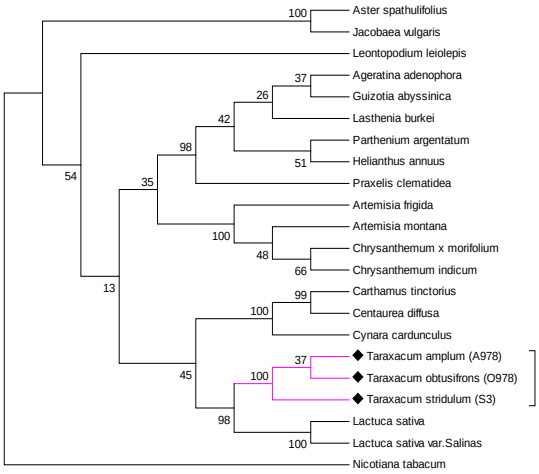
*trn*C-*pet*N



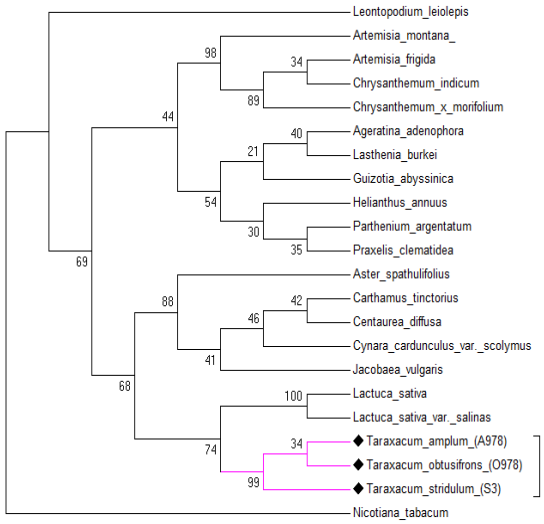
*ndh*D-*ccs*A



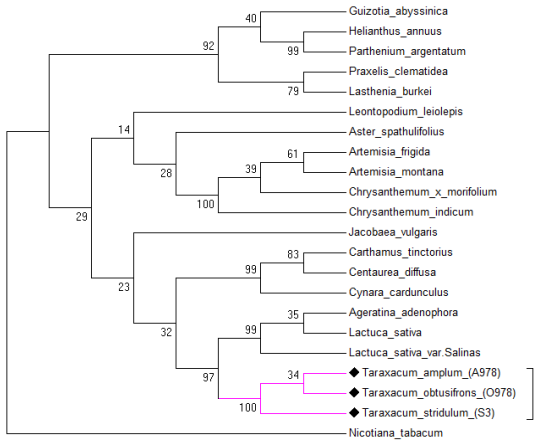
5’*rps*16-*trn*Q



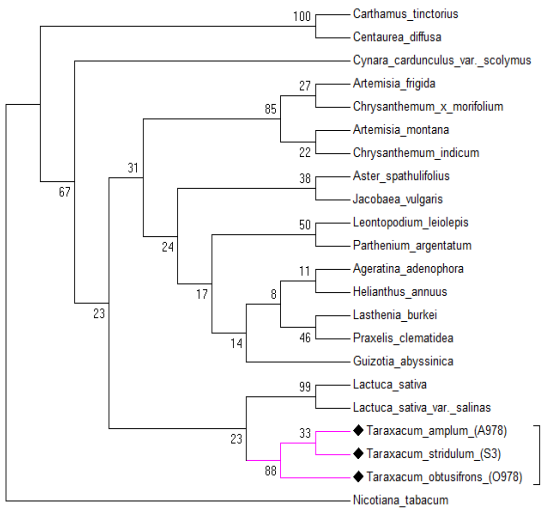
*trn*T-*psb*D



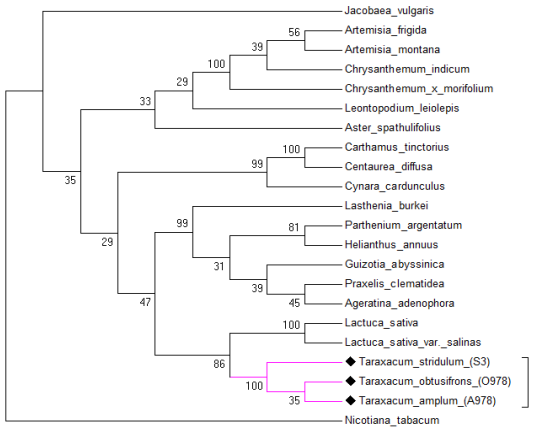
*trn*M-*atp*E



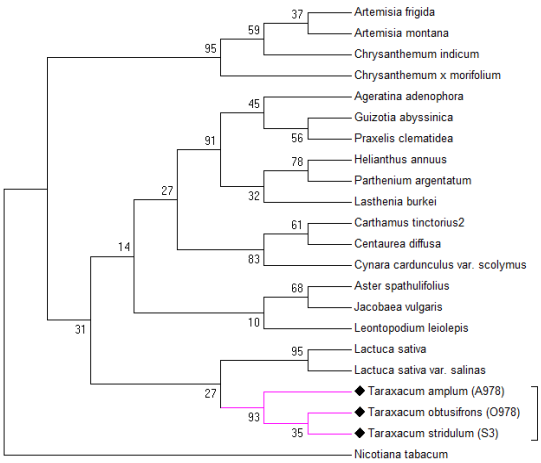
*pet*A-*psb*J



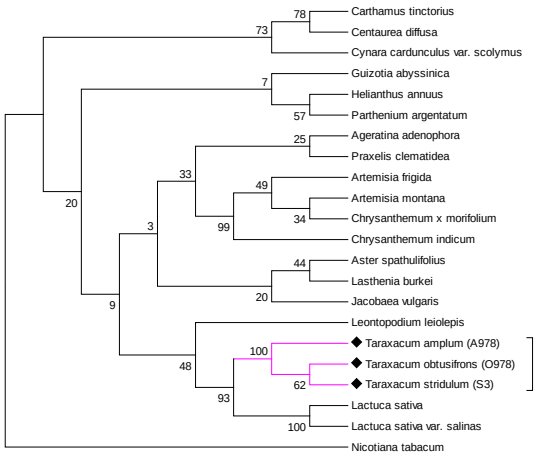
*pet*B-*pet*D



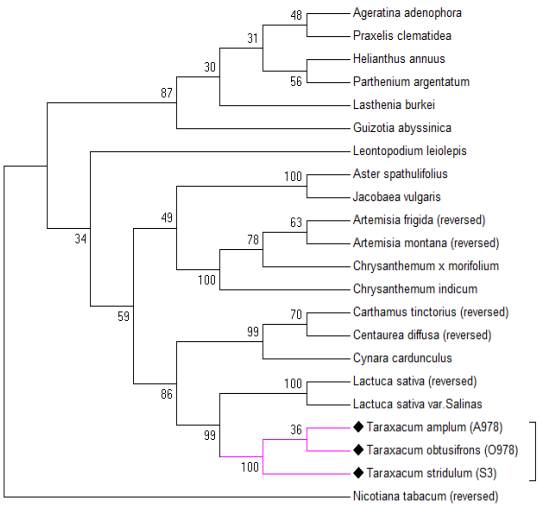
*rps*16-intron



*psb*Z-*trn*G

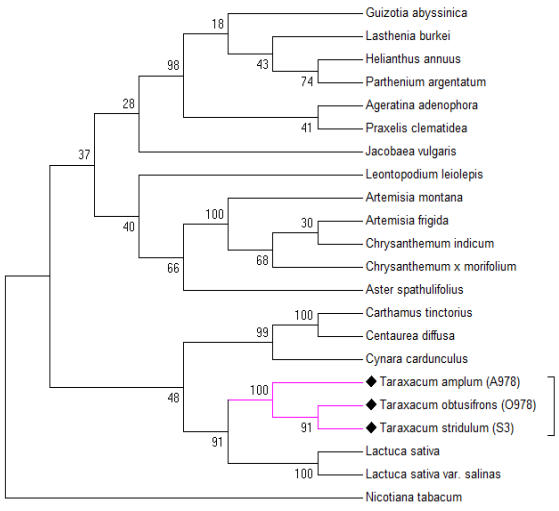


*trn*F-*ndh*J

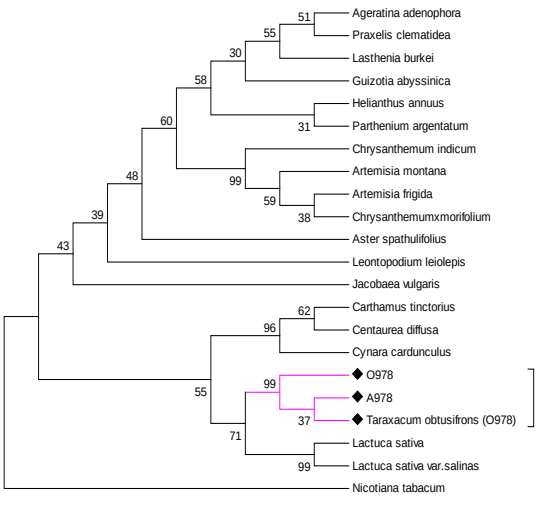


*rpl*32-*trn*L-UAG

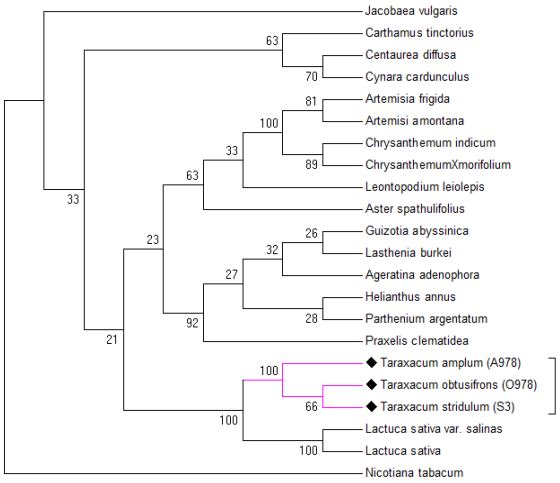
Coding region



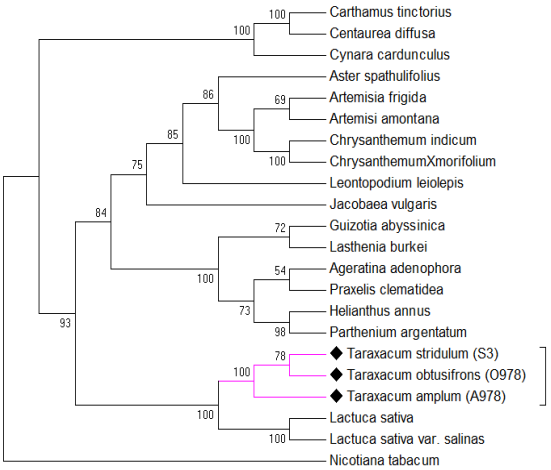
*psb*B



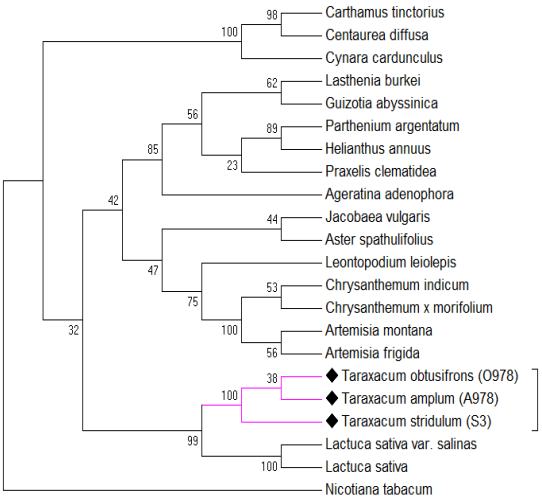
*trn*L



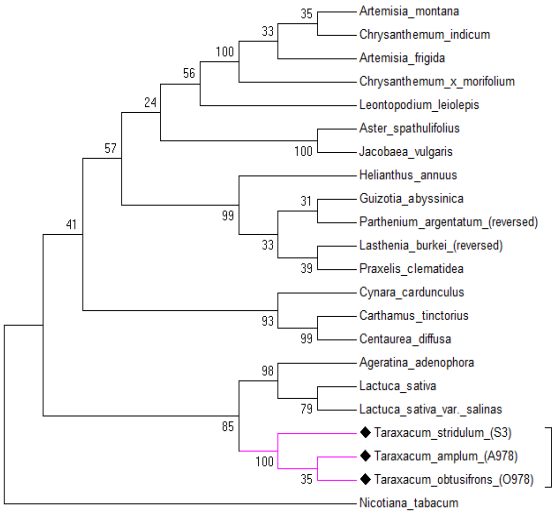
*acc*D



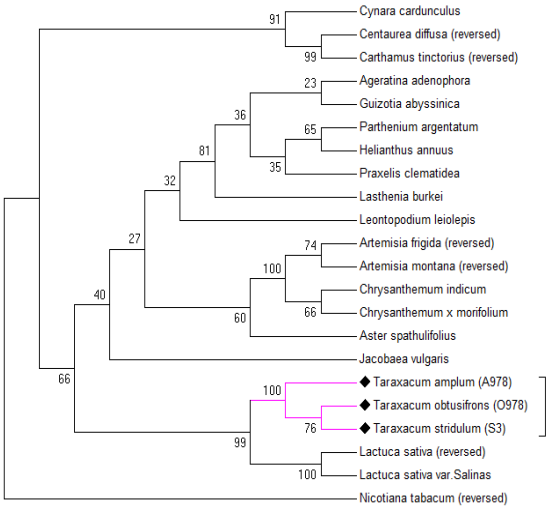
*rpo*C2



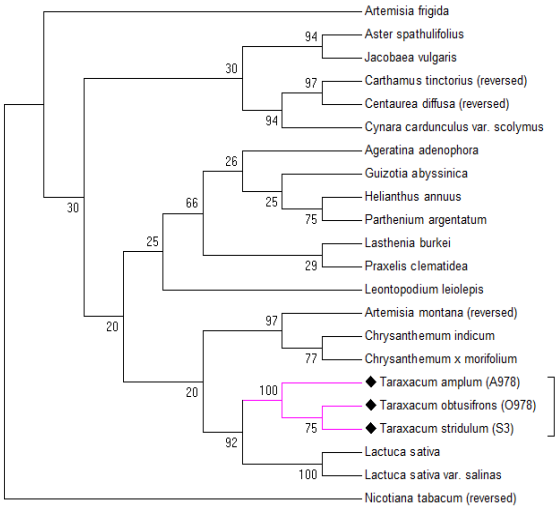
*rpo*C1



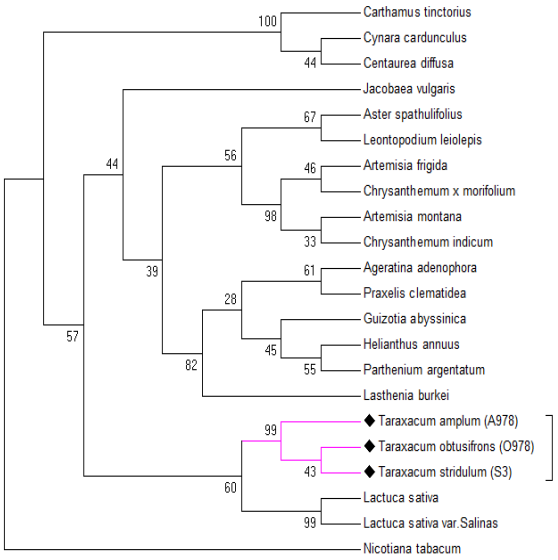
*clp*P



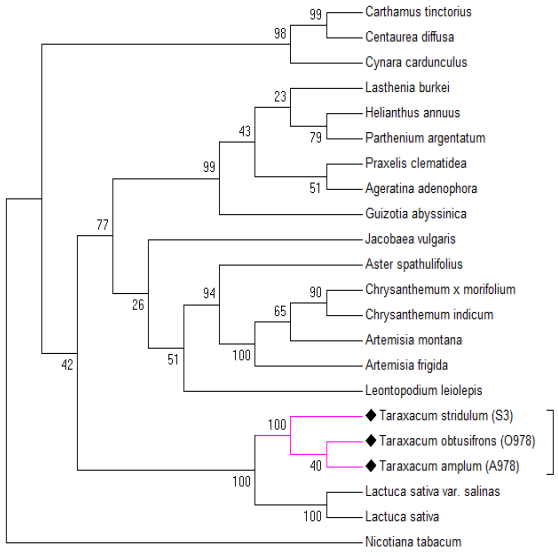
*ndh*D



*ycf*1

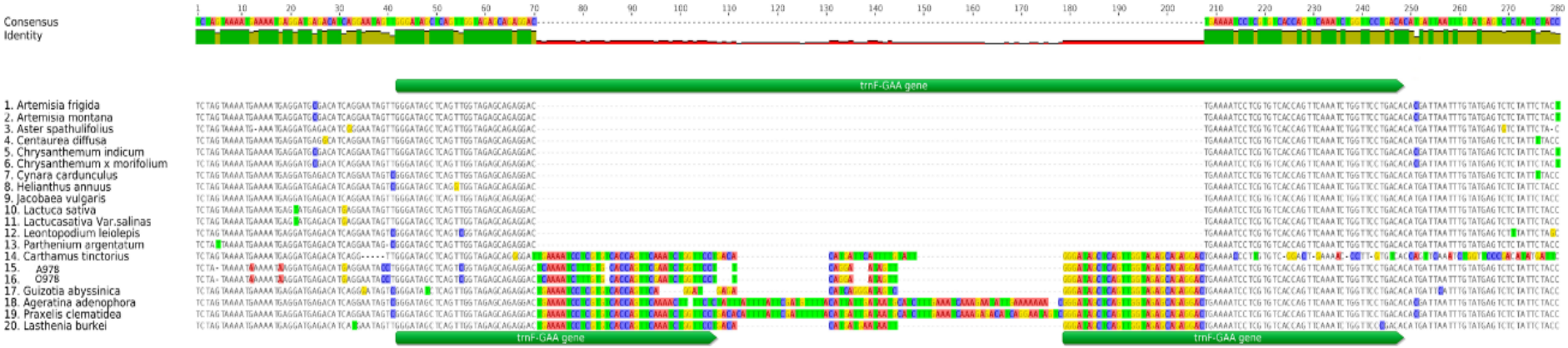

Carthamus tinctorius
Cynara cardunculus
Centaurea diffusa
Jacobaea vulgaris
Aster spathulifolius
Leontopodium leiolepis
Artemisia frigida
Chrysanthemum x morifolium
Artemisia montana
Chrysanthemum indicum
Ageratina adenophora
Praxelis clematidea
Guizotia abyssinica
Helianthus annuus
Parthenium argentatum
Lasthenia burkei
◆ Taraxacum amplum (A978)
◆ Taraxacum obtusifrons (O978)
◆ Taraxacum stridulum (S3)
Lactuca sativa
Lactuca sativa var.Salinas
Nicotiana tabacum

*rpl*22


Carthamus tinctorius
Centaurea diffusa
Cynara cardunculus
Lasthenia burkei
Helianthus annuus
Parthenium argentatum
Praxelis clematidea
Ageratina adenophora
Guizotia abyssinica
Jacobaea vulgaris
Aster spathulifolius
Chrysanthemum x morifolium
Chrysanthemum indicum
Artemisia montana
Artemisia frigida
Leontopodium leiolepis
◆ Taraxacum stridulum (S3)
◆ Taraxacum obtusifrons (O978)
◆ Taraxacum amplum (A978)
Lactuca sativa var. salinas
Lactuca sativa
Nicotiana tabacum

*ndh*F

Supplementary Figure 4.4. Alignment of *trn*F-GAA sequence of investigated Asteraceae.



In the CD ….

Supplementary Table 4.1. Maximum Likelihood fits of 24 different nucleotide substitution models for 22 accessions using the whole chloroplast genome plus 40 genic and inter-genic regions. Evolutionary analyses were conducted in MEGA6 [47]. Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best [1], and were used for the trees in S3 Fig. As noted in MEGA6, "non-uniformity of evolutionary rates among sites may be modelled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+I). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown. For estimating ML values, a tree topology was automatically computed. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position." There were a total of 136267 positions in the whole genome dataset, and the number of positions in the separate alignments for each region is shown (total number of positions in the dataset). Abbreviations: GTR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor.

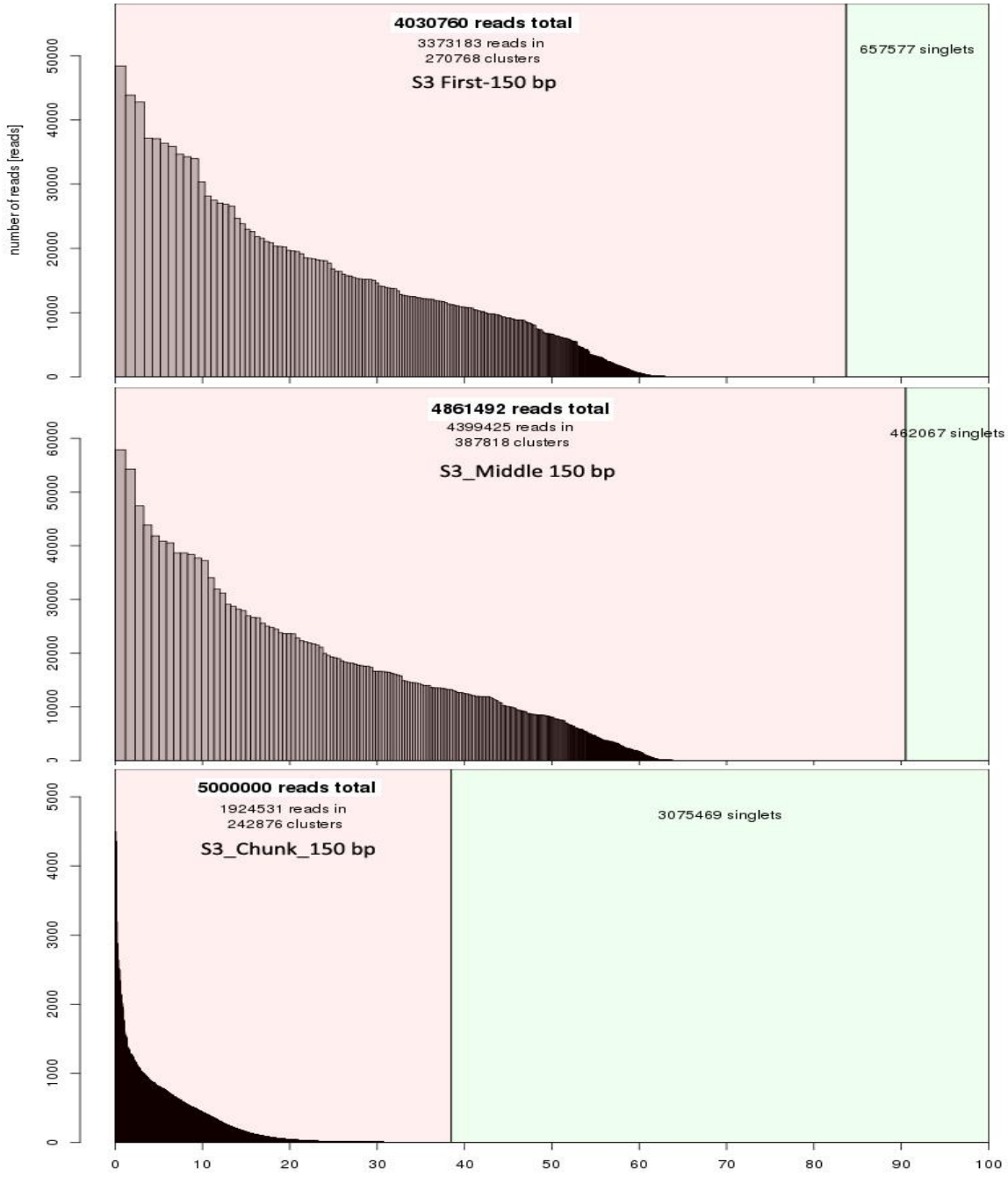
Supplementary Table 4.2. Characteristics of plastomes of 21 different accessions of 16 Asteraceae genera.

Supplementary Table 4.3. Repetitive motif abundance in *Taraxacum* and *Lactuca* plastomes computed by Reputer and Tandem Repeat Finder.

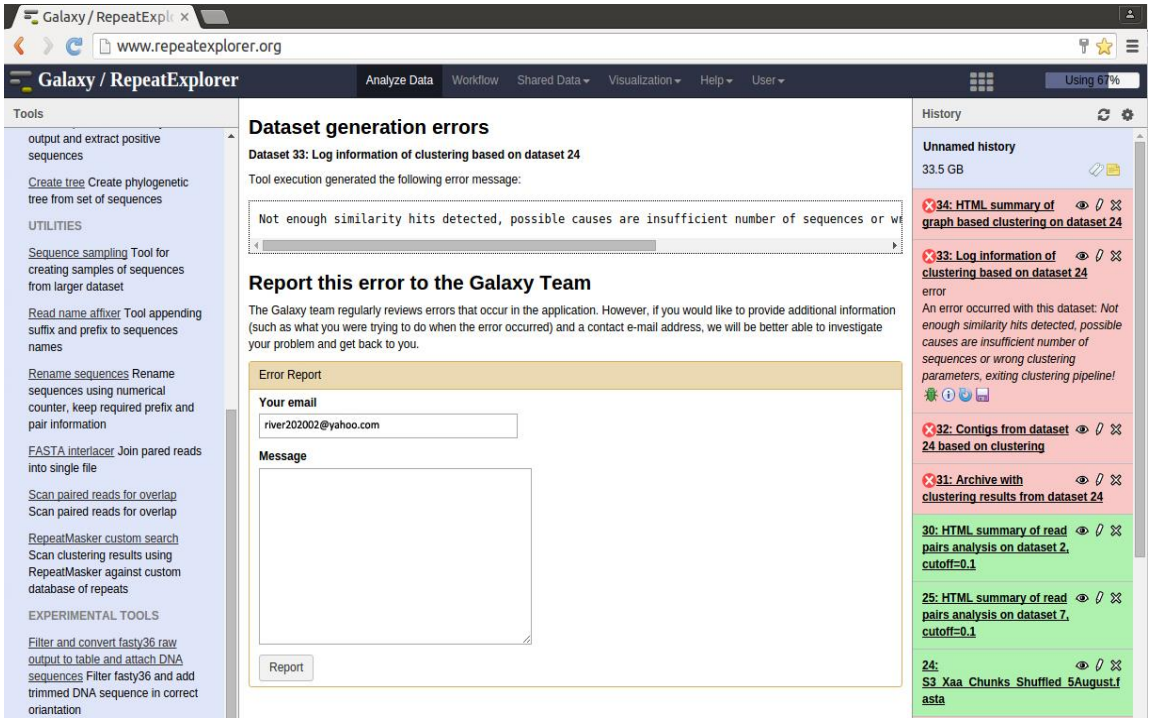
Supplementary Table 4.4. Codon usage and codon-anticodon recognition pattern of the 21 Asteraceae plastomes calculated by http://www.bioinformatics.org/sms2/codon_usage.html. Absolute numbers and values recalculated as per mille (1/1000) and proportion are shown with a heat map gives relative usage of each codon.

9.3 Appendices from Chapter 5

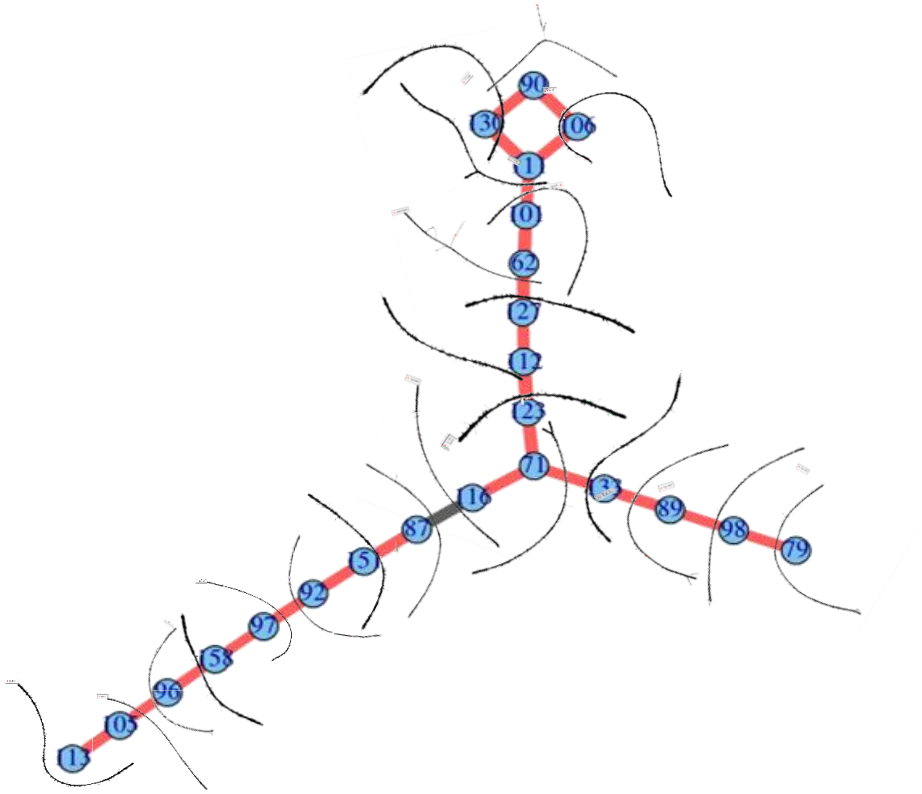
Supplementary Figure 5.1. Distribution of clusters based on *Taraxacum* S3 reads of (first 150bp, Middle 150 bp, chunk part to form 150bp), by size and class of repetitive element histograms, shows the result of clustered using RepeatExplorer. Each bar in the histograms shows the individual size (height) of each cluster and the size relative to the sampled genome (width). The Y-axis shows both the percentage of the reads and number of reads in the clusters and the X-axis shows their cumulative content. Moreover, single-copy and unclustered sequences are reflected to the right of the vertical bar. Bars are coloured according to the type of repeat present in the cluster, as determined by the similarity search.



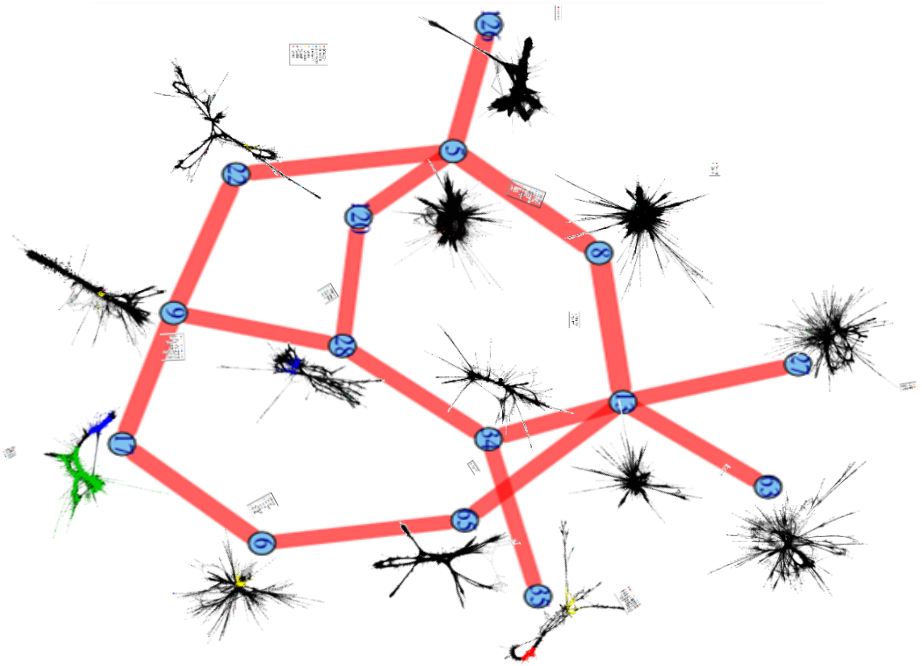
Supplementary Figure 5.2. Show the error message of uploaded randomized sequences by RepeatExplorer.



Supplementary Figure 5.3. Shows the graph of grouped RepeatExplorer clusters by cut-off value 0.1. Only groups with more than two clusters are shown graph. Connection through mates is labelled by red if similarity hits exist between clusters, otherwise connection is shown as grey.

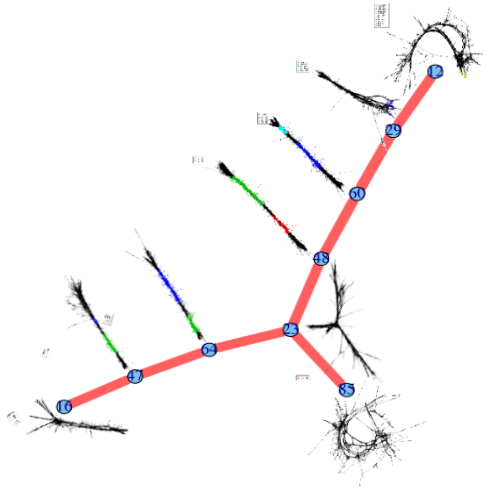


Group 1 – Low_complexity

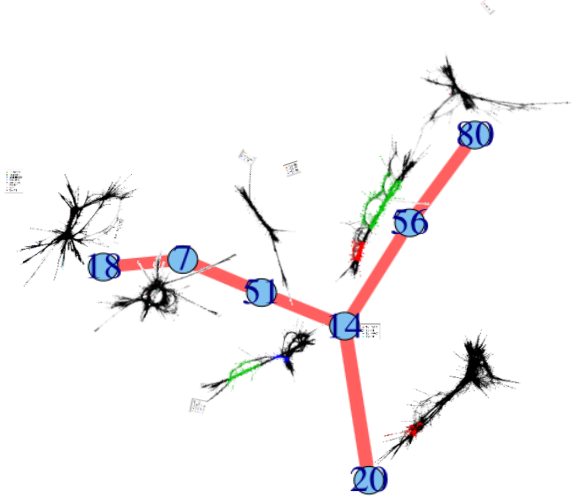


Group 2 – gypsy/chromovirus

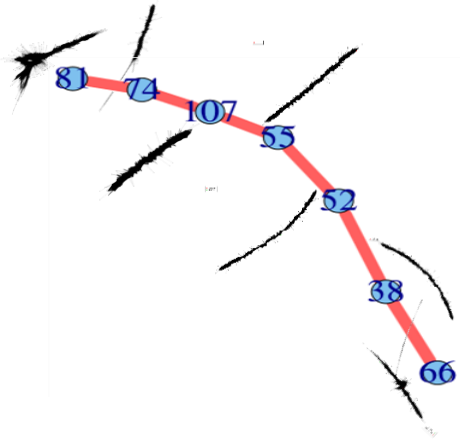
Supplementary Figure 5.3 continue ….



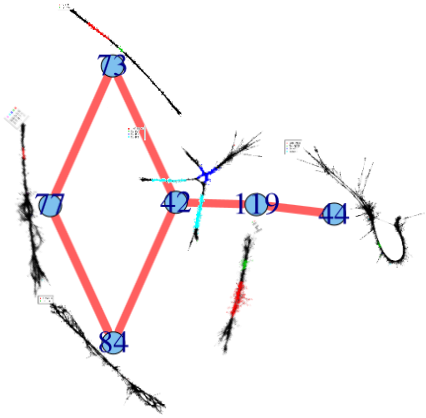
Group 3 – copia/maximum-SIRE



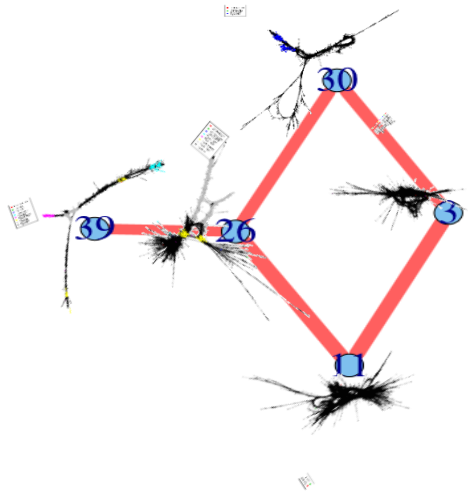
Group 4 - copia/ maximum-SIRE



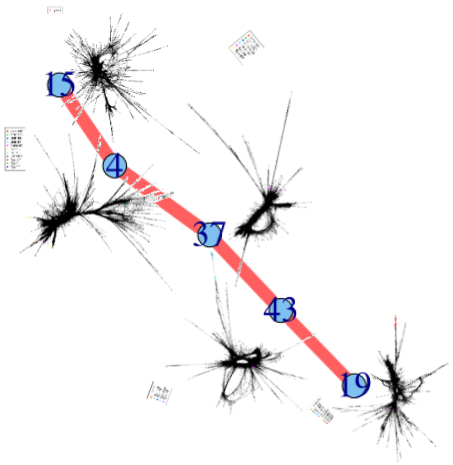
Group 5 - rRNA/45S



Group 6 - copia/SIRE

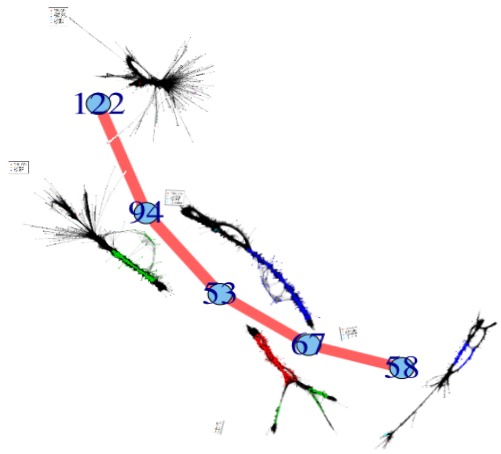


Group 7 - gypsy/chromovirus

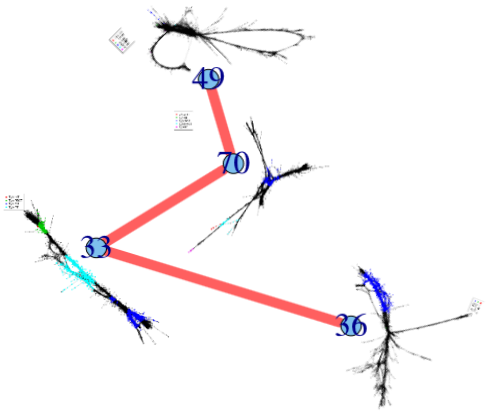


Group 8 - Low_complexity

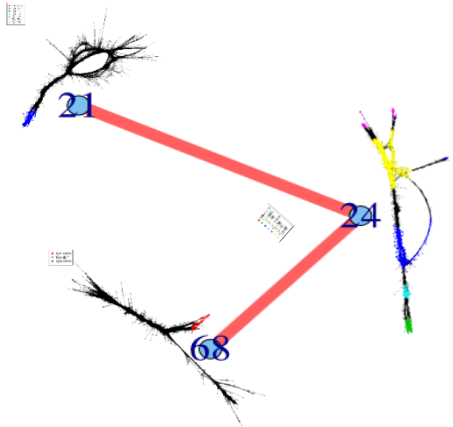
Supplementary Figure 5.3 continue …



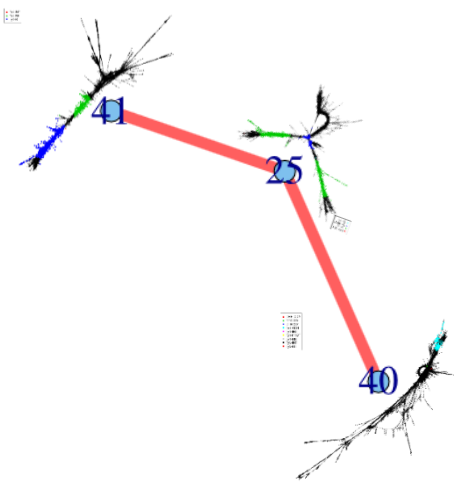
Group 9 - copia/Bianca



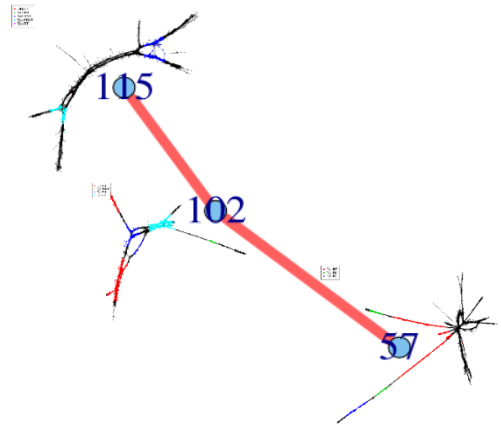
Group 10 - gypsy/Tat



Group 11 – copia/Angela

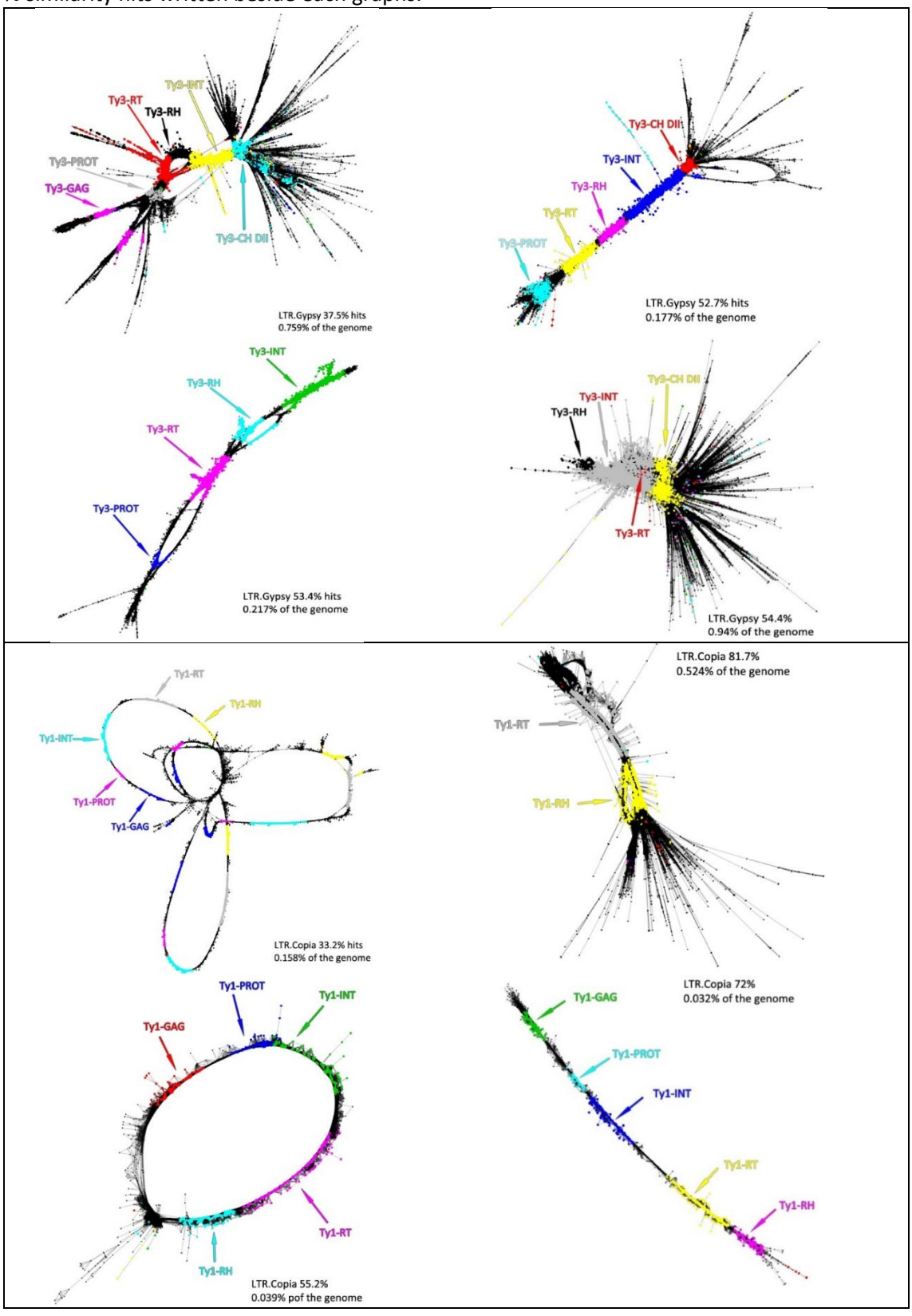


Group 12 – copia/SIRE

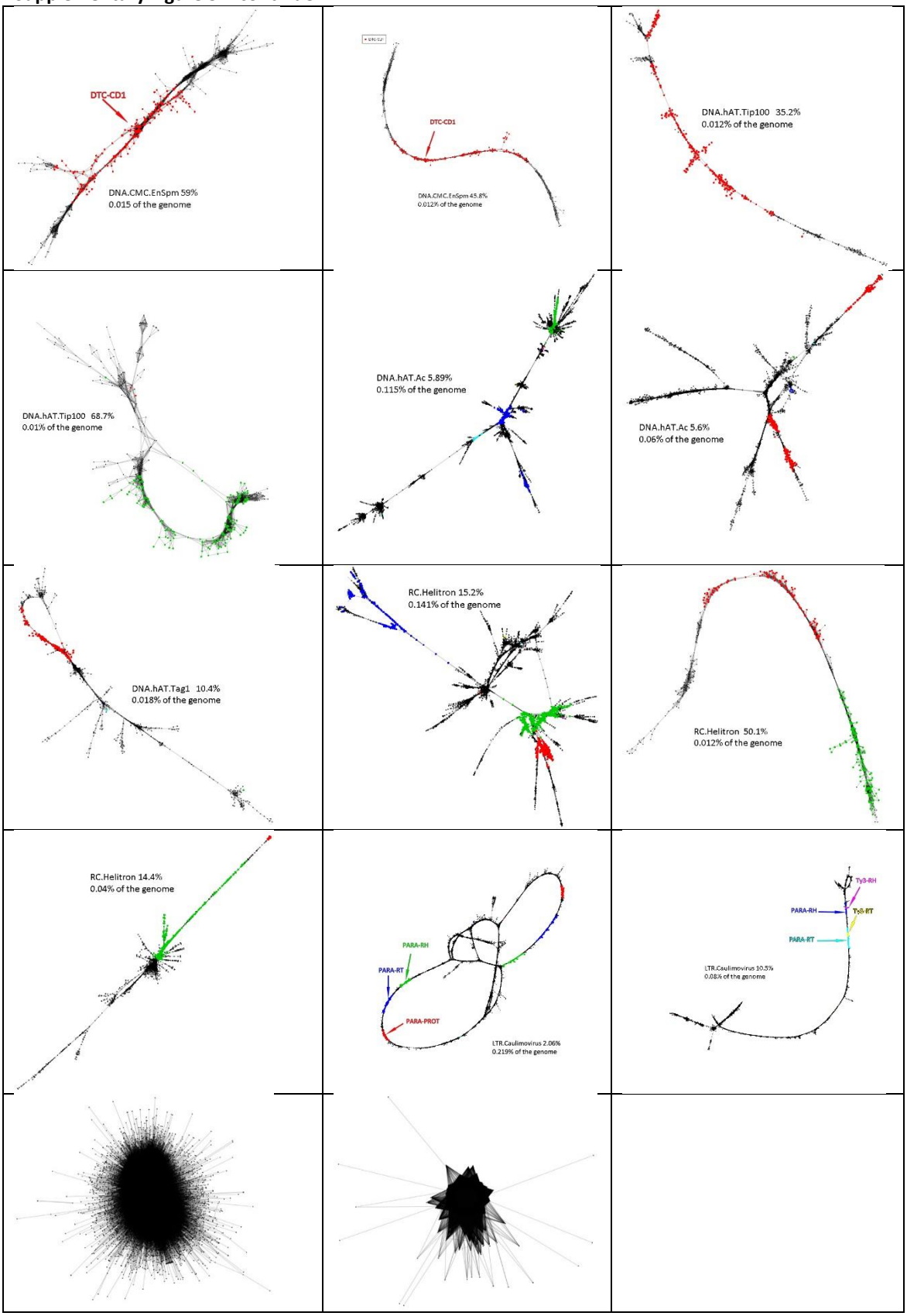


Group 13 - gypsy/Athila

Supplementary Figure 5.4. Show the pattern shape of different repetitive DNA clusters from three *Taraxacum* microspecies. Repetitive domain have been characterized and labelled on each graphs and name of repetitive DNA represented by the graph, genomic proportions, and % similarity hits written beside each graphs.



Supplementary Figure 5.4 continue…



In the CD ….

Supplementary Table 5.1. The analysing data statistic to calculate percentage of mono-, di-, and Tri-nucleotide sequences comparisons between different clusters of repetitive DNA from RepeatExplorer outcomes.

Supplementary Table 5.2. Numerical data for Figure 5.7, show the coverage of each probe used in *in situ* hybridizations.