Research Networking Programmes

Short Visit Grant □ or Exchange Visit Grant ☒

(please tick the relevant box)

Scientific Report

Scientific report (one single document in WORD or PDF file) should be submitted online within one month of the event. It should not exceed eight A4 pages.

Proposal Title: Mining molecular markers from EST data bases to study threatened plants.

Application Reference N°: 4379

1) Purpose of the visit

As stated in the project proposal the aims of the visit were: i) to mine EST databases in the search of SSR motifs to develop EST-SSR for all the genera of threatened plants listed by IUCN; ii) to identify gaps in the available information (i.e. taxa that require further research); and iii) to empirically test the transferability and variability of EST-SSRs markers between species within one of the selected genus.

2) Description of the work carried out during the visit

All EST sequences available in GenBank for those genera included in the red list of the International Union for Conservation of Nature and Natural Resources (IUCN) were downloaded and mined for SSR motifs. Each fasta file containing EST sequences was analyzed using iQDD software with the following settings: sequences larger than 100bp were mined for dimer, trimer, tetramer, pentamer and hexamer motifs with a minimum length of 20 bp. Primers were searched using default settings, i. e. an optimal length of 20 bp, Tm 60ºC and designed for SSR with a PCR product between 90-320 bp. Results were summarized regarding the number of available EST sequences, number of SSR found, repetition length and type of motif. Moreover, these results were sorted in seven broad taxonomic groups, florideophyceae, cariophyceae, ferns, lycopodiaceae, gymnosperms, mocots and dicots in order to provide more detailed insights for each particular group. Finally, a list including all the genera successfully scanned for SSRcontaining information of total number of SSR found, as well as number of dimers, trimer, tetramers, pentamers and hexamers, will be provided as supplementary material in the manuscript. This way, every researcher interested in a specific genera would have access to the EST-SSR primers information upon request.

In comparison with classical microsatellites, EST-SSR markers are more likely to be related with genetic regions responsible for quantitative variation (i. e. coding regions). Therefore, since natural selection acts directly on phenotypes rather than on genotypes, EST-SRR can act as functional markers with high relevance in conservation studies. Unfortunately, most threatened taxa are non-model organism and well annotated genomes are not available for comparison, preventing the determination of the location of the EST-SSRs within the genome (i.e. intergenic regions, introns, UTRs or exons). Aiming to minimize this impediment, the EST sequences data set of two genera with well-known annotated genomes were analized and used as a control for the IUCN red list genera. Namely, the genera Arabidopsis was selected as a control for dicots and Oryza was used as a standard for monocots. EST sequences for both genera were downloaded and searched for SSR with iQDD following the same criterion as for the genera included in the IUCN red list. Afterwards, iQDD output files were used as input for the blast analysis using the default parameters specified in the NCBI website. In those cases where a positive hit was found (i.e. at least 98% of coincidence with the NCBI reference database for that genera) the coincident sequences were downloaded and aligned in Geneious 6.1.6 with the iQDD output sequences. Using the gene information from the BLAST search, the positions of the SSR and its primers were determined (i.e. located in UTRs, exons or in a non-coding regions (intergenic region, intron).

EST-SSR have proved to be a valuable tool in plant studies (Wöhrtmann et al. 2011, Tabbasam et al. 2013). Most of these studies have target particular species, mainly those with economic relevance (Varshneya et al. 2005, Simko 2009, Mishra et al. 2011). Therefore, the present project represents the first attempt to standardize EST-SSR as suitable markers in conservation studies of threatened non-model species. Besides, it is also important to highlight that these project encompasses a vast number of genera. Even if multiple studies have already showed the empirical performance of these markers, three genera from the IUCN red list were selected to test the designed primers. These genera (Trifolium, Centaurea and Allium) comprehended four species (T. angulosum, T. saxatile, C. valesiaca and A. angulosum) with ten individuals per species. A selection of 12 EST-SSRs markers for each genera were tested for amplification using the M13 tail method developed by Schuelke (2000). For those EST-SSR that successfully amplified, the PCR product was further analyze for polimorphism. Finally, transferability was also examined between two species of the same genera (T. angulosum and T. saxatile).

Bibliography:

- Mishra, R K; Gangadhar, B H; Yu, J W; Kim, D H; Park, S W (2011). Development and Characterization of EST Based SSR Markers in Madagascar periwinkle (Catharanthus roseus) and Their Transferability in Other Medicinal Plants. Plant Omics Journal 4 (3): 154-162.
- Simko, I (2009). Development of EST-SSR markers for the study of population strucuture in Lettuce (Lactuca sativa L.). Journal of Heredity. 100(2): 256-262.
- Schuelke, M. (2000). An economic method for the fluorescent labelling of PCR fragments. Nature Biotechnology. 18, 233 - 234.
- Tabbasam, N; Zafar, Y and Mehboob-ur-Rahman (2013) Pros and cons of using genomic SSRs and EST-SSRs for resolving phylogeny of the genus Gossypum. Plant Systematics and Evolution. DOI 10.1007/s00606-013-0891-x.

- Varshneya, R S; Sigmunda, R; Börnera, A; Korzunb, V; Steina, N; Sorrellsc, M E; Langridged, P and Graner, A (2005). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant science 168: 195-202.
- Wöhrmann T and Weising K (2011). In silico mining for simple sequence repeat loci in a pineapple expressed sequence tag database and cross-species amplification of EST-SSR markers across Bromeliaceae. Theoretical and Applied Genetics. 123(4):635-47.

3) Description of the main results obtained

The EST collection from the two control genomes (i.e. Oryza and Arabidopsis) were screened for perfect SSR using iQDD. The Oryza data set was originally constituted by 1342281 EST sequences. However, after excluding sequences shorter than 100bp and redundancy only 3030 EST sequences remained for the primer search and 581 SSR were retrieved from their analysis. The Arabidopsis data set encompassed 1532829 EST sequences. After excluding sequences shorter than 100bp and redundancy only 899 EST sequences remained for the SSR primer search. Within these 899 EST sequences, 151 SSR were retrieved. Despite the larger number of EST sequences available for Arabidopsis, Oryza produced a much larger number of SSRs. This is explained by the short length of many of the Arabidopsis EST sequences and by the high levels of redundancy, in fact, previous studies that evaluated the EST data set of both species found a lower number of SSR for Arabidopsis than in Oryza (Victoria et al. 2011).

As mononucleotide repeat polymorphisms are difficult to interpret, only di-, tri-, tetra-, penta- and hexanucleotides were considered as potential candidates for EST-SSR (Table 1 and 2). For both genera trimers were the most common motif, encompassing more than half of the SSR; dimers were also abundant while tetra and pentamers were scarce in both genera (Table 1 and Table 2) (Gao et al. 2003, Kantety et al. 2002, Poncet et al 2006). Both trimers and dimers are known to be favoured in higher plants in comparison with algae or mosses (Victoria et al. 2011). As expected in vascular plants, the most abundant dimer motif was AG/CT (Fig. 1 and Fig. 3) (Victoria et al. 2011; Morgante et al. 2002; Kantety et al. 2002; Temnykh et al. 1999) while, low frequencies of AT were observed in both genera (Morgante et al. 2002; Kantety et al. 2002). Trimer repeats differed between genera. The trimer repeat CCG was the most abundant in Oryza, followed by other GC-rich repeats as expected in mocots (Fig. 2) (Gao et al. 2003, Kantety et al. 2002, Victoria et al. 2011). In comparison, AAG was the most common repeat in the Arabidopsis EST-SSR collection and in general, the GC-rich motifs were not dominant (Fig. 4) (Victoria et al. 2011). Regarding the distribution within the genome, di, tetra and pentamers were mainly located within non-coding regions and UTRs, while tri and hexamers were more common in coding regions (Table 1 and Table 2). The high abundance of dimers in non-coding regions and in UTRs, as well as the predominance of trimers in conding regions, has been previously reported (Gao et al. 2003; Wang et al 1994). More than 75% of the CCG SSR found in Oryza were related to coding regions. Interestedly, CCG repeats have been found to be involved in many gene functions as stress resistance, transcription regulation, metabolic enzyme biosynthesis, and so on (Gao et al. 2003). Finally, tetramers and pentamers were underrepresented in both genera and were located mainly in non-coding regions and UTRs. Hexamer motifs displayed an intermediated frequency and they were manly located in exons. These results suggest that for those studies targeting molecular markers that migth be under selective pressure, trimers would be the best option because they are mainly located within exons and they are more polymorphic than hexamers. However, this does not mean that the remaining markers should be disregarded. In fact, markers located in UTRs migth also have an evolutionary role because UTRs have a regulatory function in gene expression. Eventually, the EST-SSR located in non-coding regions are likely to be more polymorphic than those located in exons and UTRs and they would behave as neutral markers. Therefore, this approach can also be regarded as a cheaper and less time-consuming alternative technique for SSR development comparing with classical methods for SSR development.

The IUCN red list for plants includes 23 classes with 381 families. In the bdEST (NCBI) there are EST sequences available for 13 classes and 141 families. Within these 141 families, there are 257 plant genera with EST sequences available in the dbEST that were analyzed (Table 3). A total of 14498726 EST sequences were screened for SSR obtaining 17076 microsatellites with their respective primer pairs. More than eighty six per cent of the analyzed genera rendered any SSR. Those genera with a reduced number of sequences, or with sequences shorter than 100bp, did not provide any SSR. Because ETS-SSR are not as polymorphic as classical SSR, we set a minimum of fifty SSRs as a very conservative threshold to obtain enough polymorphism for population genetic studies. This threshold was based in results from previous studies (Chen et al 2006, Eujayl et al 2004, Wohrmann & Weising 2011). More than one third of the genera passed the threshold, meaning that a large number of EST-SSR were designed for 70 genera (Table 3). Since EST-SSR are very conservative, they have been proved to be transferable among species, this means that in 70 genera we could target a very large number of different species. Overall, seven groups were analyzed: Florideophyceae, Cariophyceae, Ferns, Lycopodiaceae, Gymnosperms, Mocots and Dicots (Fig. 5). Florideophyceae, cariophyceae, ferns and lycopodiaceae had one single genus with more than 50 SSR. Therefore, generalizations about the distribution and type of motif in these groups cannot be done without risk and further research should be done. Still, and in agreement with previous studies, gymnosperms revealed a high proportion of hexanucleotides and AT/TA were the most common motifs (Table 3) (Victoria et al. 2011). As expected, trimers were the most common motifs in monocot and dicots followed by dimers while tetramers and hexamers were scarce (Fig. 5). In higher plants (mocost and dicots) AG/GA was the most common motif (Victoria et al. 2011). However, differences between monocots and dicots were found for trimers. In monocts the CCG/GCC motifs were highly represented while AAG/GAA were more common in dicots (Victoria et al. 2011). Finally, preliminary results from the empirical test revealed that the primers designed for the EST-SSR amplified in 40-60% of the cases and the PCR product has the expected size. Besides, they were 100% transferable and 70% polymorphic between species of the same genus (e.g. Trifollium) and in 20-40% they were also polymorphic within species.

In summary, even if only one third of all the genera analyzed rendered 50 or more EST-SSR, the results for these 70 genera encompass a large number of non-model species with conservation concern. Besides, as it was said, 50 SSR is a very conservative threshold and several studies have found enough polymorphic EST-SSR testing smaller numbers. Following the results of the control genomes, studies with conservation purposes should be focused in trimers because they are highly likely to be located within exons and are more abundant and polymorphic than hexamers. Finally, the empirical essay demonstrated that they are valuable markers in studies involving species of the same genera and are also polymorphic enough to be suitable for population studies with one single species. In conclusion, the in silico mining of EST sequences for SSR development in threatened plants seems a suitable option.

Bibliography:

- Chen, C; Zhou, P; Choi, YA; Huang, S and Gmitter, FG Jr (2006). Mining and characterizing microsatellites from citrus ESTs. Theoretical and Applied Genetics. 112(7):1248-57.
- Eujayl, I; Sledge, MK; Wang, L; May, GD; Chekhovskiy, K; Zwonitzer, JC and Mian, MA (2004). Medicago truncatula EST-SSRs reveal cross-species genetic markers for Medicago spp. Theoretical and Applied Genetics. 108(3):414-22.
-Gao, L; Tnag, J; Li, H and Jia, J (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. Molecular breeding. 12(3): 245-261.
- Kantety, RV; La Rota, M; Matthews, DE and Sorrells, ME (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. PLant Molecular Biology. 48(5-6):501-10.
- Victoria, FC; da Maia, LC and Costa de Oliveira, A (2011). In silico comparative analysis of SSR markers in plants. BMC Plant Biology. 11:15
- Wang, Z; Weber, JL; Zhong, G and Tanksley, SD (1194). Survey of plant short tandem DNA repeats. Theoretical and Applied Genetics. 88(1):1-6.
- Wöhrmann, T and Weising, K (2011). In silico mining for simple sequence repeat loci in a pineapple expressed sequence tag database and cross-species amplification of EST-SSR markers across Bromeliaceae. Theoretical and Applied Genetics. 123(4):635-47.

4) Future collaboration with host institution (if applicable)

As a result from this collaboration and at this crucial stage of my career (last months as PhD student), Professor's Koch research group at Heidelberg University emerges as the ideal host institution for a postdoctoral visit. Therefore, after the visit we plan to submit postdoctoral grant applications in the near future to continue our work together.

5) Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant)*

The four months stay in the currently laboratory of Prof. Koch provided one manuscript currently under preparation that in the following three months we expect to submit. The manuscript would probably be submitted to the journal of Theoretical and Applied Genetics.

6) Other comments (if any)

Please, for tables and figures see Anex.

	Intergenic	Intron	UTR	Exon	No match	Total
Di	25 (24.75)	27 (26.73)	17 (16.83)	1 (0.99)	31 (32.67)	101 (17.10)
Tri	57 (18.10)	17 (5.40)	49 (15.56)	133 (42.22)	59 (19.05)	315 (53.30)
Tetra	9 (20.93)	6 (13.95)	14 (32.56)	5 (11.63)	9 (20.93)	43 (7.28)
Penta	17 (28.33)	6 (9.52)	22 (36.67)	0 (0.00)	15 (25.00)	60 (10.32)
Hexa	11 (17.74)	3 (4.84)	10 (16.13)	22 (35.48)	16 (25.81)	62 (11.67)
Total	119 (20.48)	59 (10.15)	112 (19.28)	161 (27.71)	130 (22.36)	581 (100)

Table 1: Number and distribution of 581 EST-SSRs motifs found in 1342281 Oryza EST sequences downloaded from dbEST. During the SSRs search only EST sequences larger or equal to 100bp, and SSRs motif with 20 or more pair of bases were considered. Numbers between parentheses correspond with percentages.

	Intergenic	Intron	UTR	Exon	No match	Total
Di	3 (10.71)	5 (17.86)	15 (57.14)	1 (3.57)	3 (10.71)	27 (17.87)
Tri	8 (8.00)	2 (2.00)	22 (22.00)	64 (64.00)	4 (4.00)	100 (66.23)
Tetra	1 (12.50)	0	2 (25.00)	0	5 (62.50)	8 (5.30)
Penta	0	0	3 (75.00)	0	1 (25.00)	4 (2.65)
Hexa	1 (8.33)	0	1 (8.33)	10 (83.33)	0	12 (7.95)
Total	13 (8.61)	10 (6.62)	43 (28.48)	75 (49.67)	13 (8.61)	151

Table 2: Number and distribution of 151 EST-SSRs motifs found in 1532829 Arabidopsis EST sequences downloaded from dbEST. During the SSRs search only EST sequences larger or equal to 100bp, and SSRs motif with 20 or more pair of bases were considered. Numbers between parentheses correspond with percentages.

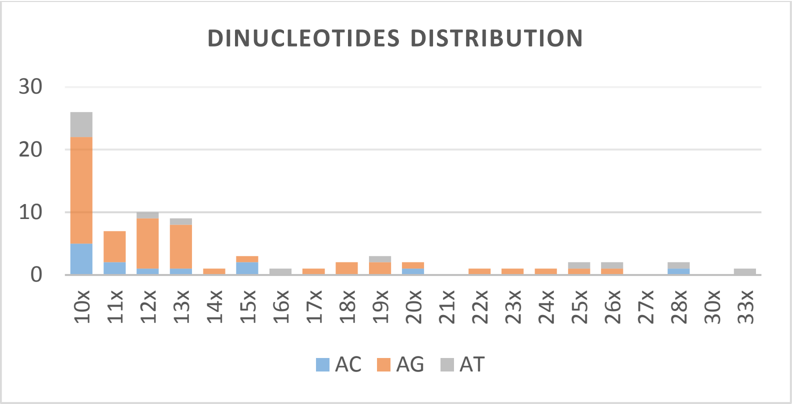

Fig 1: **Fig 3**: Dinucleotides distribution obtained from EST sequences using iQDD software that had a blast hit in the Oryza sativa (japonica cultivar-group) reference database from the NCBI. Dinucleotides are sorted by motif (AC in blue, AG in orange and AT in grey). Length of the repeated motif in axis X and number of the dinucleotides in axis Y.

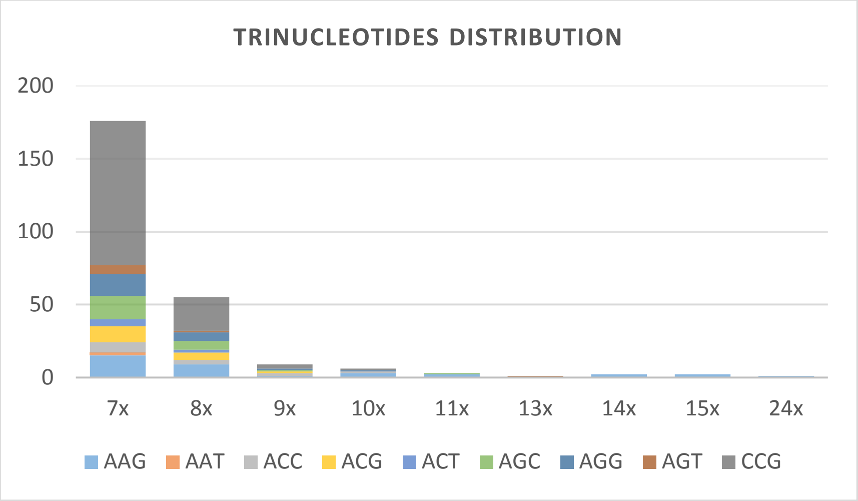

Fig 2: Trinucleotides distribution obtained from EST sequences using iQDD software that had a blast hit in the Oryza sativa (japonica cultivar-group) reference database from the NCBI. Trinucleotides are sorted by motif (each color matches one motif type). Length of the repeated motif in axis X and number of the dinucleotides in axis Y.

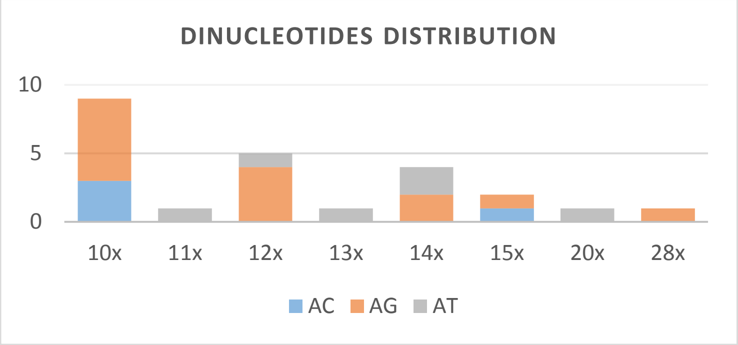

Fig 3: Dinucleotides distribution obtained from EST sequences using iQDD software that had a blast hit in the Arabidopsis thaliana reference database from the NCBI. Dinucleotides are sorted by motif (AC in blue, AG in orange and AT in grey). Length of the repeated motif in axis X and number of the dinucleotides in axis Y.

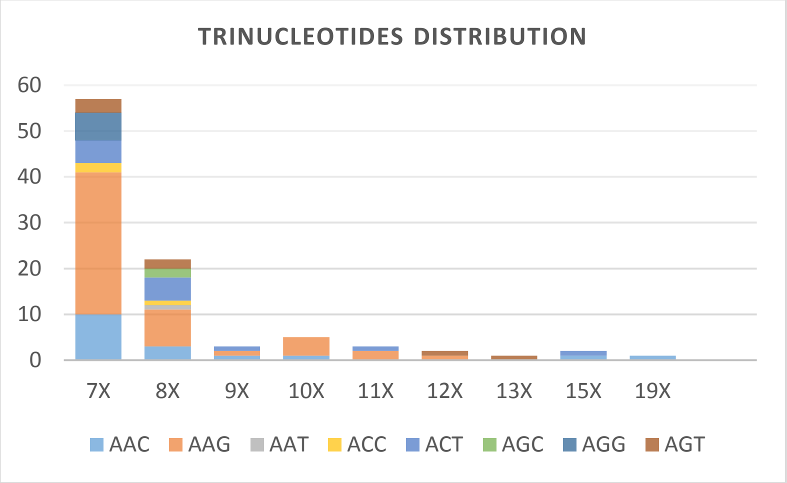

Fig 4: Trinucleotides distribution obtained from EST sequences using iQDD software that had a blast hit in the Arabidopsis thaliana reference database from the NCBI. Trinucleotides are sorted by motif (AC in blue, AG in orange and AT in grey). Length of the repeated motif in axis X and number of the dinucleotides in axis Y.

Group	Shared genera	Genera with SSR	Genera with ≥50 SSR	EST seqs.	Dimers	Trimers	Tetramers	pentamers	hexamers	Total	Common motifs
Florideophyceae	1	1	1	88280	16 (8.0)	77 (38.5)	39 (19.5)	38 (19.0)	30 (15.0)	200	TGA/CAG/CAT/GAT
Cariophyceae	2	2	1	16645	2 (8.0)	10 (40.0)	2 (8.0)	1 (4.0)	10 (40.0)	25	AAG/GTC
Ferns	5	3	1	35665	129 (81.65)	18 (11.39)	3 (1.89)	2 (1.27)	6 (3.80)	158	AGC/CAG/AG/GA
Lycopodiophyta	3	3	1	101292	20 (10.53)	122 (64.21)	15 (7.89)	7 (3.68)	26 (13.68)	190	GCA/ACG/CAG
Gymnosperms	18	15	3	1191184	144 (22.26)	145 (7.89)	30 (5.26)	58 (10.18)	193 (33.86)	570	AT/TA
Monocot	58	37	12	3197142	598 (19.24)	1395 (44.88)	296 (9.52)	323 (10.39)	496 (15.96)	3108	AG/GA/CCG/GCC/
Dicot	170	161	51	9868518	4339 (33.83)	4898 (38.19)	775 (6.04)	778 (6.07)	2035 (15.87)	12825	AC/GA/AAG/AGA/GGA
Total	**257**	**222 (86.38)**	**70 (31.53)**	**14498726**	**5248 (30.73)**	**6665 (39.03)**	**1160 (6.79)**	**1207 (7.07)**	**2797 (16.37)**	**17076**	

Table 5: Number of SSRs motifs found in 257 genera included in the IUCN red list with EST sequences in the dbEST. During the SSRs search only EST sequences larger or equal to 100bp, and SSRs motif with 20 or more pair of bases were considered. Numbers between parentheses correspond with percentages.

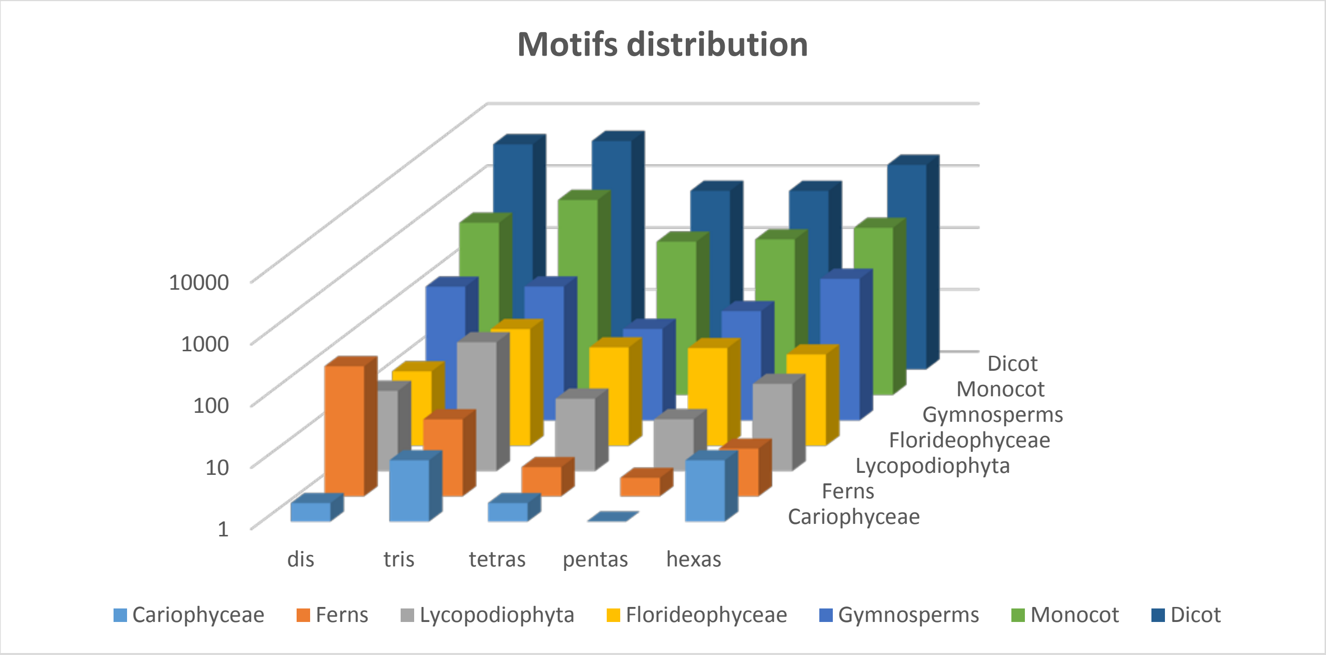

Fig 5: Distribution of SSR motif in 222 IUCN red list genera grouped in: Florideophyceae, Cariophyceae, Ferns, Lycopodiophyta, Gymnoperms, Monocot and Dicots. The axis Y (logarithmic scale) represents the number of motif. Axis X indicates the type of motis (di, tris, tetra, penta or hexa) and each colour indicates a different group (light blue Cariophyceae, orange for ferns, grey for Lycopodiophyta, yellow for florideophyceae, blue for gymnosperms, green for monocot and dark blue for dicots).