BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(17)

# Distribution and characterization of simple sequence repeats in *Gossypium raimondii* genome

**Changsong Zou, Cairui Lu, Youping Zhang & Guoli Song***

State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China; Guoli Song – Email: cs.zou@163.com; Phone: +86-(372)-2562375; FAX: +86-(372)-2562256; *Corresponding author

Received August 20, 2012; Accepted August 20, 2012; Published September 11, 2012

**Abstract:**
Simple sequence repeats (SSRs) can be derived from the complete genome sequence. These markers are important for gene mapping as well as marker-assisted selection (MAS). To develop SSRs for cotton gene mapping, we selected the complete genome sequence of *Gossypium raimondii*, which consisted of 4447 non-redundant scaffolds. Out of 775.2 Mb sequence examined, a total of 136,345 microsatellites were identified with a density of 5.69 kb per SSR in the *G. raimondii* genome leading to development of 112,177 primer pairs. The distributions of SSRs in the genome were non-random. Among the different motifs ranging from 1 to 6 bp, penta-nucleotide repeats were most abundant (30.5%), followed by tetra-nucleotide repeats (18.2%) and di-nucleotide repeats (16.9%). Among all identified 457 motif types, the most frequently occurring repeat motifs were poly-AT/TA, which accounted for 79.8% of the total di-nt SSRs, followed by AAAT/TTTA with 51.5% of the total tetra-nucleotede. Further, 18,834 microsatellites were detected from the protein-coding genes, and the frequency of gene containing SSRs was 46.0% in 40,976 genes of *G. raimondii*. These genome-based SSRs developed in the present study will lay the groundwork for developing large numbers of SSR markers for genetic mapping, gene discovery, genetic diversity analysis, and MAS breeding in cotton.

**Key words**: *Gossypium raimondii*, Simple Sequence Repeats (SSRs), distribution, molecular marker

**Background:**
Simple sequence repeats (SSRs) are tandemly repeated DNA motifs (1-6 bp long) which are present in both protein coding and non-coding regions of DNA sequences, and show a high level of length polymorphism due to mutations of one or more repeats. SSRs are easy to use and analyze by virtue of their multiallelic nature, reproducibility, high abundance and extensive genome coverage **[1, 2]**. The traditional methods of developing SSR markers are usually time consuming and labor-intensive. Generally these processes involve genomic library construction, hybridization with the repeated units of nucleotides and sequencing of the clones. The computational approach for developing SSR markers from the genome sequence provides a better platform than the conventional approach. Several bioinformatic tools for the identification of microsatellites in genomic sequences have been developed. The most commonly used tools for SSR search are: SSRIT **[3]**, TROLL **[4]**, MISA **[5]**, SSRFinder **[6]**, Modified Sputnik I and II **[7, 8]**, and SciRoKo **[9]**. SciRoKo is a user-friendly software tool for the identification of microsatellites in genomic sequences **[9]**.

Cotton (*Gossypium spp.*) is a major world agricultural crop, and the annual planting area is about 3,300 million hectares **[10]**. In recent years, molecular marker technology has been widely applied to such studies on cotton as genetic mapping **[11-13]**, genetic diversity analysis **[14]**, MAS **[15, 16]**, and gene tagging **[17]**. Due to the facts that the cotton genome is relatively large, with a 1C content of 2,250 Mb, and that intraspecific molecular polymorphism in this species is low, there is a major preoccupation for more highly polymorphic genetic markers for marker-assisted breeding programs. To date, approximately 17,000 pairs of SSR primers have been developed from four cotton species (*G. arboretum* L., *G. barbadense* L., *G. hirsutum* L., and *G. raimondii Ulbrich*) **[18]**. However, rare of them are able to

 

represent the large cotton genome adequately. In this study, the frequency and distribution of SSRs in the *G. raimondii* genome were characterized.

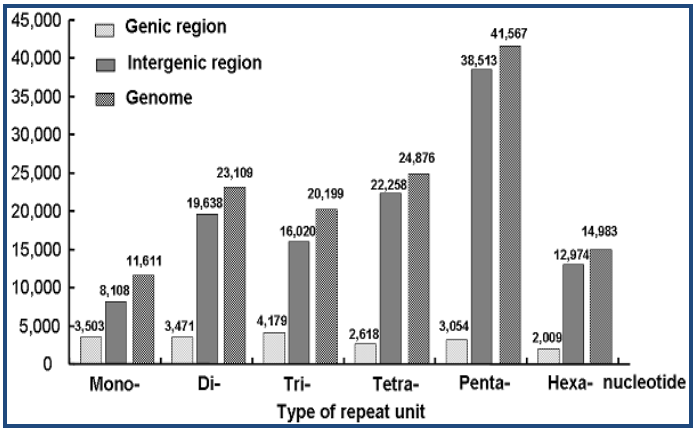


**Figure 1**: Frequency distribution of different repeat types identified in the *G. raimondii* genome

**Methodology:**

***Data source***

The genome sequence and annotation information of *G. raimondii* were download from the CGP (http://cgp.genomics.org.cn/page/species/mapview.jsp).

***SSR scanning and analysis***

The genome was scanned for SSRs 1oci with program software SciRoKo 3.4 (SSR Classification and Investigation by Robert Kofler) **[9]**. The parameters were set for detection of mono, di, tri, tetra, penta, and hexa -nucleotide (nt) motifs with a minimum of 15, 7, 5, 3, 3 and 3 repeats, respectively (under the mismatched and fixed penalty search mode). Initially, each SSR was considered to be unique and was subsequently classified according to theoretically possible combinations. The motif association statistic requires the standardizations. During standardization, the reverse complements of microsatellite motifs were considered, and similar microsatellite motifs are grouped together. For example, a poly-A repeat is equivalent to a poly-T repeat on a complementary strand, and an AAG is equivalent to AGA and GAA in different reading frames and to CTT, TCT and TTC on a complementary strand. Thus, there are two possible combinations for mono-nt repeats, four for di-nt repeats, and ten for tri-nt repeats, 33 for tetra-nt repeats, 102 for penta-nt repeats, and 350 for hexa-nt repeats. In this study, we defined two genomic location categories as genic (5'-Utr, exon, intron, and 3'-Utr) and intergenic regions. To locate the distribution of SSRs in different genomic regions, the position of SSRs were compared with the genome annotation by Perl scripts. To describe the abundance of SSRs in different genomic regions, we calculated the "relative abundance" (RA) by dividing the number of SSRs by the mega base-pair (MB) of sequences in our analyses.

***Primer designing***

Primer pairs were designed from the obtained SSR sequences by using Primer3. Perl scripts were used to operate Primer3 core code for batch designing primer. The major parameters for primer design were as follows: primer length, for which we selected 17-27 bp, with 20 bp being optimal, PCR product sizes of 100-250 bp, an optimum annealing temperature of 57°C, and a GC content of 30%-65%, with 50% being optimal. Then the SSRs were searched for both forward and reverse primers.

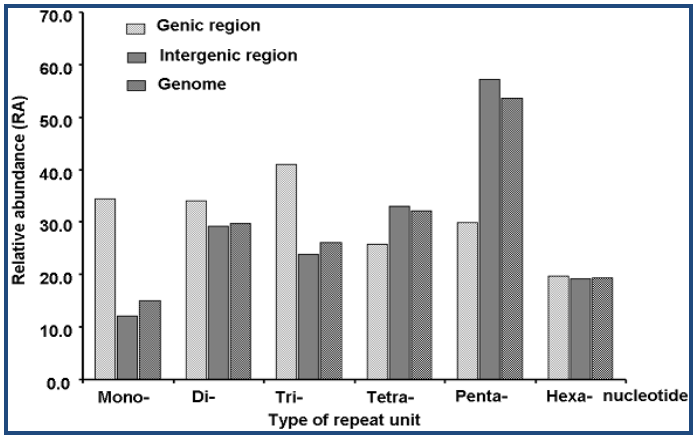


**Figure 2:** Genome-wide distribution and relative abundance of SSR types by their unit size. Each bar represents the relative abundance of the SSR types in different genome locations. Relative abundance = number of SSR type/region size in mega bases (Mb)

**Discussion:**

A total of 136,345 microsatellites were identified in the 775.1 Mb (containing 4447 scaffolds) genomic DNA sequence of G. raimondii using the SciRoko programs. With the help of core primer3 and Perl scripts, 112,177 primer pairs were obtained (Data was not shown). Among the SSRs we analyzed, 113,766 (83.4%) were perfect repeats, 22,579 (16.5%) were mismatched repeats. The results showed that SSRs were abundant in the *G. raimondii* genome with about one SSR every 5.69 kb **Table 1 (see supplementary material)**. The most abundant microsatellite was the penta-nt repeats of which 41,567 (30.5% of the SSRs) were identified, followed by the tetra-nt repeats (24,876, 18.2%) and di-nt repeats (23,109, 16.9%). The numbers of mono-nt, tri-nt and hexa-nt repeats were 11,611 (8.5%), 20,199 (14.8%), and 14,983 (11.0%), respectively **(Figure 1)**. The SSR loci were classified by repeat type and frequency of repeats per motif **Table 2 (see supplementary material)**. We found 457 types of repeat motifs in these SSRs. Among the SSR groups (standardization), the most abundant repeat motif type was poly-A/T in mono-, ploy-AT/TA in di-, poly-AAT/TTA in tri-, poly-AAAT/TTTA in tetra-, poly-AAAAT/TTTTA in penta-, and poly-AAAAAT/ TTTTTA in hexa-nucleotides. For each SSR type, the less frequency the SRR has with number of repeats the more.

These SSRs put insight into the frequency distribution of different types of nucleotide repeats in *G. raimondii.* More SSRs were found in the intergenic regions (64.1%) than in the genic regions (35.9%) **Figure 1 & Table 3 (see supplementary material)**. The different SSR repeat units showed obviously differential or non-random distributions in the different genomic locations. The microsatellite analysis showed that the distribution of SSRs in exonic, intronic and intergenic regions of the genome were non-random and strongly biased, probably reflecting the functional significance of SSRs. In general, the relative abundances of 3-UTR, 5-UTR, and intron region were considerably higher than that of intergenic region **Table 1 (see supplementary material)**; the tri-nt repeats were the most abundant SSR type in the genic region, whereas, penta-nt

repeats were the most abundant SSR type in the intergenic region **(Figure 2)**.The relative abundances of the tri-nt SSRs in the Coding Sequence (CDS) regions were 51.3 per Mb, which significantly higher other SSR types. The enhanced frequency of tri-nt SSRs in the coding regions might indicate the effects of selection against possible frameshift mutations.

In an attempt to analyze the differential distribution of SSRs more clearly, we characterized the distribution of the SSR types in each repeat unit across the different genomic locations **Table 4 & Table 2 (see supplementary material).** The results showed that the distribution of the different SSR types in the genome was non-random. For instance, of the two possible types of mono-nt SSRs, poly-A/T was the predominant form with 10,141 loci, about 89.6% of the total mono-nt loci. Of the ten possible of tri-nt SSRs, the poly-AAT/TTA accounted for 54.8% of the total tri-nt loci, followed by ploy-AAAT/TTTA accounted for 51.5% of the total terta-nucleotide. In genome, the most frequently occurring repeat motifs were poly-AT/TA, which accounted for 79.8% of the total di-nt SSRs **Table 2 (see supplementary material)**. In the genic region, the most frequently occurring tri-nt repeat motifs were poly-AAG/TTC, which accounted for 27.1% of the total tri-nt SSRs in CDS region **Table 2 (see supplementary material)**. Ignoring the mono-nucleotide repeats, the di- and tri- nucleotide repeat motifs with the highest frequencies were poly-AT/TA and -AAG/TTC in the genic region, respectively, which were identical with the previous reports **[19]**.

Currently a number of studies are being reported regarding the development of EST-SSRs in cotton species using the computational tools **[15, 19]**. Microsatellite markers are very important for studing genetic mapping, genetic diversity analysis, molecular marker-assisted breeding (MAS) **[13-16]**. In the present study, 18,834 microsatellites were detected from the total protein-coding genes, and the frequency of gene containing SSRs was 46.0% in 40,976 genes of *G. raimondii*. Although *G. raimondii* seeds contain no valuable fibers, the epidermal seed trichomes grow thickly. As one of the allotetraploid cotton donors, the D-subgenome has contributed important quantitative trail loci (QTLs) and/or genes to fiber development **[20]** With the help of gene function annotation, the putative functions of the genes could lead to find the important functional domain markers (FDM) related to gene ontology study such as stress response and fiber development, and develop the important FDM related to genetic diversity analysis and MAS for breeding in cotton species.

**Conclusion:**

Cotton commonly known as fiber crop is a plant of great commercial value. Up to now many works have been reported regarding the application of molecular markers in this plant for genetic mapping, gene discovery, genetic diversity analysis, and MAS. As such, the cotton research community has made efforts to develop many portable markers to overcome the problem of low DNA polymorphism rates among various cultivated cotton breeding programs (http://www.cottonmarker.org/). SSRs are the most powerful genetic markers for genetic linkage analysis, diversity study and marker assisted selection. High-resolution mapping in cotton has not been got because of limited DNA polymorphism within a cotton species. To explore the genetic make-up of cotton, inter-species variability, evolutionary relationship, development and application of molecular markers are of immense importance. The genome-based SSRs developed in the present study will shed light into the discovery of the information. This investigation is laying the groundwork for developing large numbers of SSR markers in cotton. The growing collection of portable markers in cotton provides a cost-effective tool for genome mapping and gene discovery to understand and improve the cotton species.

**Acknowledgement:**

This work was supported by a grant from the Specical Forehead Investigation for 973 Program (Grant No. 2011CB111511), and also by the Central Public-interest Scientific Institution Basal Research Fund, China (Grant No. SJB1210). The authors are grateful to Dr. Qin Li for providing the necessary facilities about analyze the SSRs in genome.

**References:**

**[1]** Kantety RV *et al. Plant Mol Biol*. 2002 **48**: 501 [PMID: 11999831]
**[2]** Morgante M & Olivieri AM, *Plant J*. 1993 **3**: 175 [PMID: 8401603]
**[3]** Temnykh S *et al. Genome Res*. 2001 **11**: 1441 [PMID: 11483586]
**[4]** Castelo AT *et al. Bioinformatics.* 2002 **18**: 634 [PMID: 12016062]
**[5]** Thiel T *et al. Theor Appl Genet*. 2003 **106**: 411 [PMID: 12589540]
**[6]** Gao L *et al. Mol Breed.* 2003 **12**: 245
**[7]** Morgante M *et al. Nat Genet.* 2002 **30**: 194 [PMID: 11799393]
**[8]** La Rota M *et al. BMC Genomics.* 2005 **6**: 23
**[9]** Kofler R *et al. Bioinformatics.* 2007 **23**: 1683 [PMID: 17463017]
**[10]** http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1282.
**[11]** Chee P *et al. Genome*. 2004 **47**: 449 [PMID: 15190362]
**[12]** Yu J*Z et al. J Anim Sci*. 2012 **2**: 43 [PMID: 2384381]
**[13]** Lin L *et al. BMC Genomics*. 2010 **11**: 395 [PMID: 20569427]
**[14]** Liu S *et al. Crop Sci.* 2000 **4**: 1459
**[15]** Zhang T *et al. Theor Appl Genet.* 2003 **106**: 262 [PMID: 12582851]
**[16]** Guo W *et al. Acta Agronomica Sinica.* 2005 **31**: 963
**[17]** Shen X *et al. Mol Breed.* 2005 **15**: 169
**[18]** Blenda A *et al. BMC Genomics*. 2006 **7**:132
**[19]** Wang C *et al. Chinese Science.* 2006 **51**: 557
**[20]** Kohel RJ *et al. Euphytica.* 2001 **121**:163

**Edited by P Kangueane**

**Citation: Zou *et al***. Bioinformation 8(17): 801-806 (2012)

BIOINFORMATION

open access

## Supplementary material:

**Table 1:** Summary of the SSR detection using SciRoKo

| Motif | Counts | SSR Length | | Average mismatches | Kb/SSR | GC-content (%) |
|---|---|---|---|---|---|---|
| | | Average (bp) | Standard deviation | | | |
| Mononucleotide | 11611 | 18.41 | 6.88 | 0.18 | 66.76 | 11 |
| Dinucleotide | 23109 | 22.19 | 20.77 | 0.21 | 33.54 | 12 |
| Trinucleotide | 20199 | 26.89 | 34.4 | 0.5 | 38.38 | 16 |
| Tetranucleotide | 24876 | 20.26 | 34.81 | 0.26 | 31.16 | 12 |
| Pentanucleotide | 41567 | 18.41 | 9.22 | 0.16 | 18.65 | 21 |
| Hexanucleotide | 14983 | 23.67 | 14.65 | 0.28 | 51.74 | 19 |
| Total | 136345 | 21.22 | 23.06 | 0.25 | 5.69 | 16 |

**Table 2:** Frequency and distribution of 136345 SSRs identified in *Gossypium raimondii* genome

| SSR motif | Number of repeats | | | | | | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | >20 | |
| A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3610 | 2255 | 1255 | 718 | 382 | 247 | 2205 | 10414 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 309 | 214 | 185 | 103 | 78 | 50 | 483 | 1197 |
| AT | 0 | 0 | 0 | 0 | 0 | 4504 | 3332 | 2410 | 2444 | 1533 | 1187 | 946 | 643 | 436 | 307 | 221 | 155 | 89 | 394 | 18432 |
| AG | 0 | 0 | 0 | 0 | 0 | 731 | 525 | 313 | 309 | 205 | 144 | 146 | 113 | 97 | 73 | 60 | 65 | 32 | 268 | 2982 |
| AC | 0 | 0 | 0 | 0 | 0 | 460 | 305 | 197 | 182 | 108 | 88 | 60 | 47 | 36 | 31 | 30 | 28 | 16 | 99 | 1687 |
| CG | 0 | 0 | 0 | 0 | 0 | 4 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 318 | 8 |
| AAT | 0 | 0 | 2083 | 1776 | 1821 | 843 | 845 | 588 | 591 | 501 | 419 | 296 | 277 | 176 | 186 | 135 | 128 | 87 | 413 | 11070 |
| AAG | 0 | 0 | 1155 | 950 | 688 | 377 | 283 | 181 | 141 | 93 | 83 | 44 | 33 | 22 | 38 | 21 | 17 | 15 | 111 | 4236 |
| ATC | 0 | 0 | 521 | 337 | 252 | 162 | 108 | 78 | 56 | 51 | 23 | 15 | 13 | 7 | 13 | 7 | 2 | 3 | 47 | 1664 |
| AAC | 0 | 0 | 369 | 251 | 158 | 90 | 55 | 38 | 33 | 17 | 17 | 14 | 10 | 6 | 3 | 1 | 1 | 3 | 33 | 1097 |
| ACC | 0 | 0 | 276 | 161 | 98 | 60 | 43 | 19 | 12 | 10 | 5 | 4 | 5 | 3 | 0 | 2 | 1 | 1 | 3 | 702 |
| AGG | 0 | 0 | 194 | 210 | 74 | 55 | 27 | 20 | 12 | 5 | 3 | 3 | 3 | 2 | 0 | 1 | 0 | 0 | 2 | 610 |
| AGC | 0 | 0 | 126 | 108 | 60 | 45 | 22 | 9 | 10 | 4 | 3 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 11 | 393 |
| ACT | 0 | 0 | 60 | 40 | 36 | 15 | 9 | 9 | 9 | 4 | 2 | 0 | 8 | 1 | 1 | 2 | 0 | 1 | 10 | 207 |
| CCG | 0 | 0 | 50 | 42 | 21 | 12 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 129 |
| ACG | 0 | 0 | 21 | 34 | 19 | 9 | 1 | 2 | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 91 |
| AAAT | 0 | 7201 | 3237 | 1432 | 514 | 222 | 96 | 47 | 25 | 16 | 5 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 5 | 12803 |
| AAAG | 0 | 2174 | 733 | 398 | 220 | 129 | 69 | 45 | 15 | 11 | 8 | 4 | 6 | 5 | 2 | 2 | 1 | 0 | 6 | 3826 |
| AATT | 0 | 1729 | 736 | 280 | 203 | 44 | 15 | 14 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 104 | 3030 |
| ATAC | 0 | 471 | 233 | 158 | 85 | 62 | 53 | 42 | 31 | 34 | 30 | 26 | 19 | 18 | 7 | 9 | 5 | 1 | 102 | 1386 |
| AAAC | 0 | 676 | 199 | 90 | 23 | 12 | 1 | 3 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1008 |
| AATG | 0 | 492 | 218 | 54 | 30 | 7 | 12 | 4 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 826 |
| AACC | 0 | 150 | 111 | 9 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 275 |
| ATGC | 0 | 170 | 63 | 23 | 4 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 266 |
| AATC | 0 | 156 | 36 | 11 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 208 |
| ATAG | 0 | 85 | 46 | 26 | 13 | 8 | 4 | 5 | 0 | 3 | 2 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 3 | 198 |
| AACT | 0 | 99 | 56 | 19 | 7 | 4 | 6 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 195 |
| AAGG | 0 | 103 | 30 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 140 |
| AGGG | 0 | 66 | 27 | 19 | 7 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 125 |
| ATCC | 0 | 74 | 28 | 10 | 6 | 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 124 |
| Other Tetra-* | 0 | 268 | 116 | 37 | 16 | 8 | 5 | 9 | 5 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 466 |
| AAAAT | 8101 | 4138 | 1256 | 337 | 126 | 43 | 14 | 5 | 6 | 2 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 14032 |
| AAATT | 3416 | 903 | 296 | 82 | 28 | 12 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4743 |
| AAAAG | 2808 | 1228 | 400 | 137 | 58 | 26 | 15 | 5 | 5 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4685 |
| CCCGG | 1580 | 729 | 203 | 50 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2569 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AATAT | 1080 | 395 | 209 | 119 | 63 | 59 | 31 | 20 | 8 | 16 | 5 | 5 | 2 | 0 | 1 | 1 | 0 | 0 | 3 | 2015 |
| AATCG | 712 | 724 | 385 | 61 | 15 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1906 |
| AAAAC | 801 | 356 | 82 | 33 | 7 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1281 |
| AATTC | 525 | 170 | 70 | 49 | 11 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 830 |
| AAATG | 561 | 141 | 39 | 6 | 8 | 5 | 5 | 4 | 2 | 3 | 2 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 780 |
| AATCT | 577 | 64 | 10 | 5 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 661 |
| AAGAG | 477 | 101 | 32 | 12 | 6 | 3 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 634 |
| AAATC | 307 | 150 | 40 | 22 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 523 |
| ACCGC | 413 | 71 | 9 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 499 |
| AGGGG | 381 | 50 | 8 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 443 |
| ATATC | 206 | 108 | 41 | 27 | 14 | 13 | 4 | 2 | 7 | 3 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 6 | 438 |
| AAAGG | 307 | 66 | 22 | 7 | 8 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 414 |
| ACCCG | 235 | 112 | 35 | 12 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 397 |
| AACCC | 258 | 90 | 16 | 9 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 382 |
| AAACC | 177 | 47 | 19 | 5 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 252 |
| AATGC | 164 | 65 | 14 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 245 |
| AAACT | 142 | 40 | 15 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 203 |
| AACAT | 123 | 36 | 11 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 176 |
| AATAC | 108 | 33 | 11 | 7 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 164 |
| AAAGC | 100 | 39 | 21 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 162 |
| AATGT | 114 | 29 | 8 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 152 |
| AATCC | 91 | 32 | 8 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 135 |
| AGAGG | 78 | 27 | 8 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 118 |
| AATAG | 82 | 18 | 10 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 115 |
| AAGGG | 62 | 34 | 10 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 110 |
| AACAC | 80 | 21 | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 108 |
| AATGG | 58 | 32 | 4 | 1 | 2 | 2 | 1 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 105 |
| ATAGC | 84 | 12 | 3 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 103 |
| AGAGC | 85 | 16 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 102 |
| Other Penta-* | 1357 | 485 | 143 | 54 | 21 | 11 | 3 | 3 | 2 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 2085 |
| AAAAAT | 2057 | 1001 | 268 | 100 | 39 | 25 | 10 | 13 | 5 | 2 | 2 | 3 | 0 | 1 | 0 | 1 | 0 | 2 | 1 | 3529 |
| AAAAAG | 787 | 594 | 181 | 85 | 45 | 20 | 12 | 2 | 4 | 6 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1740 |
| AAAATT | 514 | 343 | 100 | 27 | 8 | 6 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1001 |
| AAAATG | 287 | 158 | 90 | 25 | 20 | 21 | 6 | 4 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 614 |
| AAATAT | 290 | 184 | 53 | 21 | 10 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 563 |
| AAAAAC | 187 | 150 | 29 | 19 | 11 | 3 | 4 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 412 |
| AAATTT | 233 | 127 | 33 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 399 |
| AATTAT | 182 | 142 | 32 | 8 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 368 |
| AAAATC | 95 | 106 | 50 | 29 | 15 | 8 | 5 | 4 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 7 | 327 |
| ATATAC | 99 | 99 | 49 | 22 | 8 | 11 | 6 | 2 | 6 | 3 | 4 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 315 |
| AACCCT | 120 | 97 | 34 | 16 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 274 |
| AAAACT | 66 | 149 | 32 | 13 | 10 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 273 |
| AATATT | 103 | 58 | 24 | 6 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 196 |
| AAATTG | 118 | 43 | 7 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 171 |
| AAATGG | 32 | 48 | 39 | 17 | 11 | 10 | 2 | 3 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 166 |
| AAAAGG | 71 | 44 | 17 | 5 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 143 |
| AACAGT | 35 | 62 | 13 | 6 | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 127 |
| AAGATG | 61 | 41 | 14 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 123 |
| AAAACG | 83 | 28 | 5 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 119 |
| AAATTC | 78 | 30 | 6 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 117 |
| Other Hexa-* | 1653 | 1348 | 563 | 227 | 103 | 44 | 22 | 16 | 5 | 8 | 1 | 5 | 1 | 3 | 1 | 0 | 0 | 0 | 6 | 4006 |

* Motifs with less 100 SSRs were not listed

**Table 3:** Relative abundance of SSR types in different regions of the *G. raimondii* genome

| SSR types | Genic region(101.9Mb) | | | | | | | | | | Intergenic region (673.3Mb) | | Genome (775.2Mb) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CDS (45.2Mb) | | 3-UTR (3.9Mb) | | 5-UTR (3.9Mb) | | Intron (48.9Mb) | | | | | | |
| | count | RA* | count | RA | count | RA | count | RA | count | RA | count | RA | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mono | 17 | 0.4 | 345 | 88.5 | 468 | 120.0 | 2673 | 54.7 | 8108 | 12.0 | 11611 | 15.0 |
| Di | 30 | 0.7 | 245 | 62.8 | 385 | 98.7 | 2811 | 57.5 | 19638 | 29.2 | 23109 | 29.8 |
| Tri | 2318 | 51.3 | 192 | 49.2 | 328 | 84.1 | 1341 | 27.4 | 16020 | 23.8 | 20199 | 26.1 |
| Tetra | 58 | 1.3 | 188 | 48.2 | 319 | 81.8 | 2053 | 42.0 | 22258 | 33.1 | 24876 | 32.1 |
| Penta | 159 | 3.5 | 233 | 59.7 | 419 | 107.4 | 2243 | 45.9 | 38513 | 57.2 | 41567 | 53.6 |
| Hexa | 737 | 16.3 | 116 | 29.7 | 227 | 58.2 | 929 | 19.0 | 12974 | 19.3 | 14983 | 19.3 |
| Total | 3319 | 73.4 | 1319 | 338.2 | 2146 | 550.3 | 12050 | 246.4 | 117511 | 174.5 | 136345 | 175.9 |

*Relative abundance = number of SSR type / region size in mega bases (Mb)

**Table 4:** Distribution of SSRs types in different genic regions of the *G. raimondii* genome

| SSR motifs | Genic region repeat number | | | | Total |
|---|---|---|---|---|---|
| | CDS | 3-UTR | 5-UTR | Intron | |
| A | 13 | 304 | 420 | 2287 | 3024 |
| C | 4 | 41 | 48 | 386 | 479 |
| AT | 11 | 175 | 106 | 1690 | 1982 |
| AG | 17 | 46 | 227 | 654 | 944 |
| AC | 1 | 24 | 52 | 465 | 542 |
| CG | 1 | 0 | 0 | 2 | 3 |
| AAG | 628 | 31 | 161 | 354 | 1174 |
| AAT | 33 | 73 | 68 | 637 | 811 |
| ATC | 500 | 51 | 22 | 133 | 706 |
| ACC | 371 | 10 | 16 | 22 | 419 |
| AAC | 182 | 17 | 25 | 135 | 359 |
| AGC | 242 | 0 | 8 | 23 | 273 |
| AGG | 206 | 2 | 16 | 12 | 236 |
| Other Tri-* | 156 | 8 | 14 | 23 | 201 |
| AAAT | 3 | 50 | 110 | 856 | 1019 |
| AAAG | 13 | 39 | 90 | 270 | 412 |
| ATAC | 11 | 37 | 18 | 326 | 392 |
| AAAC | 3 | 15 | 27 | 191 | 236 |
| AATT | 1 | 8 | 9 | 97 | 115 |
| Other Tetra-* | 27 | 39 | 72 | 312 | 444 |
| AAAAT | 2 | 42 | 87 | 684 | 815 |
| AAAAG | 27 | 58 | 95 | 433 | 613 |
| AAAAC | 5 | 24 | 25 | 158 | 212 |
| AAATT | 1 | 7 | 15 | 146 | 169 |
| Other Penta-* | 124 | 102 | 202 | 817 | 1245 |
| AAAAAT | 5 | 11 | 45 | 224 | 285 |
| AAAAAG | 9 | 26 | 47 | 152 | 234 |
| Other Hexa-* | 723 | 78 | 135 | 554 | 1490 |

* Motifs with less 100 SSRs were not listed