INTERNATIONAL JOURNAL OF APPLIED BIOLOGY AND PHARMACEUTICAL TECHNOLOGY

www.ijabpt.com ISSN : 0976-4550

Volume-7, Issue-1, Jan-Mar-2016 Coden IJABFP-CAS-USA Copyrights@2016

Received: 7th Nov 2015 *Revised: 26th Dec 2015* *Accepted: 4th Jan 2016*

Research article

CARRAGHEEN MOLECULAR MARKER DATABASE (CAMM-DB): A COMPREHENSIVE DATABASE FOR CARRAGHEEN (*CHONDRUS CRISPUS*) MOLECULAR MARKERS

Jyotika Bhati, Tanmaya Kumar Sahu and Pankaj Kumar Pandey

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi-110012, India

ABSTRACT: Objective: CaMM-Db has been developed to manage the molecular marker information on *Chondrus crispus* (carragheen) and to make it accessible to the biological community. The database contains derived microsatellites and SNPs (Single Nucleotide Polymorphisms) from carragheen genomic sequences. The purpose for which carragheen is used, is also achieved by other Indian red seaweeds like, *Gracilaria* and *Hypnea* species. Except carragheen, the genome sequences of none of the red seaweeds are available in public domain. Thus, an insight into carragheen genome will help enable the biotechnologists to work on carragheen and the other red seaweeds.

Methods: Keeping the above in view, CaMM-Db was developed using the carragheen genomic sequences from the databases of National Center for Biotechnology Information (NCBI) for microsatellite determination and SNP discovery. Till date, no SNPs for *Chondrus crispus* are submitted to NCBI, thus, here an attempt has been made to discover SNPs from carragheen genomic sequences.

Results: This database provides information on different types motifs categorized based on different properties. The database is further integrated with Primer3 to facilitate the generation of suitable primers of interest for wet lab experimentation.

Conclusion: As it is the first database on the molecular markers in carragheen genome, it can be used as a valuable resource for the scholars indulged in genetic research on carragheen and other Indian seaweeds.

Key words: Red algae; Red seaweeds; SNPs; ESTs; SSRs; Primers

*Corresponding author: Pankaj Kumar Pandey, Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi-110012, India, E-mail: singh.jyotika@gmail.com

INTRODUCTION

Red seaweed or *Chondrus crispus* (commonly called carragheen) is widely available in coastal ecosystems and commonly called as Irish moss or carragheen moss. It is found in abundance along the Atlantic coast of Europe and North America. It is economically important as food as well as a source of gelling agent. It is used as one of the important industrial source of the carrageenan in the ice cream and food processing industries for thickening and stabilizing purposes. In different parts of the world it has different industrial applications. Though the cultivation of *Chondrus crispus* is rare in India, but similar seaweeds like *Gracilaria* and *Hypnea* species are used for the same purpose (Baghel RS et al, 2010, Gupta RS, Desa E. 2001, Valderrama D et al, 2013) in the country. Till date the genome of any Indian red seaweed is not available in public domain and *Chondrus crispus* is the only species of red seaweed sequenced till date. Thus, the genome of *Chondrus crispus* will give insights into various biological and metabolic systems of marine red algae along with its adaptations to the marine environment (Collen J et al, 2013).

Its genome size is 105 Mb with 9606 coding genes on a single chromosome. *Chondrus crispus* is investigated scientifically for the study of various stress responses, photosynthesis mechanisms in algae and also for the metabolic process like carrageenan biosynthesis. Although, its genome is publically available, unfortunately, the genomic sequences have received a little attention by the researchers.

Study of its genes and genome may open a new vista of research by providing new insights into the complex evolutionary forces that shape the eukaryotic genomes at molecular level. Further, the identification and study of molecular markers from genomic sequences will provide initial input and impetus to study its genes and genome for further applications.

The molecular markers are the short DNA sequences that include both Single Nucleotide Polymorphisms (SNP) and multi-base pair variations. An SNP is variation of a single nucleotide (A, T, C or G) in the genomic sequence that differs among members of a biological species or paired chromosomes.

SNPs are quite useful in many areas of biological research especially in Genome-Wide Association Studies (GWAS) as high-resolution markers for mapping the genes related to diseases or normal traits. SNPs without an observable impact on the phenotype are still useful as genetic markers in GWAS, because of their quantity and the stable inheritance over generations (Thomas PE et al, 2011).

Microsatellites are one of important molecular markers having high mutation rate that generate and maintain extensive length polymorphisms (Tautz D, Renz M. (1984). This property of microsatellite, makes it a powerful genetic marker for a variety of applications such as population genetics, genetic linkage mapping, parentage assignment, molecular breeding and allele mining etc. (Jarne P, Lagoda PJ. 1996, Bruford MW, Wayne RK. 1993). Also, it is evenly dispersed throughout eukaryotic genomes (Sarika, Arora V et al, 2012).

In the present study, two types of molecular markers are identified from carragheen genomic sequencesi.e, microsatellites and SNPs. The microsatellites or Simple Sequence Repeats (SSRs) were identified by taking ESTs, complete genes and contig sequences of carragheen from NCBI whereas SNPs were identified using only the available EST sequences. In addition, a molecular marker database (CaMM-Db) was also developed including the information on the derived microsatellites and SNPs. CaMM-Db is a unique database which provides information on the type of SSR repeats, simple and compound microsatellites, along with the characteristic of repeats like size, region and pattern etc. It is expected that this database will be a valuable resource for the biological community in many aspects of carragheen genetic research world-wide. (Benson DA et al, 2012).

DATA SOURCE

Genomic sequences of *Chondrus crispus* were downloaded from NCBI (Benson DA et al, 2012) in FASTA format. The microsatellite markers were identified using MIcroSAtellite tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). The output of MISA was processed using PERL scripts (http://pgrc.ipk-gatersleben.de/misa/download/) to identify the molecular markers and other metadata from the output files. Further the processed information was populated in the database according to the database schema. Besides, EST sequences were also used for SNP discovery.

DESIGN AND DEVELOPMENT OF DATABASE

In order to design the database, MySQL (version 5.5.16) was used. Tables were created and relationships among tables were established using normalization concepts. The unique, primary and foreign keys were created based on the third normal form of the database.

Tables were designed to store information about microsatellites and SNPs with detailed molecular information. The repeats of all microsatellites sequences (obtained using the repeat analysis program of 'MISA') were also updated in the corresponding tables. The detailed workflow for the development of the database is shown in Figure 1.

Four different tables for ESTs, Contigs, Genes and SNPs were designed. The SNP table has 6 attributes viz., snp_id, est_id, position, allele, left and right flanking sequences. Rest three tables have 12 attributes i.e., accession, contig id, repeat type, motif sequence, motif type, length, size, start and end coordinates, left and right flanking sequences and SSR sequences. These four tables together constitute the CaMM-Db.

WEB INTERFACE

Multi-user web application is provided by WAMP server that allows creating web applications with Apache 2 server, PHP (Hypertext Pre Processors, version 5.4) script and MySQL database.

The database tier is provided by the MySQL (version 5.5.16) whereas web interface was developed by using HTML (Hypertext Mark-up Language) and PHP. Further, Java Script was used for client side validations. The Web interface is equipped with different tools for searching, viewing and analyzing these molecular markers (Figure 2).

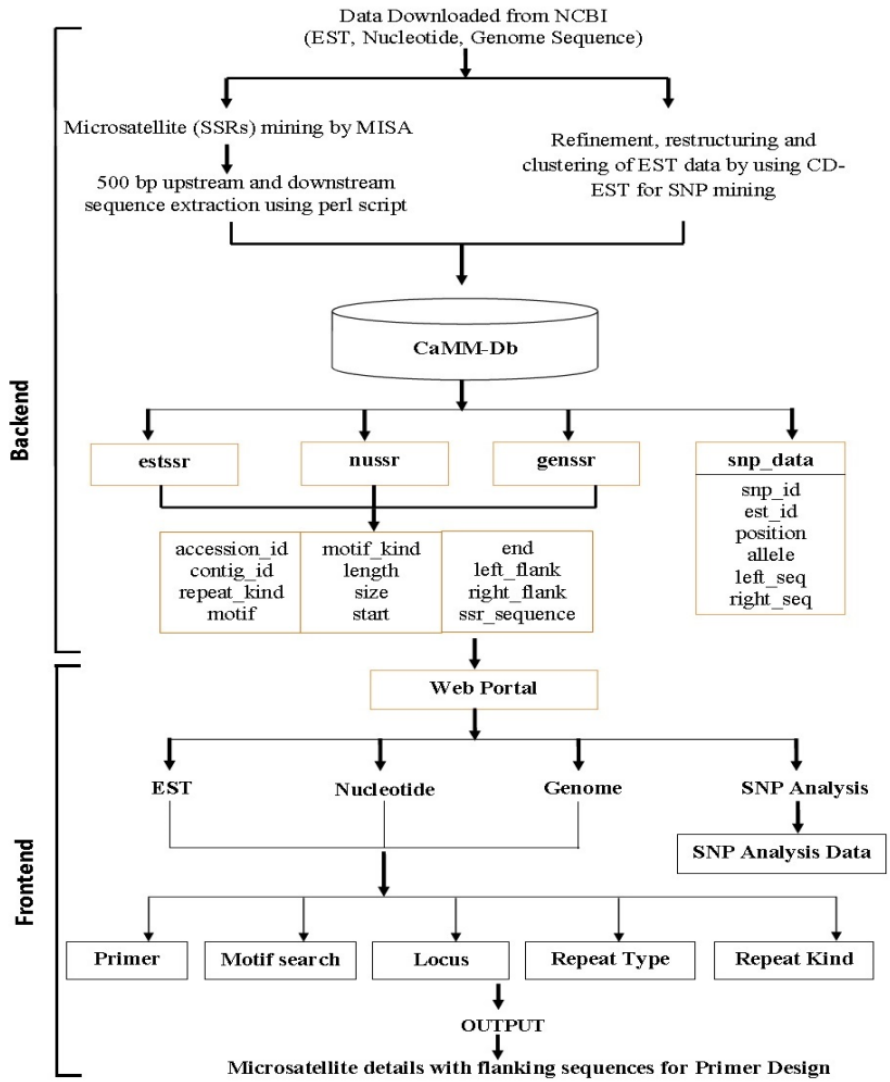

Figure 1. Architecture and data flow representation in 'CaMM-Db'.

IDENTIFICATION OF SNP'S

Generally, SNPs found in non-coding and junk region of the genome are not of much importance. However, their presence in coding regions having non-synonymous nature are vital from structural and functional view point as these are expected to contribute to the phenotype of the organism. Presently, number of SNPs detected from the expressed region of the carragheen genome is negligible. Therefore, an *in silico* approach has been followed in this study to detect the possible SNPs from expressed region of the carragheen genome.

Initially, a total of 4120 EST sequences were downloaded from NCBI (http://ncbi.nlm.nih.gov). In order to cluster these sequences based on identities, all the sequences were submitted to the CD-hit suite (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est) (Huang Y et al, 2010). The clusters were then made based on ten different levels of identities (90-99%) with the expectation that almost similar kind of sequences will be clustered together. Further, sequences in each cluster (with minimum of ten sequences) were independently aligned with Mega 6 software (Tamura K et al, 2013).

The output of Mega has been analysed through a developed PERL script for the identification of nucleotide position where, the occurrences of two different alleles is more than 35%. Again, these positions were checked for all the levels of identities and the positions with similar kind of allelic variations at all the levels of identities were considered as putative SNPs. Besides, with the help of another Perl script the EST IDs, SNP positions, alleles and flanking sequences were extracted and uploaded in the suitable tables of the database.

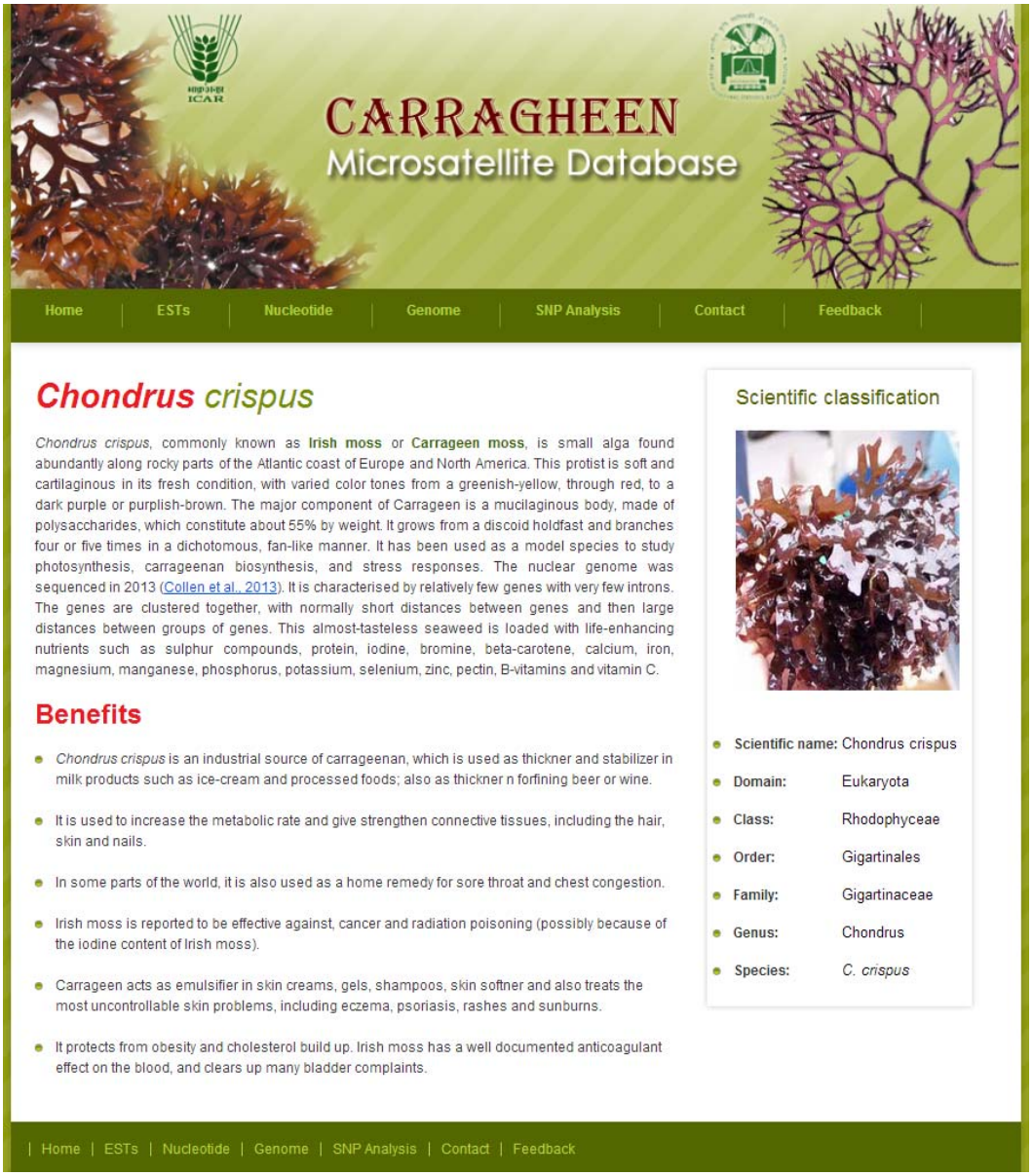

Figure 2.Homepage of CaMM-Db.

INFORMATION CONTENT OF THE DATABASE

CaMM-Db can be accessed to extract microsatellites based on motif type (mono-, di-, tri-, tetra-, penta- and hexamer), repeat motif and repeat kind (simple and compound) (Figure 3a). The ease of data search will enable researcher to select markers of their choice at desired region on a chromosome which may be coding or non-coding. Each sequence ID is linked to the main source i.e., NCBI. The system also provides information about the length of the SSR, start and end coordinates on the sequence and the complete SSR sequence (Figure 3b).

The database is further integrated with Primer3 for generation of suitable primers for wet lab experimentation. After selecting desired SSR markers based on customized search, the desired marker can be further processed for primer designing. The user may go for primer designing with default parameters as provided in this system or modify according to his requirement.

The flanking regions from both sides of the markers can be selected ranging from 100bp to 500bp (Figure 3c). The output of the system gives five best primers along with the melting temperature, product size and GC content (Figure 3d).

A total number of 1900 putative SNPs were identified from the expressed region of the genome. Out of which 580, 502, 462 and 356 were [-/G], [-/C], [-/A] and [-/T] type of indels respectively. The web interface for SNP search facilitates the user to search the SNPs based on SNP type and its position on a particular EST sequence. Figure 4 shows the result page of SNP search.

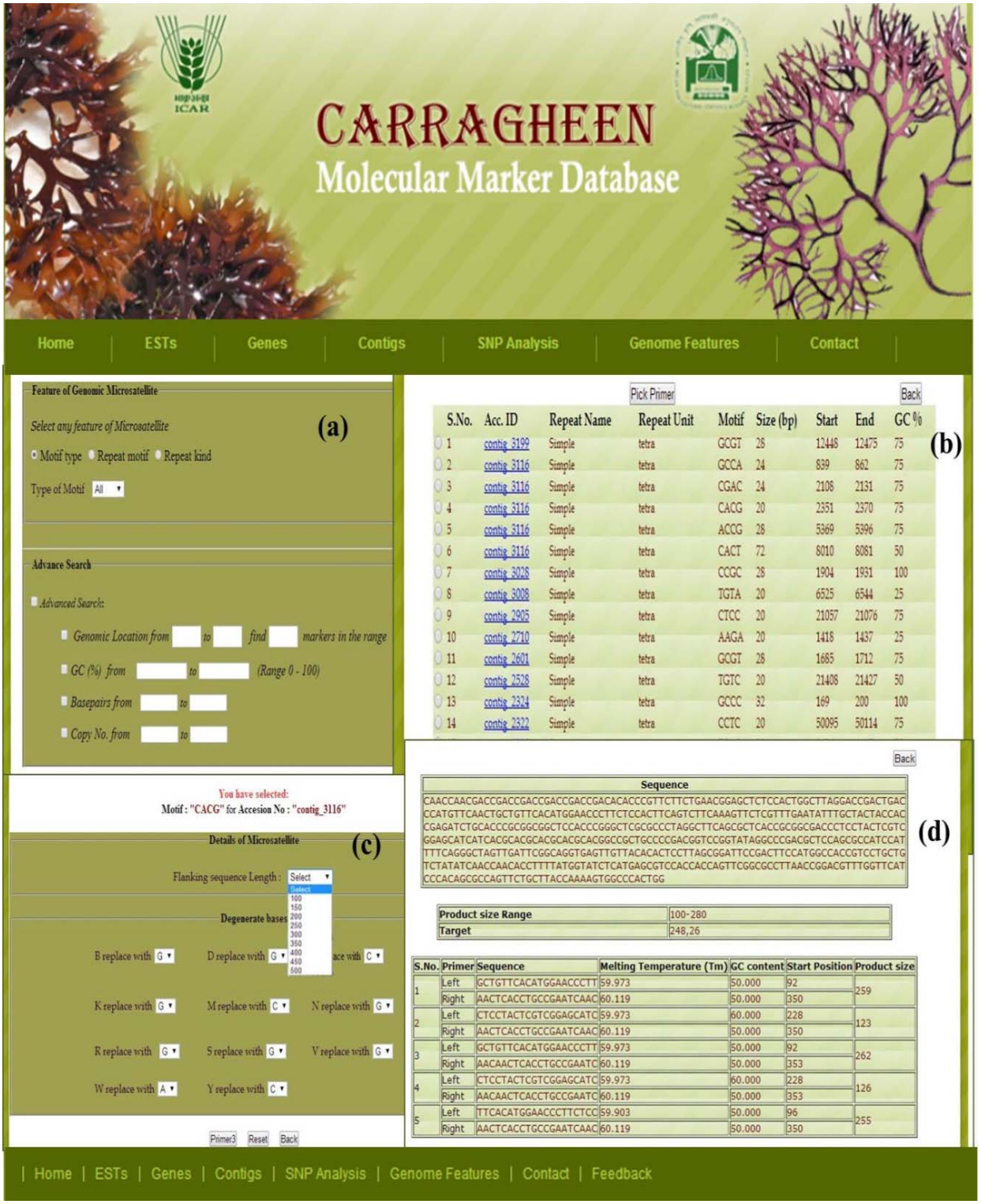

Figure 3. The web layout for SSR analysis and primer design of 'CaMM-Db' (a) search page for SSR mining (b) repeat analysis output (c) detail property of desired Primer and (d) SSR specific primer design output.

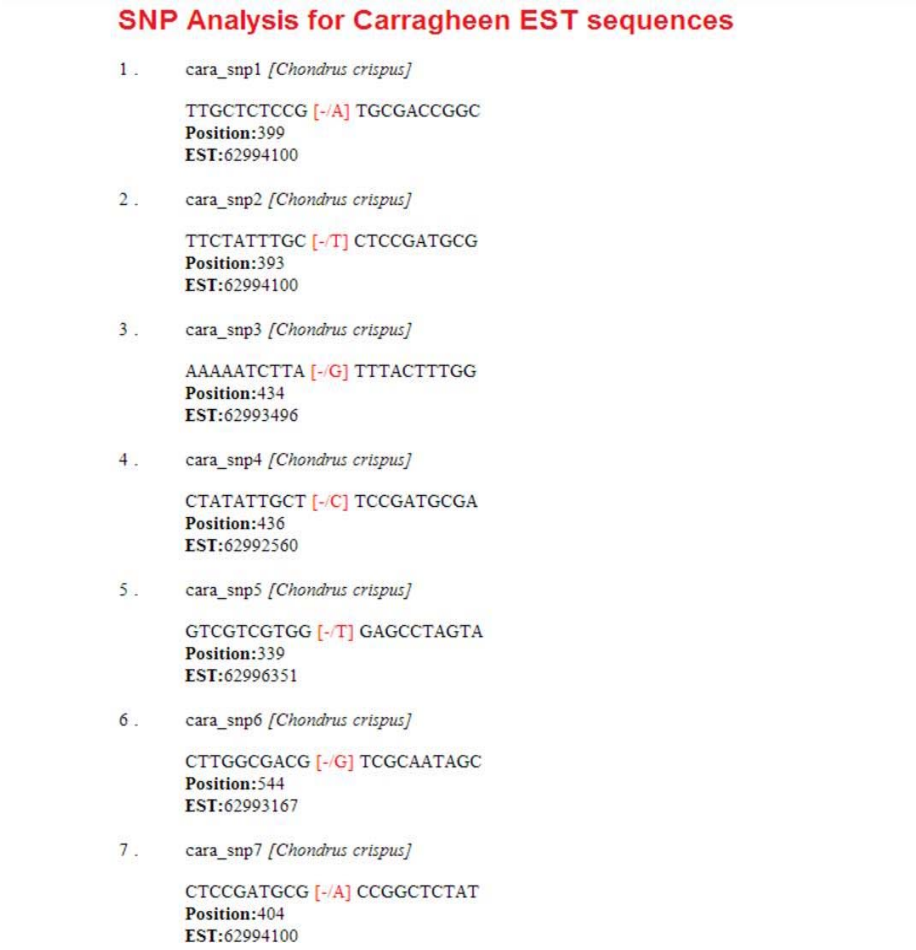

Figure 4. SNP analysis result of CaMM-Db.

UTILITY OF THE DATABASE

CaMM-Db stores 18660 SSR records of three types of the genomic sequences of carragheen genome (Table 1). This database would be of great help for researchers working on algal genomes, focusing mainly on molecular markers studies as *Chondrus crispus* is considered as model species for seaweed research. This also incorporates primer designing tool which facilitate the identification of polymorphism within the population. These studies lead to better understanding of functional importance of microsatellite makers.

Table-1: Details of the SSRs mined.

	EST	Complete Genes	Contigs
Monomer	87	6721	3174
Dimer	44	2766	1163
Trimer	45	2095	875
Tetramer	0	112	40
Pentamer	0	102	44
Hexamer	2	182	56
Compound SSR	19	789	344
Total	**197**	**12767**	**5696**

SNPs are good genetic markers due to their low heterozygosity, whereas microsatellites are good markers for studies of genetic linkage as they have a high heterozygosity (Miller JM et al, 2014). The high chance of mutability of microsatellites becomes a problem while considering allelic associations within populations. In contrast, due to low heterozygosity, SNPs offer a better chance of identifying marker-marker or marker-phenotype linkage disequilibrium. SNPs are also widely used in GWAS studies to establish the linkage between genetic variation and phenotypic traits (Stranger BE et al, 2011).

ANALYSIS OF CARRAGHEEN GENOMIC SEQUENCES

The complete data available in NCBI was analysed to get an overview of the carragheen genome. It was observed that almost 90% SSR markers were of simple type in all the three types of nucleotide sequences and rest of SSRs belongs to the compound type (Figure 5).

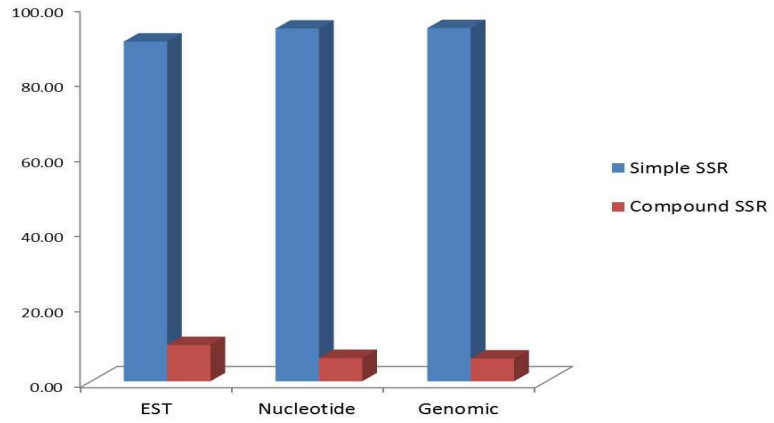

Figure 5. Distribution of SSR marker types.

Also, the monomer repeat type was found to be pre-dominant in all three sequence types followed by dimer (Figure 6).

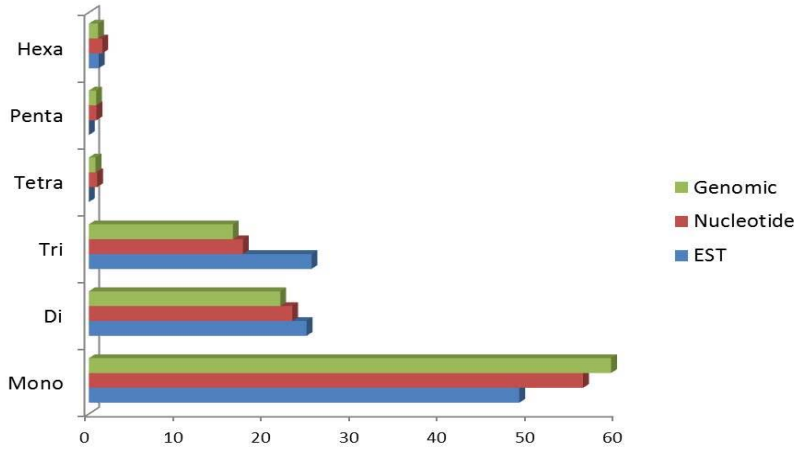

Figure 6. Distribution of SSR repeat types.

The distribution of monomer repeat types i.e., A/T or G/C (Figure 7) and dimer repeat types i.e., AT/TA, AG/GA/CT/TC, AC/CA/TG/GT and GC/CG (Figure 8) is calculated in all three sequence types.

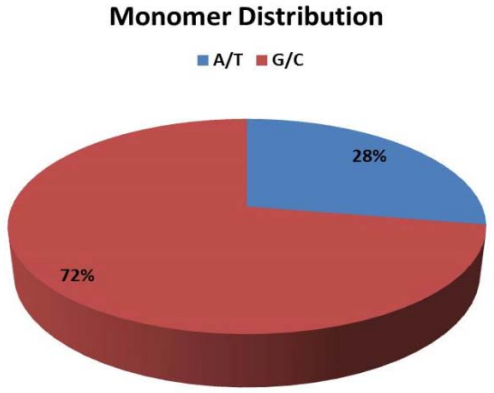

Figure 7. Monomer distribution of SSR.

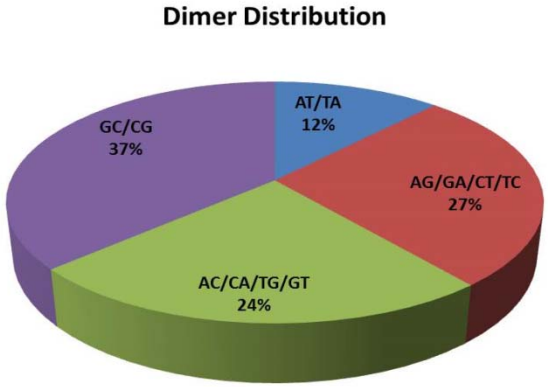

Figure 8. SSR Dimer distribution.

CONCLUSION

CaMM-Db is a database of molecular markers from carragheen genome containing 18,660 SSR markers and 12,512 putative SNPs. This database is expected to help the community of algal researcher. It provides valuable genomic information on carragheen at a single platform depending on what type of sequence need to be examined *viz.*, EST, genes or Contigs. The study also presents the frequent occurrence of a particular microsatellite repeat in Carragheen to discover new possibilities of research in these repeats.

SNPs analysed could be used for genome wide association studies as high-resolution markers in gene mapping related to diseases or normal traits. SNPs can provide powerful contributions to the population genetics studies to probe the evolutionary history of populations in unprecedented detail. This repository along with the included primer designing tool can play a key role in cutting edge areas of research by assisting with marker selection, linkage mapping, population genetics, evolutionary studies, genetic relatedness among the species and genetic improvement programmes of important Indian red seaweeds.

AVAILABILITY AND REQUIREMENT

CaMM-Db is freely available at URL http://webapp.cabgrid.res.in/carragheen/ for research and academic use.

ACKNOWLEDGEMENTS

Authors acknowledge the World Bank funded National Agricultural Innovation Project (NAIP), ICAR grant 30(68)/2009/Bioinformatics/NAIP/O&M. The scientific advice of Dr. MNV Prasad Gajula is thankfully acknowledged.

REFERENCES

Baghel RS, Kumari P, Bijo AJ, Gupta V, Reddy CR, Jha B. (2010). Genetic analysis and marker assisted identification of life phases of red alga *Gracilaria corticata* (J Agardh). Mol Biol Rep, 38, 4211-4218.

Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. (2012). Gene Bank. Nucleic Acids Res, 40, D48-D53.

Bruford MW, Wayne RK. (1993). Microsatellites and their application to population genetic studies. Curr Opin Genet Dev, 3, 939-943.

Collen J, Porcel B, Carre W. (2013). Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the *Archaeplastida*. Proc Natl Acad Sci USA, 110, 5247–5252.

Gupta RS, Desa E. (2001). *The Indian Ocean - a perspective*. Chp 16: Seaweed Resources, Vol 2, 563-584.

Huang Y, Niu B, Gao Y, Fu L, Li W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 26, 680-682.

Jarne P, Lagoda PJ. (1996). Microsatellites, from molecules to populations and back. Trends Ecol Evol, 11, 424-429.

Miller JM, Malenfant RM, David P, Davis CS, Poissant J, Hogg JT, Festa-Bianchet M, Coltman DW. (2014). Estimating genome-wide heterozygosity: effects of demographic history and marker type. Heredity, 112, 240-247.

Sarika, Arora V, Iquebal MA, Rai A, Kumar D. (2012). PIPEMicroDB: microsatellite database and primer generation tool for pigeonpea genome. Database: The Journal of Biological Databases and Curation,Vol. 2013 doi:10.1093/database/bas054.

Stranger BE, Stahl EA, Raj T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics, 187(2), 367-383.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol, 30, 2725-2729.

Tautz D, Renz M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res, 12, 4127-4138.

Thomas PE, Klinger R, Furlong LI, Hofmann-Apitius M, Friedrich CM. (2011). Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. BMC Bioinformatics, 12, S4.

Valderrama D, Cai J, Hishamunda N, Ridler N. (2013). Social and economic dimensions of carrageenan seaweed farming. Fisheries and Aquaculture Technical Paper No. 580. Rome, FAO.

ISSN : 0976-4550

INTERNATIONAL JOURNAL OF APPLIED BIOLOGY AND PHARMACEUTICAL TECHNOLOGY

Email : ijabpt@gmail.com

Website: www.ijabpt.com