International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 5 (2013): 234-238
Research Article
Open Access

ISSN: 2320-0634

SNP analysis in core mitochondrial genes of Chlorophytes

Himani Kuntal and Vinay Sharma*

Banasthali University, Banasthali- 304022, Rajasthan, India

*Corresponding author: Vinay Sharma; e-mail: vinaysharma30@yahoo.co.uk

Received: 11 August 2013 Accepted: 18 August 2013 Online: 01 September 2013

ABSTRACT

Single nucleotide polymorphisms (SNPs) represent the most frequent type of genetic polymorphism and thus provide a high density of markers near the locus of interest. However, the mining of SNPs and significance of SNPs in organellar genomes has not been completely understood. In the present work, mitochondrial genomes were investigated for the distribution and pattern of SNPs. In recent past, the availability of organelle genome sequences has allowed us to understand the organization of SNPs in their genic and intergenic region. Most of the SNPs in mitochondrial genes are neutral with respect to protein structure therefore can be used to study divergence in closely related species. In this study, the SNPs were identified and categorized in six mitochondrial genes of five species of class Chlorophyceae of green algae and many of their properties such as DNA polymorphism, codon usage bias, conserved DNA regions, indels etc. were measured. The data revealed that nad2 gene exhibited highest degree of polymorphism and significantly the indels were also observed in great amount in same. This work constitutes the first report of an exhaustive comparison of mitochondrial SNPs in algal species and has revealed important information thereon.

Keywords: SNP, Mitochondrial genomes, Chlorophytes

INTRODUCTION

Nuclear and mitochondrial DNA is thought to be of separate evolutionary origin;the latter DNA being derived from circular genomes of bacteria that were engulfed by early ancestors of today's eukaryotic cells. In most multicellular organisms, mitochondria are normally inherited from mother [1]. Biologist can determine and then compare mitochondrial DNA sequences among different species and use the comparison to build an evolutionary tree for the species examined [2].

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide A,T,G,or C in the genome differs between members of a biological species [3]. Almost all common SNPs have only two alleles [4]. The genomic distribution of SNPs is not homogenous. SNPs are found throughout the genome, e.g. in exons, introns, intergenic regions, in promoters or enhancers [5]. SNPs usually occur in non-coding regions more frequently than in coding regions or in general where natural selection is acting[6]. A SNP in coding region may directly impact a relevant protein, an intronic SNP can influence splicing [7], and a SNP in a promoter can influence gene expression [8]. The degree to which each kind of SNP influences phenotypic expression is likely to receive a great deal of attention as more and more SNPs are identified and studied. The application area of SNPs is broad.Primarily they act as genetic markers and are used for gene identification and genetic mapping [9]. Besides this major application, SNPs are also utilized in species/strain identification and in diagnosis or risk profiling [5, 10]. In past, we have extensively worked on data mining of organelle genomes and have particularly studied the distribution and pattern of simple sequence repeats (SSRs) in the chloroplast and mitochondrial genomes [11-13].

In present study, mitochondrial genes of organisms belonging to Chlorophyta (a group of green algae) were analyzed for SNP identification and characterization of their frequency and distribution patterns among the genes.

MATERIALS AND METHODS

Data source

The coding regions of mitochondrial genomes of the five species of phylum Chlorophyta were retrieved from the GenBank (http://www.ncbi.nlm.nih.gov /GenBank) database of NCBI. The species used in the analysis and their accession numbers are listed below (Table 1).

Table 1. Chlorophyta species used in single nucleotide polymorphism analysis

Name of species	Accession no.	References
Chlamydomonas reinhardtii	NC_001638.1	[14]
Chlamydomonas eugametos	NC_001872.1	[15]
Dunaliella salina	NC_012930.1	[16]
Polytomella capuana	NC_010357.1	[17]
Scenedesmus obliquus	NC_002254.1	[18]

Methods

Gene profile was created for the selected species of Chlorophyta. Six genes namely nad1, nad2, nad4, nad6, cob and cox1, were found common in all selected species.Multiple sequence alignment was performed for individual gene sequence dataset using ClustalW [19].Results were analyzed using DnaSP v5 [20] software for identification of SNPs, DNA polymorphism, polymorphic sites, conserved DNA regions, indel polymorphism and codon usage patterns.

RESULTS AND DISCUSSION

Polymorphic sites

The common gene sequences were analyzed for polymorphic sites, the number of monomorphic sites, the number of polymorphic sites segregating for two, three, or four nucleotides, the total number of parsimony-informative sites (sites that have a minimum of two nucleotides that are present at least twice), and non-informative sites (singleton sites) (Table 2).

Table 2. Polymorphic sites detected by DnaSP in the six genes

Genes	nad1	nad2	nad4	nad6	cob	cox1
Number of sites	973	1683	1545	666	1209	1609
Monomorphic sites	244	187	58	108	378	607
Singleton variable sites	310	506	151	187	416	535
Parsimony informative sites	217	328	54	126	301	361
Singleton variable sites (two variants)	191	238	85	58	235	310
Singleton variable sites (three variants)	93	208	52	63	129	166
Singleton variable sites (four variants)	26	60	14	26	52	59
Parsimony informative sites (two variants)	122	164	39	68	170	205
Parsimony informative sites (three variants)	95	164	55	58	131	156

The data represented maximum monomorphic sites in cox1 gene sequence and minimum in nad4 gene but also maximum singleton sites in cox1 gene and minimum in nad4 genes. The same proportional data is available for parsimony informative site. This result was also biased due to length of sequence.

The data showed the cox and cob genes as most conserved having the maximum number of conserved sites although the singleton variable sites are higher than parsimony informative site in each case representing higher mutational frequency.

DNA Polymorphism

DNA polymorphism analysis for all six genes showed uniformity for the highest and the lowest values. A total of 1326 sites were analyzed in nad2 gene that showed the highest values for nucleotide diversity (pi, the average number of nucleotide differences per site between two sequences and its sampling variance) and theta (θ) (Waterson estimator; estimating population mutation rate) from total number of mutations (η) and theta (θ) (from number of polymorphic sites, S), that are 0.513, 0.623, 0.392 respectively. The highest values for variance of theta (in no recombination condition, 0.038) and standard deviation of theta (in no recombination condition, 0.195) belonged to nad2. As an exception however, the highest values for the number of polymorphic sites, the total number of mutations that were 896 and 1336 found in cox1 and nad4 showed 0.00061 and 0.024 values for the variance of theta (free recombination) and standard deviation (free recombination) respectively; these being the highest values. The total number of sites analyzed in nad4 and cox1 were 1545 and 1609 respectively. The lowest values for nucleotide diversity (π), theta (θ) from eta (η), theta from polymorphic sites, variance of theta (in no recombination condition), variance of theta (θ) (in free recombination condition), standard deviation of theta (θ) in both conditions (no recombination and free recombination) were 0.360, 0.426, 0.286, 0.020, 0.00009, 0.412, 0.009 found in cox1 gene. The nad4 gene showed exceptionally lowest number of polymorphic sites (245) and total number of mutations (1336) (Table 3).

Table 3. DNApolymorphism and their variances

Genes	nad1	nad2	nad4	nad6	cob	cox1
Selected region	1-973	1-1683	1-1545	1-666	1-1209	1-1609
Number of sites	973	1683	1545	666	1209	1609
Total number of sites	771	1021	303	421	1095	1503
Number of polymorphic sites	527	834	245	313	717	896
Total number of mutations	767	1326	380	486	1081	1336
Nucleotide diversity	0.40739	0.51361	0.49637	0.46057	0.39991	0.36001
Theta (per site) from Eta	0.47751	0.62339	0.60198	0.55411	0.47386	0.42667
Theta (per site) from S	0.32809	0.39209	0.38812	0.35686	0.3143	0.28615
Variance of theta (no recombination)	0.02674	0.0381	0.03766	0.03176	0.0245	0.02029
Standard deviation of theta (no recombination)	0.16353	0.19521	0.19408	0.17822	0.15653	0.14245
Variance of theta (free recombination)	0.0002	0.00018	0.00061	0.00041	0.000138	0.00009
Standard deviation of theta (free recombination)	0.01429	0.01358	0.0248	0.02017	0.01174	0.00956

Insertion-deletion (indel) Polymorphism

In evolutionary studies, indel simply refers to the mutation class that includes both insertions, deletions, and the combination of both. The highest number of indel sites was detected in nad2 and the lowest number in nad4. Equal numbers of indel events were analyzed for gene nad1 and cox1. Indel haplotypes were found to be common for all genes (5) except nad4 (2). The same trend applied in case of haplotype diversity. InDel Diversity per site, Pi (i) which is the analogue of Pi,and the nucleotide diversity were computed as k(i)/m, where m is the net number of positions analyzed. It was found to be maximum for nad2 and the minimum value was shown for cox1. Theta was calculated from the indel events and was found to be highest for nad2 (18.72). A statistical test, Tajima's D was performed. The purpose of the test was to distinguish between a DNA sequence evolving randomly and one evolving as a non-random process. The highest value of Tajima's D was calculated for cox1 (-0.20) and nad4 showed the lowest value (-1.16) (Table 4).

Table 4. InDel (Insertion-Deletion) Polymorphism

Genes	nad1	nad2	nad4	nad6	cox1	Cob
Number of sites	973	1683	1545	666	1609	1209
Total number of sites	771	1021	303	421	1503	1095
Total number of InDel sites analysed	73	88	11	28	45	46
Total number of InDel sites	202	662	1242	245	106	114
Total number of InDels events analysed	13	39	7	8	12	17
Total number of InDels events	23	111	45	37	23	27
Number of InDel Haplotypes	5	5	2	5	5	5
InDel Haplotype Diversity	1	1	0.4	1	1	1
InDel Diversity	5.6	17.4	2.8	3.6	5.6	7.6
InDel Diversity per site, Pi(π)	0.007	0.015	0.009	0.008	0.0036	0.0066
Theta (per sequence) from I	6.24	18.72	3.36	3.84	5.67	8.16
Tajima's D value	-0.74	-0.53	-1.16	-0.44	-0.2	-0.5

Codon Usage Bias

There are two codon bias measures- ENC (Effective Number of Codons) and CBI (Codon Bias Index). ENC quantifies the "effective" number of codons that are used in a gene. For the nuclear universal genetic code, the value of ENC ranges from 20 (only one codon is used for each amino acid; i.e., the codon bias is maximum) to 61 (all synonymous codons for each amino acid are equally used; i.e., no codon bias). Among all the six genes, the maximum value of ENC was shown by nad1 (51.62) indicating the synonymous use of amino acids and minimum value was shown by nad6 gene (35.504) (Fig. 1).

The second measure, i.e. CBI indicates the deviation from the equal use of synonymous codons. CBI values range from 0 (uniform use of synonymous codons) to 1 (maximum codon bias). Here the maximum CBI value was shown by nad6 gene (0.64) while the minimum value was detected in cob gene (0.385) and nad4 (0.386) (Fig. 2).

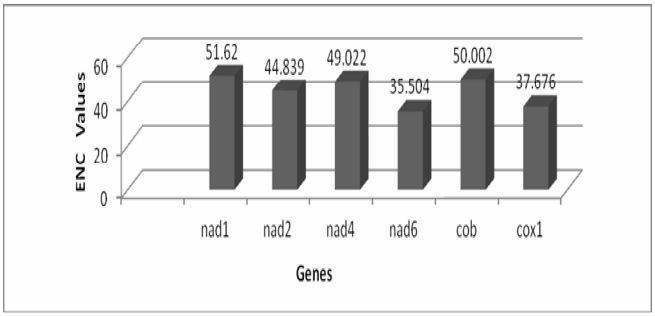

Figure 1. Effective number of codon values of genes

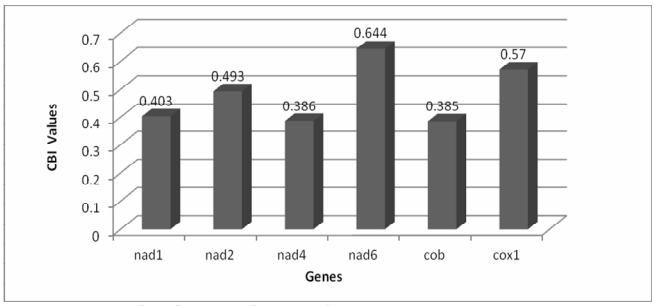

Figure 2. Codon bias indexes of genes

The highest G+C contents in coding region (G+Cc) were detected in cox1 gene (0.429) while the lowest value was reported in nad6 gene (0.378). The same condition was seen in case of G+C contents in 2nd coding position (G+C2). Here also cox1 possessed maximum G+C2 content and nad4 contained the minimum. G+C contents in 3rd coding position were detected highest for nad1 gene (0.403) and the lowest value was observed for nad2 (0.323) (Fig. 3).

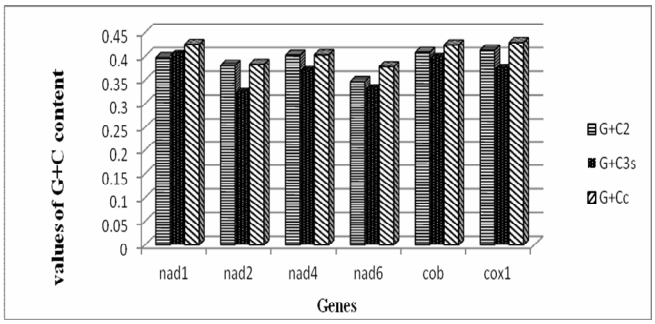

Figure 3. G+C contents of the genes

Conserved DNA regions

It represents the value of conservation along the whole sequence as well as in different regions of the individual sequence. It divides the sequences into different overlapping regions and calculates the conservation depending on the decided conservation threshold. The number of regions in which the sequences are divided depends on the length of the sequences. 917 polymorphic sites were found with 0.402 sequence conservation in cox1 gene that contained 1534 analyzed sites. Here the number of sites analyzed and the value of sequence conservation were highest for cox1 gene. The number of polymorphic sites in nad4 gene was 972 that showed a very little difference from cox1 gene despite low values for sequence conservation and conservation threshold that was 0.271 and 0.37 respectively. The lowest numbers of sites analyzed and number of polymorphic sites were 489 and 366 respectively in nad6 gene. Despite of moderate number of analysed sites, the nad4 gene showed the lowest values for sequence conservation and conservation threshold that were 0.183 and 0.28 respectively (Table 5).

The characteristics of SNPs in mitochondrial genomes of rice, date palm, castor beans and other species have also been investigated [21-24]. SNPs have also been characterized in fungal genomes *Neurospora crassa* in nad3, nad4L and nad5 genes, where they were detected in intronic portion only [25]. At the same time chloroplast SNPs have also been projected in bar-coding approaches [26-27]; but all previous works considered data from same genus.

Our study has revealed that NAD subunits are more polymorphic than other core genes of Chlorophyte group with the indels present in great amount in nad2 gene. To the best of our knowledge, ours is the first report where an exhaustive and comprehensive analysis of SNPs in mitochondrial genomes of algal species has been carried out and has revealed important information thereon.

Table 5: Sequence conservation and conservation threshold

Genes	nad1	nad2	nad4	nad6	Cob	cox1
Number of sites analysed	895	1330	1334	489	1157	1534
Number of polymorphic sites	602	1086	972	366	764	917
Sequence conservation	0.327	0.813	0.271	0.252	0.340	0.402
Conservation threshold	0.42	0.28	0.37	0.35	0.43	0.5

CONCLUSION

SNPs are abundant across prokaryotic and eukaryotic genomes and play an important role in genome evolution. We identified six common mitochondrial genes in five green algal species and investigated their properties. In Chlorophyta members, the nad2 gene showed the highest number of polymorphic sites, while the least were shown by nad6. DNA polymorphism data revealed that the nad2 exhibits maximum values for pi (π), eta (η), theta (θ),variance of theta and standard deviation of theta which clearly accounts for uniformity i.e. the one which shows maximum values for one property shows it mostly for all the other properties too and vice versa. The highest number of indel sites was detected in nad2 and the least in nad4. The statistical test Tajima's D gave negative values for all genes which indicated that low frequency polymorphism was more than the expected value. Tajima's D value was found to be least for nad4 and highest for cox1 gene. Maximum value for sequence conservation was again for nad2 while the least for nad6. Conservation threshold was highest for cox1 while least for nad2. The nad1 gene showed the highest value of ENC indicating the synonymous use of amino acids in codon usage. The maximum codon bias was detected in the nad6 gene. The overall analysis represented that NAD subunits are more polymorphic than other core genes of Chlorophyte group.

REFERENCES

1. Penman D (2002). Mitochondria can be inherited from both parents. Science. 251: 16-20.

2. Brown WM, George M and Wilson AC (1979). Rapid evolution of animal mitochondrial DNA. Proceedings of National Academy of Sciences of the United States of America. 76: 1967–1971.
3. Jehan T and Lakhanpaul S (2006). Single Nucleotide Polymorphism (SNP) - Methods and applications in plant genetics: A review. Indian J Biotechnol. 5: 435-459.
4. Stenson PD, Mort M, Ball EV et al. (2009). The Human Gene Mutation Database. Genome Med. 1: 13.
5. Schork NJ, Fallin D and Lanchbury S (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet. 58: 250–264.
6. Barreiro LB, Laval G, Quach H et al. (2008). Natural selection has driven population differentiation in modern humans. Nat Genet. 40: 340–345.
7. Krawezak M, Reiss J and Cooper DN (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet. 90: 41–54.
8. Drazen JM, Yandava CN, Dube L et al. (1999). Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. Nat Genet. 22: 168–170.
9. Kollers S, Kerstens HHD, Kommadath A et al. (2009). Mining for single nucleotide polymorphisms in pig genome sequence data. BMC Genomics.10: 4.
10. Satya R, Zavaljevski N and Reifman J (2011). SNIT: SNP identification for strain typing. Source Code Biol Med. 6: 14.
11. Kuntal H and Sharma V (2011). *In silico*analysis of SSRs in mitochondrial genomes of plants. OMICS. 15: 783-789.
12. Kuntal H, Sharma V and Daniell H (2012). Microsatellite analyses in organelle genomes of Chlorophyta. Bioinformation. 8: 255-259.
13. Sharma V, Kuntal H, Katara P et al.(2012). Identification of SSRs in EST sequences of *Lolium* and *Agrostis* species. Int JIntegr Biol. 13: 81-87.
14. Ma DP, King YT, Kim Y et al. (1992). Amplification and characterization of an inverted repeat from the Chlamydomonasreinhardtii mitochondrial genome. Gene. 119: 253-257.
15. Denovan EM, Nedelcu AM and Lee RW (1998). Complete sequence of the mitochondrial DNA of Chlamydomonas eugametos. Plant Mol Biol. 36: 285-295.
16. Smith DR, Lee RW, Cushman JC et al. (2010). The Dunaliella salina organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol. 10: 83.
17. SmithDR and Lee RW (2008). Mitochondrial genome of the colorless green alga Polytomellacapuana: a linear molecule with an unprecedented GC content. Mol Biol Evol. 25: 487-496.
18. Kuck U, Godehardt I and Schmidt U (1990). A self-splicing group II intron in the mitochondrial large subunit rRNA (LSUrRNA) gene of the eukaryotic alga Scenedesmus obliquus. Nucleic Acids Res. 18: 2691-2697.
19. Thompson J, Higgiin D and Gibson T (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nuc Acids Res. 22: 4673-4680.
20. Librado P and Rozas J (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25: 1451-1452.
21. Tian X, Zheng J, Hu S et al. Yu J (2006). The Rice Mitochondrial Genomes and Their Variations. Plant Physiol. 140: 401–410.
22. Fang Y, Wu H, Zhang T et al. (2012). A Complete Sequence and Transcriptomic Analyses of Date Palm (Phoenix dactylifera L.) Mitochondrial Genome. PLoS ONE. 7: e37164.
23. Rivarola M, Foster JT, Chan AP et al. (2011). Castor Bean Organelle Genome Sequencing and Worldwide Genetic Diversity Analysis.Castor Bean Organelle Genome Sequencing and Worldwide Genetic Diversity Analysis. PLoS ONE. 6:| e21743.
24. Gaur R, Azam S, Jeena G et al. (2012). High-Throughput SNP Discovery and Genotyping for Constructing a Saturated Linkage Map of Chickpea (Cicerarietinum L.). DNA Res. 19: 357-373.
25. McCluskey K. (2012). Variation in mitochondrial genome primary sequence among whole-genome sequenced strains of Neurospora crassa. IMA Fungus 3: 93–98.
26. Doorduin L, Gravendeel B, Lammers Y et al. (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. 18: 93–105.
27. Schroeder H, Höltken A, Fladung et al. (2011). Chloroplast SNP-marker as powerful tool for differentiation of Populus species in reliable popular breeding and barcoding approaches. BMC Proceedings 5: P56.

© 2013; AIZEON Publishers; All Rights Reserved

This is an Open Access article distributed under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*****