Title: Exploring repeats in rice genomes: Identification, Characterization and its Applications

Gourab Das[1,2], and Indira Ghosh[2,*]

[1]Bioinformatics & Computational Biology Facility (BCBF), Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Tata Memorial Centre (TMC) Sector 22, Utsav Chowk - CISF Rd, Owe Camp, Kharghar, Navi Mumbai, Maharashtra 410210.

[2]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi -110067, India

[*]Corresponding author. Email: indira0654@gmail.com

Keywords:

Repetitive sequences, Abiotic stress, *Oryza*, Pointwise mutual information, Evolution

Abstract:

Biodiversity is a fundamental property of all natural systems existing in the field biology. It refers to the underlying heterogeneity at different levels of ecology, genetics and evolution. In case of plant systems, dramatic variability has been observed during the Anthropocene at different spatial scales. Environmental stress is one of the major influencing factors behind this plant biodiversity. Huge genetic diversity has been also demonstrated across varieties of important crop species like rice. Repetitive sequences which are a major contributor of genomic diversity in polyploidy plants have been found to occur ubiquitously in their genomes. To date diverse repeat types have been characterized in the plant genomes performing various functions starting from qualitative trait markers to genome evolution and stress management. With an objective to identify of plant stress associated genes using DNA repeat probes, a robust method has been developed. The method has been modularized into three distinct sections. First part is dedicated for identification of different types of repeats. Earlier review has suggested building a pipeline of multiples tools for capturing different types of repeats from the genome sequences. Specialized tools like TROLL, Tandem repeat finder (TRF), PHOBOS and database like REPBASE have been selected for performing this job. Second module is intended for screening of stress related genes from the published articles and databases and the last module has been designed for the association mining between genomic repetitive patterns with stress phenotypes. The method has been used to explore stress associated repeats from 9 *Oryza* species from different continents and other plants like *Arabidopsis* and *Brachypodium*. In case of *Oryza* species distribution repeats has been found to be significantly different between stress associated and housekeeping genes. More than 55% of the repeats are found to be in positive association (nPMI > 0) whereas 26% of the repeats are false-positives. These repetitive probes have been utilized in several applications. Firstly, using as molecular markers to identify stress related genes in different *Oryza* species where availability is limited. Secondly, using as a probe to reanalyzing the evolutionary lineage of *Oryza* species etc.

Introduction

Rice is one of the mostly used cereals that have been served as a staple food for approximately half of the world population [1]. Along with increasing heat stress due to uprising global warming, several other abiotic stressors including drought, salinity, chemicals, greenhouse gases and biotic stressors like bacteria, fungus viroids pose adverse influence on the growth of the rice plant and its yield [2-6]. Repetitive sequences are abundant in rice genomes [7-10] majority being transposable elements (TEs) [11-13] covering more than 70% of the genome and often play key roles in stress adaptation and disease resistance through many genetic and epigenetic regulatory mechanisms [14-15]. Earlier researches have shown the presence of tandem repeats in excess in defense responsive genes [16]. Microsatellites have become a popular probe for marking candidate genes related to salinity [17], biotic and abiotic stresses [18-21] nowadays. Moreover, influence of transposable elements (TEs) have been significantly observed in genome evolution [22], stress response [23-25], regulations through microRNAs [26], long non-coding RNAs [27] and in many more. In this context, studying repeats in rice genomes to investigate its functional association in abiotic stress management and other applications will be highly demanding and beneficial for better understanding the rice biology.

Oryza sativa, a model cultivated species of Asian rice with AA genome, is a diploid angiosperm having one embryonic leaf in their seeds (monocot) and possess 24 chromosomes (2n=24) though one-half of the rice species are allopolyploid [28-29]. In *Oryza* species, total 10 different types of genomes have been characterized which includes 6 diploids (AA, BB, CC, EE, FF, GG) and four allotetraploids (BBCC, CCDD, HHJJ, HHKK) with 48 chromosomes (2n=48) [28]. It is believed that *Oryza* genus has emerged from Gondwana-land approximately 130 million years ago [30] though the debate is still ongoing regarding the actual origin of rice. Both subspecies *O. sativa* ssp. *japonica* (short grain) and *Oryza sativa* ssp. *indica* (long grain) have been domesticated from wild grass ancestor *Oryza rufipogon* around 10000-14000 years ago which is another controversial and live topic of research [31]. *O. rufipogon* (AA genome) is the perennial progenitor of *O. sativa* and wild species *Oryza nivara* (AA genome) is thought to be the recent annual predecessor [32]. African rice *Oryza glaberrima* (AA genome) is believed to be domesticated from wild species *Oryza barthii* (AA genome) around 2000-3000 years ago [33]. Other African wild rice species *Oryza punctata* which possess BB type genome has its own importance for showing resistance to biotic stressors like bacteria, planthoppers etc. Another African wild rice *Oryza brachyantha* is the only known rice species with FF type genome. The only diploid species from Latin America is *Oryza glumaepatula* (AA genome) which has a significant similarity with *O. sativa* genome as estimated by Fuchs et al. [34]. All of these species from different geographical locations with various types of genomes indicates its rapid diversification by dynamic evolution for legitimate adaptation under stressful conditions [35-36]. This provides a suitable platform to perform genome-wide comparative analysis of repetitive elements among rice species exploring its association with several biosynthesis, biogenesis and stress related pathways.

In the present study, different types of TRs and TEs have been identified, characterized and compared among multiple wild type and cultivated genomes to mine their abundance and strength of association with stress related genes. Repeats have been also used as probes for detecting novel stress genes in rice genomes and analyzing evolutionary relationship. These repeats can be served as a knowledgebase for rice stress gene markers.

Materials and Methods

The complete workflow for the proposed research has been presented in Figure 1. It consists of three modules. A. Identification and characterization of the repetitive sequences from genomes. B. Screening of stress related genes from the published articles and databases and C. Mining association between repeats and stress.

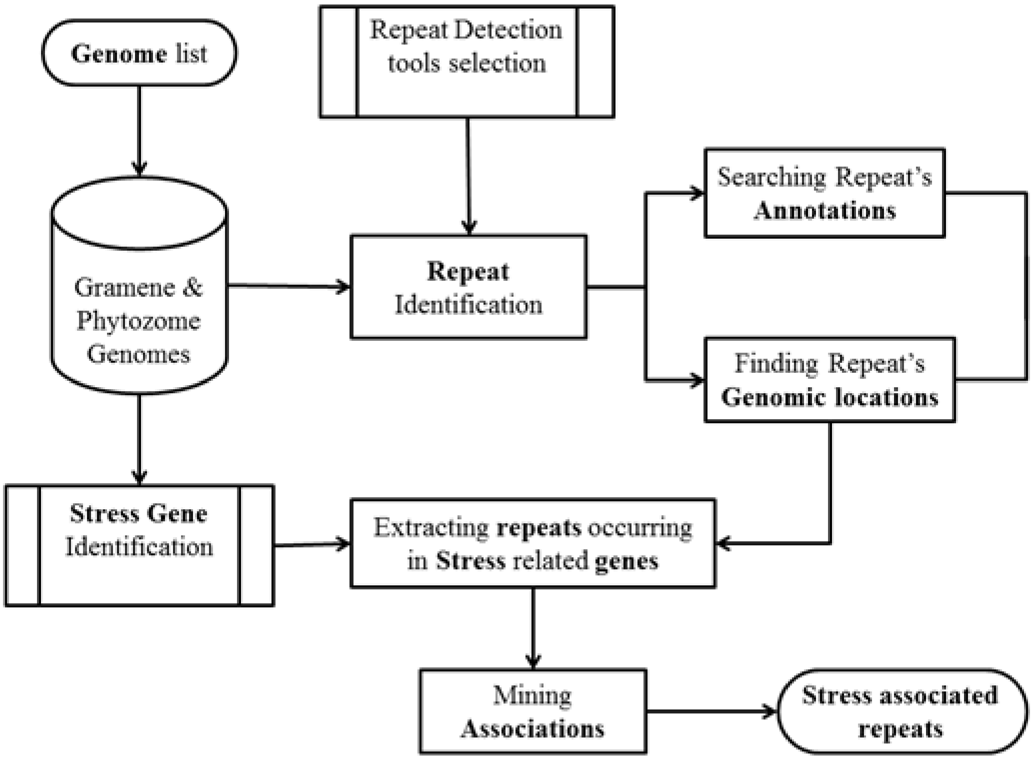

Genome list
Repeat Detection tools selection
Gramene & Phytozome Genomes
Repeat Identification
Searching Repeat's Annotations
Finding Repeat's Genomic locations
Stress Gene Identification
Extracting repeats occurring in Stress related genes
Mining Associations
Stress associated repeats

Figure 1: Basic workflow for identification, characterization of repeats in rice genomes and mining its association with stress.

Genome datasets and annotations

Chromosome sequences of 9 *Oryza* species and other model species *Brachypodium distachyon* (monocot) and *Arabidopsis thaliana* (eudicot) have been downloaded from Gramene-Ensembl database (release 34) [37] using file transfer protocol services (ftp://ftp.gramene.org/pub/gramene/). All protein-coding and non-coding RNA genes available in gff3 format have been parsed for further use.

Table 1: List of Rice and other genomes used in the present study

No.	**Common Name**	**Scientific name**	**Assembly**	**Annotation**
1	African wild rice	*Oryza barthii*	GCA_000182155.1	OGE Maker (Aug 2013)
2	Wild rice from tropical Africa	*Oryza brachyantha*	GCA_000231095.2 / OGEv1.4b	OGEv1.4
3	Wild red rice from Africa	*Oryza punctata*	GCA_000573905.1	OGE Maker (Aug 2013)
4	African rice	*Oryza glaberrima*	GCA_000147395.1 / AGI1.1 (May 2011)	2011-05-AGI (MIPS)
5	Wild rice from America (South)	*Oryza glumaepatula*	GCA_000576495.1	OGE Maker (Aug 2013)
6	Wild progenitor of Asian rice	*Oryza rufipogon*	PRJEB4137	OGE
7	Wild progenitor of Asian rice	*Oryza nivara*	GCA_000576065.1	OGE Maker (Aug 2013)
8	Asian rice	*Oryza sativa* ssp. *japonica*	IRGSP-1.0	MSU 7.0
9	Asian rice	*Oryza sativa* ssp. *indica*	ASM465v1	2010-07-BGI
Other model species selected in the present study				
1	Arabidopsis/Thale cress	*Arabidopsis thaliana*	TAIR 9	TAIR 10
2	Stiff Brome grasses	*Brachypodium distachyon*	V1.0	V1.0

Major resources of plants complete genomes include Phytozome, NCBI RefSeq, Gramene-Ensembl etc. Phytozome is the Plant Comparative Genomics portal developed by Department of Energy's Joint Genome Institute (JGI) with genome sequences and annotations of 81 green plant species including the genome of *O. sativa* ssp *japonica* only like NCBI Reference Sequence (RefSeq) database. But Gramene [38] provides the platform for inter-genus comparison with 9 *Oryza* genomes (now 11 in the recent release) so that consistency in results can be checked.

Regarding earlier map-based sequence has provided a rice genome assembly of size 389 Mbp with 95% coverage divided into 12 chromosomes [39] But, few months back a nearly complete de novo assembly of *O. indica* rice genome of estimated size of 390 Mbp and 99% coverage [40]. Table 1 shows the list of *Oryza* genomes used in the present study for comparative analysis.

Extraction of repeats

In the present study, different types of repetitive patterns have been explored in the rice genomes which include perfect, imperfect types of tandem repeats and many families of transposable elements. Tandem repeats of motif lengths 1-100 bp have been searched in the genomes using selected tools and parameters listed in Table 2. For interspersed repeats like transposable elements, the expert curetted library from Repbase [41] has been used for further research work. Many new repetitive sequence databases have developed recently [42-43] which can be also used to validate the repeat mining.

Tandem Repeats

Tandem repeats have been searched in the genomes using TROLL [44] for perfect microsatellites, PHOBOS [45] for perfect minisatellites and TRF [46] for imperfect repeats. In many of the recent papers, minimum repeat locus length criteria have been used as follows: mono: 8/12, di: 8, tri: 9, tetra: 12, penta: 15, hexa: 18 and for rest of the minisatellites at least two copies of the repeating motif. In the present study, relaxed locus length criteria for mononucleotide repeats have been used i.e. 8bp. No maximum limit has been set to include long repeats of functional importance. In case of imperfect repeats, minimum alignment score criteria are calculated by multiplying match weight of 2 with a minimum length of the locus for each repeat motif length ranging from 1-100.

Table 2: List of parameters for repeat identification

Repeat Type	Selected Tool/ Database	Motif Length	Minimum Locus length	Additional parameters
Perfect	TROLL, PHOBOS	1-100	8 for mono; 12 for rest	Not Applicable
Imperfect	TRF	1-100	Minimum 8 copies for mono, 4 copies for di and rest 3 copies at least.	Match score: 2 Mismatch and INDEL penalty: 7 Imperfection: 10%
Interspersed	Repbase v20.05	Not Applicable	Not Applicable	E-value: 10^{-6} Query coverage: 95% Sequence identity: 95%

Interspersed Repeats: Transposable elements

Transposable elements constitute the major portions of the genomes. Expert curetted libraries consisting of several families and subfamilies of retrotransposons, DNA transposons are already available in databases like Repbase. *Oryza sativa* transposon sequence library prepared by REPEATMASKER software (http://www.repeatmasker.org/) in Repbase database has been used to search these curetted families in other rice genomes using blast with 10^{-6} e-value cut-off and 95% sequence identity and query coverage [47]. Parameters used in the present work for extraction of different classes of repeats have been tabulated in Table 2.

Annotation and locations of the repeats

Annotations of repeats are parsed with respect to their locations in the genomes. For repeats occurring in the exons, UTRs and CDS regions, gene symbol, locus tag and product information of the gene have been used as the annotation of the repeat. Rice Annotation Project (RAP) [48] has taken the initiative to detect loci by mapping the sequences of transcripts and proteins to other monocot species and also to integrate annotation data from various resources like published literature, specialized databases consisting of annotations for several plant genomic elements e.g. microRNAs, small RNAs etc. As repeats are found to occur in the small and long RNA encoding genes, annotations from these databases can be used for the repeats.

Statistically significant repeats

To find the statistical significance the extracted repeats of motif length 1-100 bp, comparisons have been done chromosome-wise across rice species by generating random genomes for each chromosome of every rice species. 1-sample t-test at significance level 0.05 has been conducted to check whether repeats are randomly distributed in the genomes or not.

Comparisons among Oryza, Arabidopsis, and Brachypodium

Model dicotyledonous species *Arabidopsis* and another monocotyledonous wild grass *Brachypodium* have been selected for comparison with *Oryza*. Genomic parameters and repeat density parameters for both tandem and interspersed repeat classes have compared among all *Oryza* species with *Arabidopsis* and *Brachypodium.* To calculate whole genome size, individual sizes of all chromosomes have been added. The arithmetic mean of the %GC contents of all chromosomes is chosen for the %GC content comparison. The weighted mean of chromosome-wise gene densities per kbp has been computed. In case of calculating the %genic, %exonic and %coding region for the whole genome, respective amounts from each chromosome are added and then normalized by total genome size. While calculating repeat density parameters, repeat density in length parameter has been found same way as did for the %coding region calculation and for repeat density in number parameter weighted mean is used like gene density per kbp calculation.

Searching repeats in the stress-related genes and association mining

As mentioned before, several experimental and computational studies have been cited in literature for identifying stress-related genes in rice and other crop species [49-52]. Many microarray-based transcriptomic studies also have been conducted and available in literature [25, 53-54] for stress associated gene detection. However, the available list is still not comprehensive and with rapid changes in global climate, more genes are affected and to be associated with stress tolerance mechanisms in rice. Nevertheless, positive (stress associated) and negative (uniformly expressed in any condition) sets have been constructed from available lists of genes reported in the published articles (Figure 2). Both positive and negative sets of genes are available for only *Oryza sativa* species. Hence, comparisons of repeat distributions have been performed for this species only. A nPMI based method [55] has been utilized for finding an association between stress and the repeats.

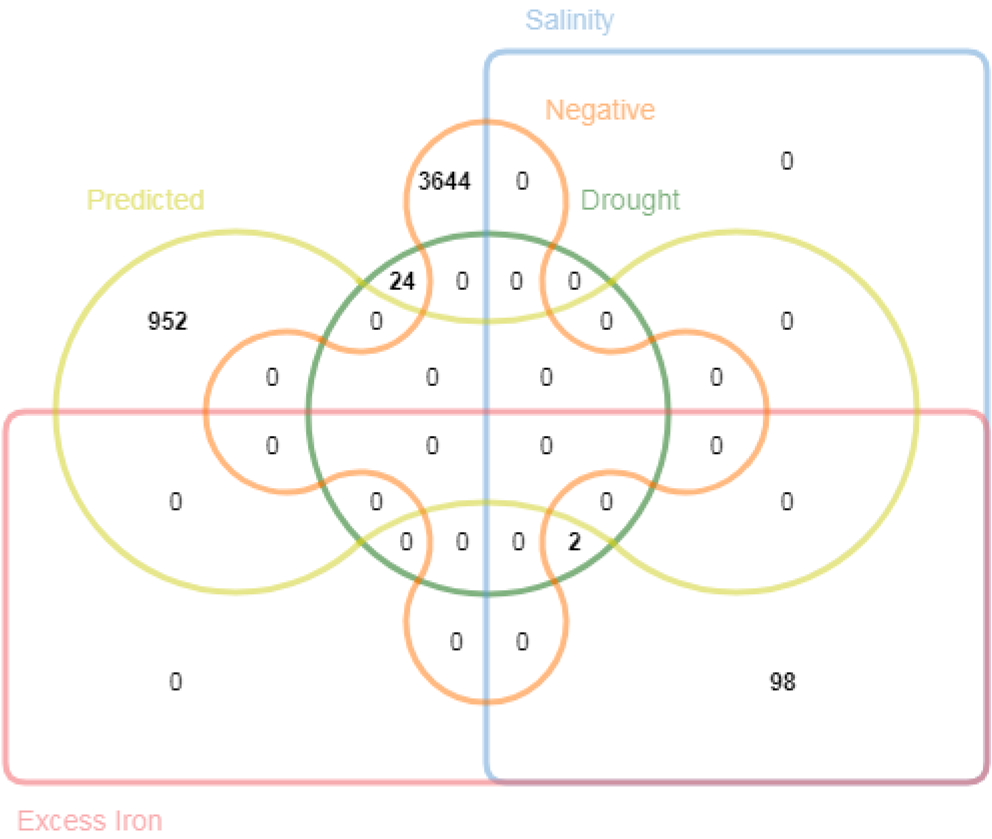


Figure 2: List of positive (stress related) and negative (housekeeping i.e. uniformly expressed under any condition). Venn diagram is created using jVenn web tool [56]. Red: Excess iron; Cyan: Salinity; Green: Drought; Yellow: Predicted stress genes; Orange: Negative or Housekeeping; Least common gene list is observed here.

Building positive set

A positive set of stress associated genes in rice has been collected from several published articles. Recently, Lakra et al. have reported 190 salinity stress-related proteins which are identified from proteomics study of contrasting rice genotypes [49]. In another study by Finatto et al., total 2525 rice genomic loci have been reported to be influenced by excess iron stress [25]. A well-known database that contains curetted drought stress-related genes for a number of plant species including *Arabidopsis*, rice etc. has been developed by Alter et al which contains 30 drought stress-related genes for rice [50]. Finally, 2692 genomic loci have been selected as gold standard (Set A) stress gene set. Many exclusive theoretical studies have been performed to identify plant stress-related genes which include predictions using orthology by Prabha et al. [52], transcription factor interaction network [51,54], gene expression analysis [57] etc. Total 952 genomic loci have been collected from aforementioned theoretical studies to construct predicted stress-related genes (Set B) for further analysis.

Constructing negative set for stress (NSS)

To date, it is still difficult to identify housekeeping genes (HK) those are uniformly expressed under any conditions. Very few rice genes have been reported as housekeeping genes. But the number is very low for comparative analysis of repeat distributions. To overcome this inadequacy, predicted housekeeping genes from Rice Expression Database (RED) [58] have been used for comparative analysis. RED is a comprehensive database that stores gene expression profiles from 284 high-quality RNA-Seq experiments performed under various treatments and a wide range of growth stages. The database enlists a set of housekeeping genes that are uniformly expressed under any condition. On the basis of statistical parameter 'τ' [59] which indicates the extent of expression breadth, HK genes have been sorted in ascending order. To balance the positive and negative classes for unbiased prediction equal no. (As of positive gene) of top HK genes having low 'tau' value and does not overlap with the positive set, have selected as the negative set. List of the positive and negative gene lists have been given in the Supplementary Table file.

Results & discussions

Rice genomes and availability of annotations

As mentioned earlier, rice is a diploid species which has 12 pairs of chromosomes with size much larger than *Salmonella* chromosome (approximately 80 times). It is the first crop genome that is sequenced after successful sequencing of the flowering model plant *Arabidopsis* which is much smaller in size than *Oryza* (approximately one-third of rice genome). Initiatives have been taken from several institutes for sequencing and assembling the aforementioned rice species. Among them, Arizona Genomics Institute (AGI) have sequenced and assembled 5 *Oryza species* including *O. barthii*, *O. glaberrima*, *O. glumaepatula*, *O. nivara* and *O. punctata*. Sequencing and assembling of *O. brachyantha*, *O. indica*, and *O. rufipogon* have been done in the three different institutes of China namely Institute of Genetics and Developmental Biology (IGDB), Chinese Academy of Sciences (CAS), Beijing, Beijing Genome Institute (BGI) and Shanghai Institutes for Biological

Sciences, CAS respectively. Model rice species *O. sativa* has been sequenced and assembled collaboratively by *MISSISSIPPI STATE* University (MSU) and International Rice Genome Sequencing Project (IRGSP) members [60]. In the year 2000, India becomes a part of IRGSP and took the challenge of sequencing chromosome 11 jointly ventured by University of Delhi South Campus (UDSC), the National Research Centre on Plant Biotechnology (NRCPB) and the Indian Agricultural Research Institute (IARI), New Delhi. Recently, Indian scientists have developed markers and detected fragrance alleles in the rice varieties [61-62] .

Two major initiatives have been taken by MSU and Rice Annotation Project (RAP) for annotating the model species *Oryza sativa* after completion of its sequencing in 2004 [63] using various *ab initio* methods and EST/mRNA alignments to produce the final gene models. MSU gene models have been also generated using automatic annotation pipeline which includes several tools like FGENESH, Genemark, Genscan, GeneSplicer, tRNAScan etc. Complete annotation and cross-references between two gene models are available at RAP database (http://rapdb.dna.affrc.go.jp) which is adopted by several databases like NCBI, Gramene etc. Annotations for other rice species have been performed by the corresponding sequencing groups either individually or in collaboration with other groups. In summary, protein-coding gene annotations have been done by evidence-based MAKER-P genome annotation pipeline mainly. Finding annotations for non-coding RNA genes have been executed using tRNAScan program [64] and transposable elements have been detected using RepeatMasker tool. For the majority of the rice species, genome assemblies and annotations are not finalized yet which is reflected from the presence of contigs in the sequences and very few number of RNA associated genes in the annotations compare to *O. sativa* and *O. indica* (Table 3).

Table 3: Availability of protein-coding, RNA associated genes and transcripts from Ensembl Plants release 37 (http://plants.ensembl.org/)

No.	Species	Protein Coding Genes	RNA associated genes	Transcripts
1	*Oryza barthii*	34575	3212	44891
2	*Oryza brachyantha*	32038	2560	34598

3	*Oryza glaberrima*	33164	41415	74579
4	*Oryza glumaepatula*	35735	3967	50860
5	*Oryza sativa indica*	40745	48978	89723
6	*Oryza nivara*	36313	2428	50788
7	*Oryza punctata*	31762	6029	47089
8	*Oryza rufipogon*	37071	3412	51004
9	*Oryza sativa japonica*	35679	56313	98663

Comparison of genomic parameters in Oryza species

Model species *Oryza sativa* has the genome of size 373.25 Mbp and average %GC content of 43.6. As shown in Figure 3A-3B, variations in genome sizes and % GC content indicate its ability to adapt to diverse environmental conditions [65]. In spite of low variance in gene density per kbp across 9 *Oryza* species (variance 0.21) (Figure 5.3C) percentages of genic regions are highly varying (Figure 3D). %Exonic and %CDS regions are also varying across 9 rice species (Figure 3E-3F). Model species *O. sativa* has the 0.244 genes per kbp of the genome which is higher than all other species except *O. glaberrima.* Another important observation is that all the domesticated rice species i.e. *O. glaberrima*, *O.indica*, *O. punctatata* and *O. sativa* have lower percentages of the genic region than their wild wild progenitors which include African wild species *O. barthii*, *O. brachyantha*, South American wild species *O. glumaepatula* and Asian wild rice *O. nivara* and *O. rufipogon*. This might be due to the fact of reduction of transposable elements in genic regions across the domesticated species [66]. However, all of the species possess comparable percentages of exonic and protein coding regions.

Different classes of repeats and their distributions in chromosomes

Both perfect and imperfect tandem repeats are abundant in rice chromosomes covering more than 5% of the genome which include both microsatellites and minisatellites present inside genes, intergenic regions and transposable elements. Model species *O. sativa* possess maximum amount of repeats (D_L ~ 6.4%) compare to other rice species followed by *O.*

rufipogon (D_L ~ 5.9%) and *O. indica* (D_L ~ 5.7%). Similarly, in the same species, *O. sativa* maximum numbers of repeats are found to occur per Mbp (D_N ~ 4182 repeats/Mbp) followed

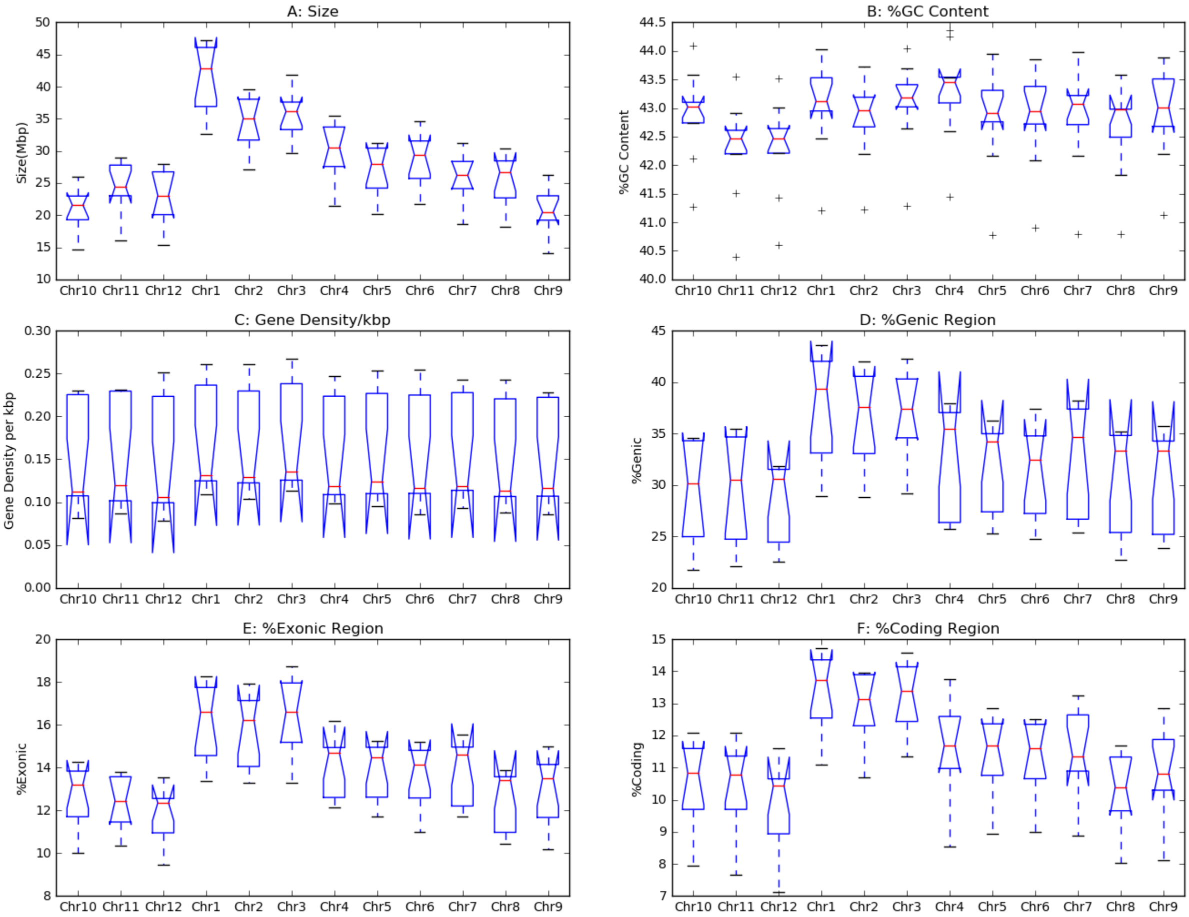


Figure 3: Comparison of genomic properties across *Oryza* species; A: Size in Mbp; B: %GC content; C: %Genic Region; D: Gene density per kbp; E: %Exonic Region; F: %Coding Region. Along X-axes of each subplot, chromosomes with number have been shown. For example, Chr1 abbreviates for chromosome number 1.

by *O. rufipogon* (D_N ~ 4060 repeats/Mbp), *O. barthii* (D_N ~ 3936 repeats/Mbp) and *O. indica*(D_N ~ 3929 repeats/Mbp) (Figure 4-5). On average, 3906 tandem repeats are occurring per Mbp of the rice genome. In case of interspersed repeats, only 1.5% of the genome is covered with this kind of repeats (median D_L ~ 1.51%) and their number of occurrences is much lower than that of tandem repeats (median D_N ~ 30 repeats/Mbp). Assembling and finishing off the rice genomes have not fully completed yet (95% coverage) and contains gaps and uncharacterized sequences. Hence, there remains a possibility that these numbers might change after gap filling and inclusion of uncharacterized parts into rice chromosomes as seen in the recently assembled rice genome with 99% coverage [40].

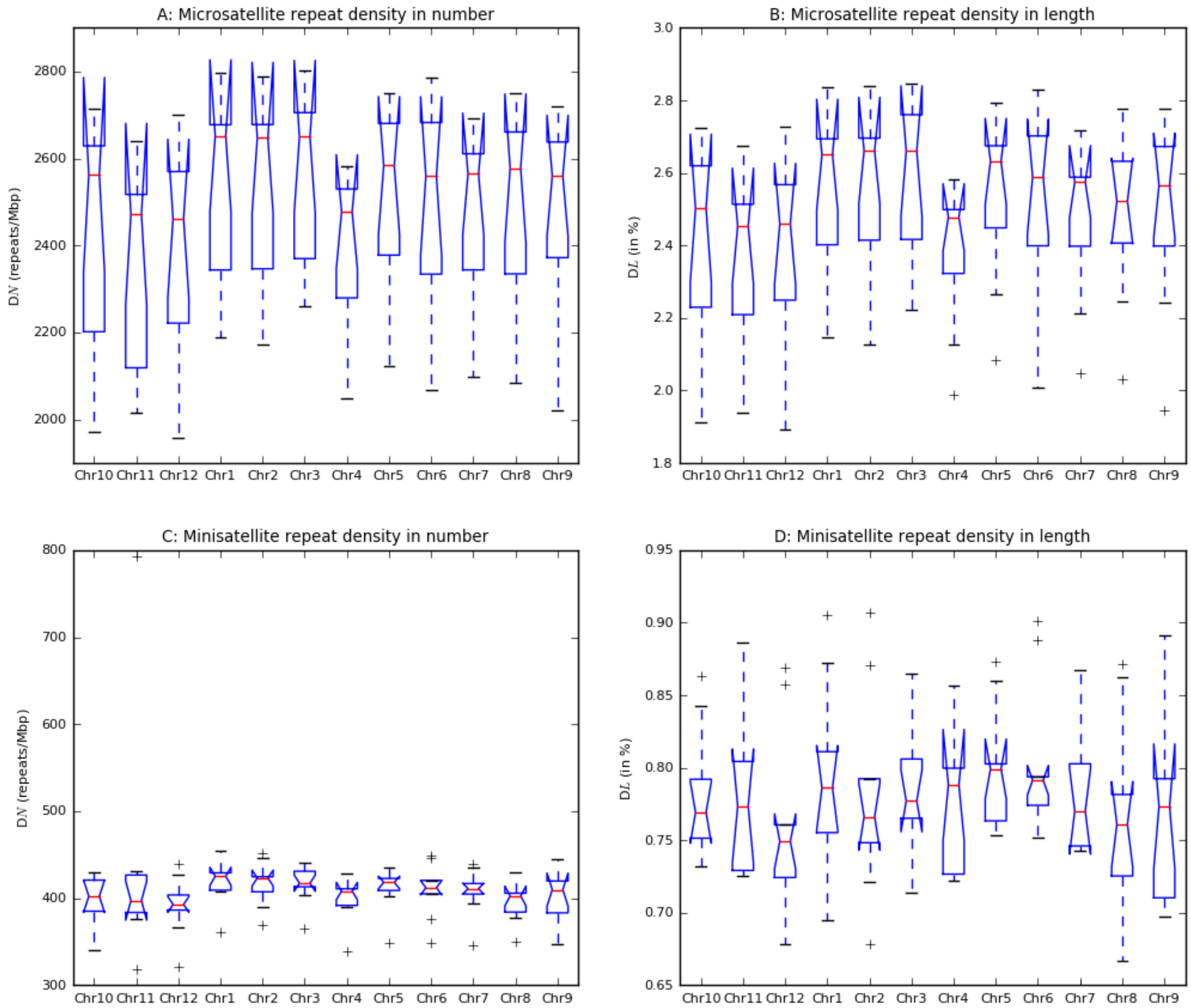


Figure 4: Comparing distributions of different classes of perfect repeats in 9 *Oryza* genomes. A: Microsatellite repeat density in number (D_N); B: Microsatellite repeat density in length parameter (D_L); C: Minisatellite repeat density in number (D_N parameter) and D: Minisatellite repeat density in length parameter (D_L).

Microsatellites

Microsatellites are the most predominant class of repeats in rice. On average more than 2500 perfect microsatellites are present per Mbp of rice genome. *O. sativa* has the highest D_N for perfect repeats (D_N ~ 2728 repeats/Mbp) and *O. punctata* has the lowest number (D_N ~ 2097 repeats/Mbp). Among the 9 species, chromosome 1 has the highest D_N value (median 2652 repeats/Mbp) whereas chromosome 12 has the least (median 2461 repeats/Mbp) (Figure 4A). In terms of D_L parameter, approximately 2.5% of genome is covered by perfect microsatellites with very low variation across species (< 0.05). Their abundance is also equal

to 2.5% of a chromosome (variance of 0.006 across chromosomes) (Figure 4B). *O. sativa* followed by *O. rufipogon* have more than 2.7% of their genome covered by perfect microsatellites whereas *O. punctata* has the least value of coverage i.e. 2.04%. Imperfect microsatellites have occurred less frequently than perfect ones (D_N ~ 690 repeats/Mbp) with large variations across species (Figure 5A). Approximately 1.2% of the rice genome is covered by imperfect microsatellites with a variance lower than 0.02. The coverage of imperfect microsatellites in each chromosome is 1.2% too and variance is below 0.001 (Figure 5B). So, microsatellites are almost equally distributed in the chromosomes but highly varying in numbers and might be of functional importance.

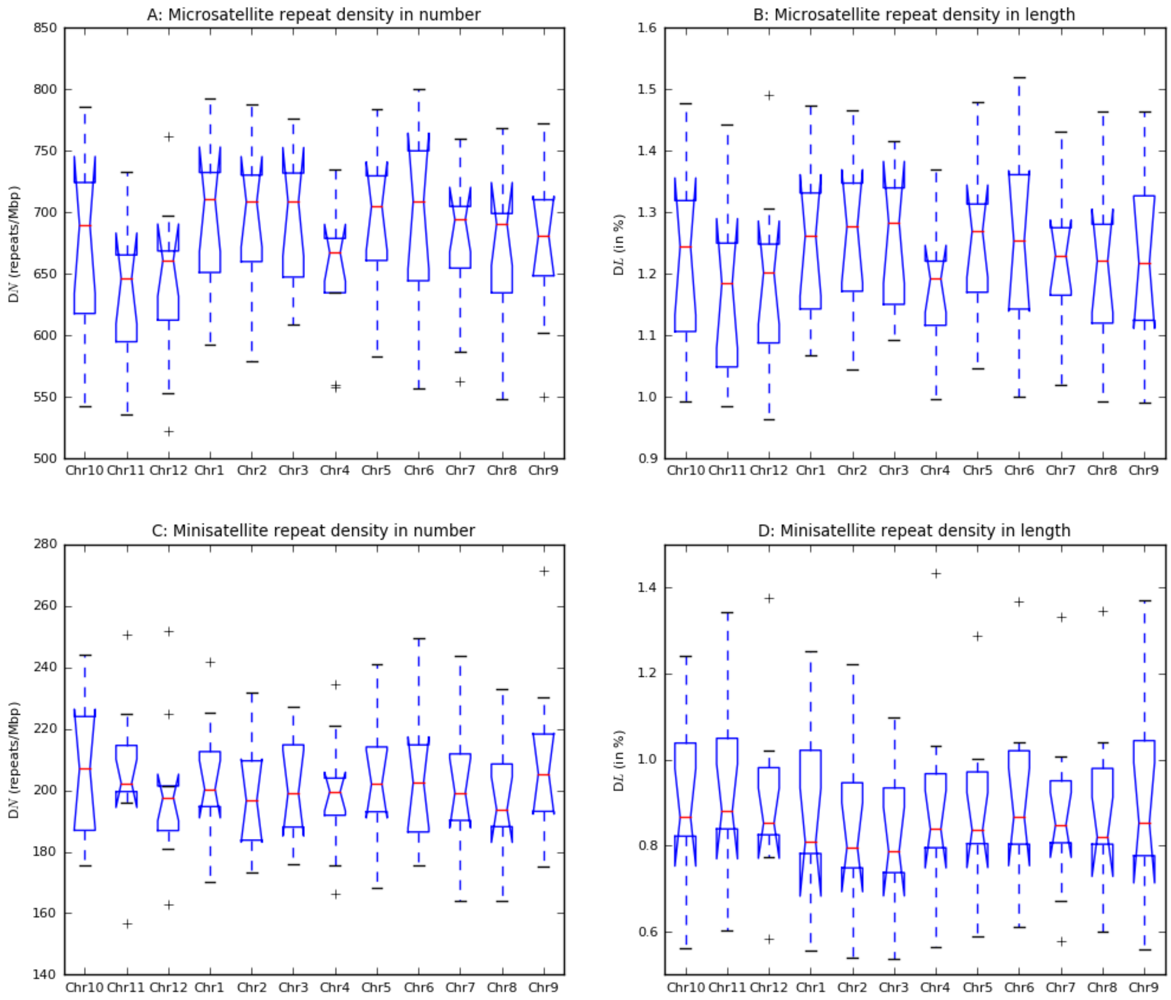


Figure 5: Comparing distributions of different classes of imperfect repeats in 9 *Oryza* genomes. A: Microsatellite repeat density in number (D_N); B: Microsatellite repeat density in length parameter (D_L); C: Minisatellite repeat density in number (D_N parameter) and D: Minisatellite repeat density in length parameter (D_L).

Minisatellites

Minisatellites are almost one fifth (in terms of numbers) of the abundance of the microsatellites. Perfect minisatellites are frequent (median D_N ~ 413 repeats/Mbp) compare to imperfect ones (median $D_N < 200$ repeats/Mbp) (Figure 4C-5C) but covers little less portion of the genome than polymorphic minisatellites (mean D_L ~ 0.8% and 0.9% respectively) (Figure 4D-5D). *O. sativa* has the maximum numbers of minisatellites repeats per Mbp whereas *O. rufipogon* has more minisatellites coverage in the genome. Like perfect microsatellites, perfect minisatellites are also equally distributed across chromosomes (mean D_L ~ 0.7% per chromosome and variance ~ 0.0001) and not varying too much across species (variance < 0.05). Similar observations have been found in case of imperfect minisatellites also. In terms of D_N parameter, perfect minisatellites are little less varying across species compare imperfect ones. So, tandem repeats are equally distributed across chromosomes but like microsatellites minisatellites are also varying in number per Mbp of the genome.

Interspersed repeats

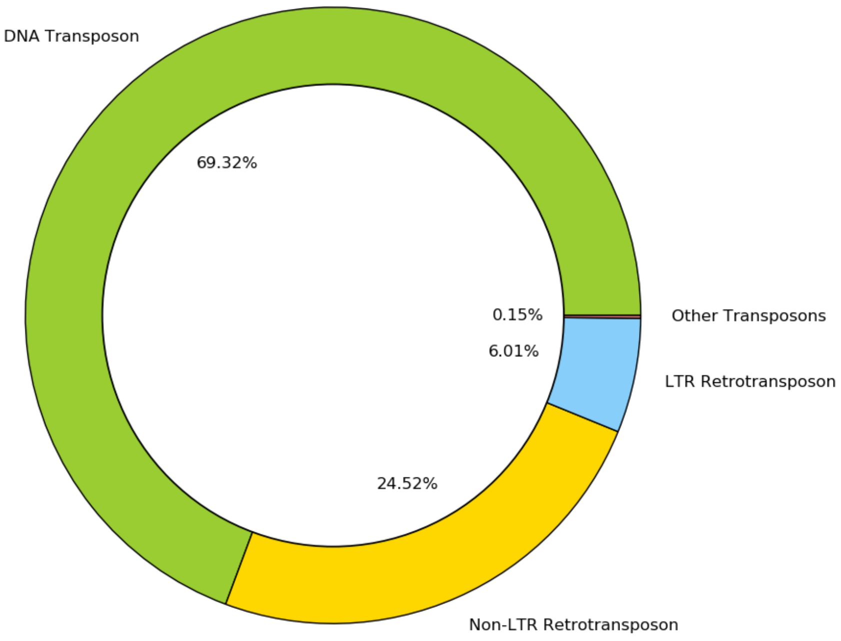


Figure 6: Different classes of Interspersed Repeats in *Oryza sativa* in Repbase v20.05.Green: DNA Transposons; Blue: LTR Retrotransposons; Yellow: Non-LTR Retrotransposons; Black: Other transposons.

As mentioned earlier in the materials and method section, interspersed repeats are collected from Repbase which is the reference database used by the tools for identifying interspersed repeats in the eukaryotic genomes. This particular database is chosen because it is a curetted, regularly updated resource and has utilized a specialized scheme for classifying and annotating transposable elements which constitutes the major part of the interspersed repeats in any eukaryotic species. In the database, transposable elements are listed for the model species *Oryza sativa* only (Figure 6). Using programs like Blast, these reference sequences can be mapped to the other rice species for finding highly homologous transposable elements that can perform similar functions across rice species. Transposable elements can occur in multiple copies in a genome. Hence, '-max_target_seqs' and'-max_hsps' parameters of blastn program have been set as default i.e. 500 and all HSPs. But to find highly similar sequences with highly similar functions, more than 95% sequence identity and query coverage with an expected value of 10^{-6} filters have been applied. Though, it have filtered and reported much fewer amounts of transposable elements (maximum D_L of 7.75% in *O. sativa*) than earlier reports (~37.5%) but the extracted elements are highly reliable for stress association analysis and prediction (Figure 7).

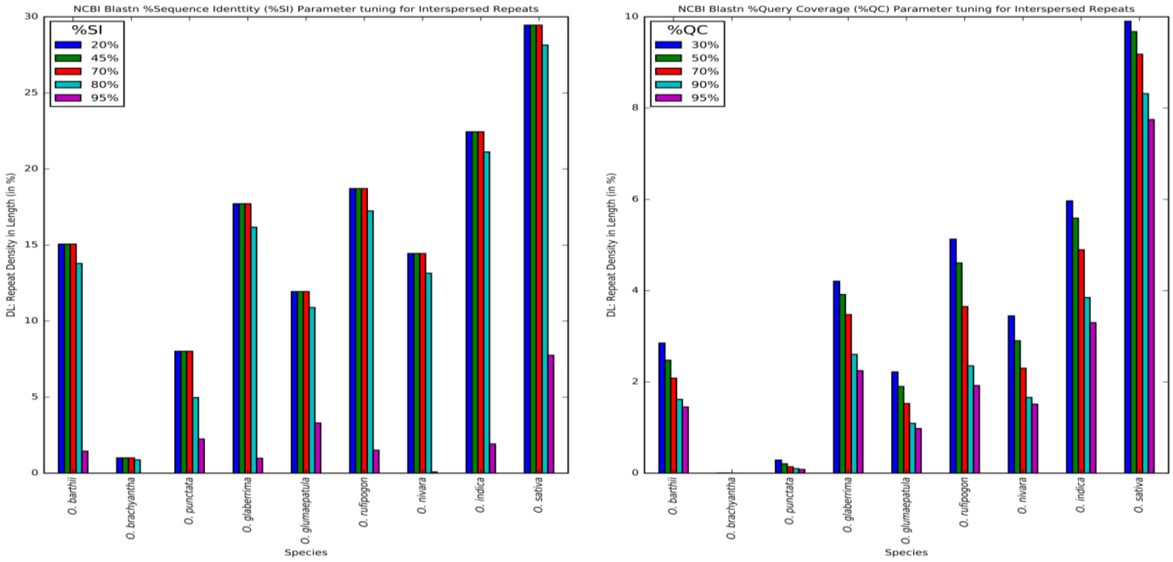


Figure 7: Comparison of blast parameters for interspersed repeat extraction from rice species. A: Percent sequence identity and B: Percent query coverage

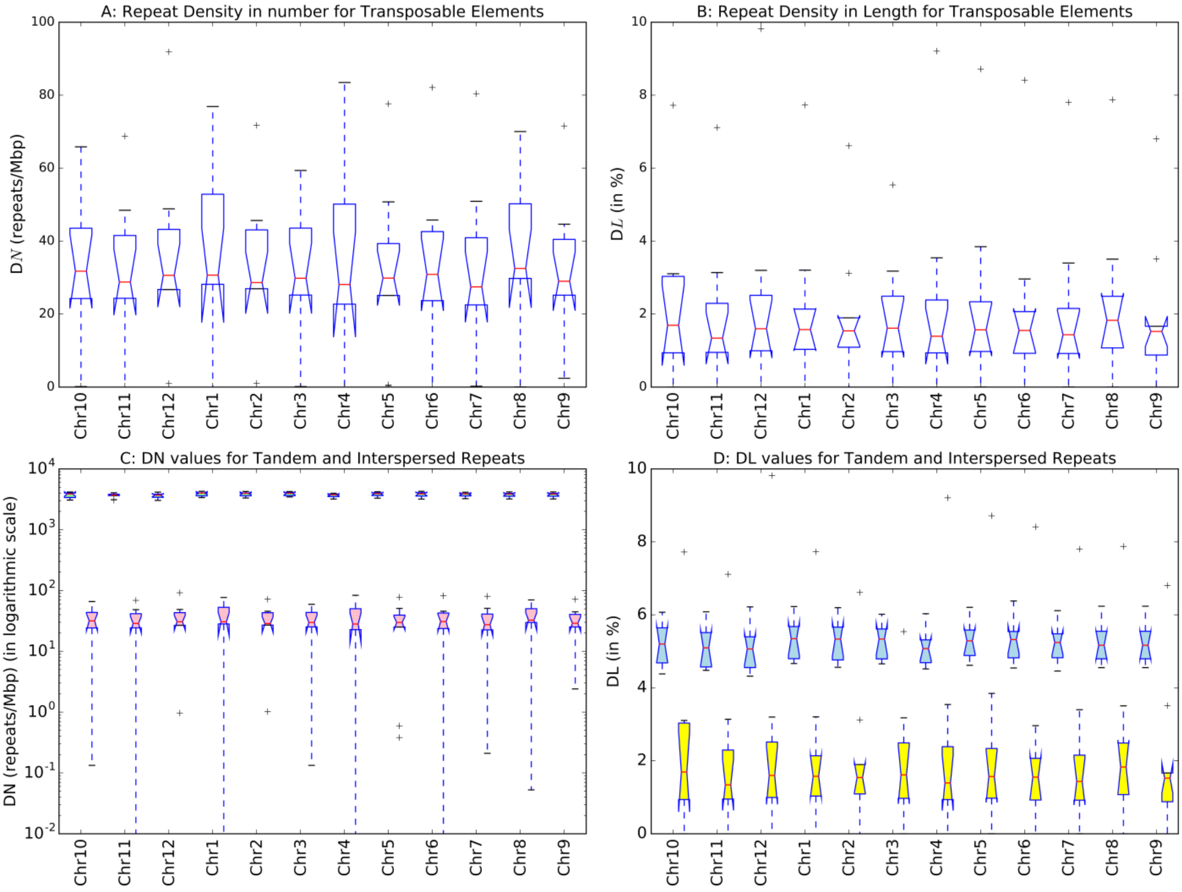


Figure 8: Distribution of Interspersed Repeats in Oryza sativa and comparison of Tandem and Interspersed Repeats in Oryza sativa on the basis of repeat density in number parameter. Pink and yellow are interspersed and green and blue are for tandem repeats

As expected, *O. sativa* possess maximum number of transposable elements (D_N ~ 75 repeats/Mbp) and *O. brachyantha* has less than 1 repeat/Mbp (Figure 8A-8B). Effect of extreme strict criteria for the selection of highly reliable interspersed repetitive elements is reflected in the current result.

Statistically significant repeats

In the frequency analysis of tandem repeats of motif lengths ranging from 1-100 bp including both perfect and imperfect repeats, the very first observation is that the perfect tri-meric repeats are occurring significantly (1sample t-test p-value < 0.05) more frequently than the other motifs followed by monomeric and di-meric repeats. Most striking observation is that perfect minisatellites do occur in the real genomes but absent in the random genomes as expected (Figure 9-10 and supplementary Figure 1A and Figure 1B). All repeats of motif lengths 1-6 bp are occurring significantly the rice genomes (1sample t-test p-value < 0.05). Imperfect repeats of motif lengths 1 to 100 nucleotides are present in the rice genomes but with less number per Mbp (Figure 11-12) than perfect ones and also statistically significant (1sample t-test p-value < 0.05). Unlike perfect minisatellites, imperfect minisatellites have been occurred significantly different from their occurrence in the random genomes. Both perfect and imperfect minisatellites are less frequent than the occurrence of microsatellites.

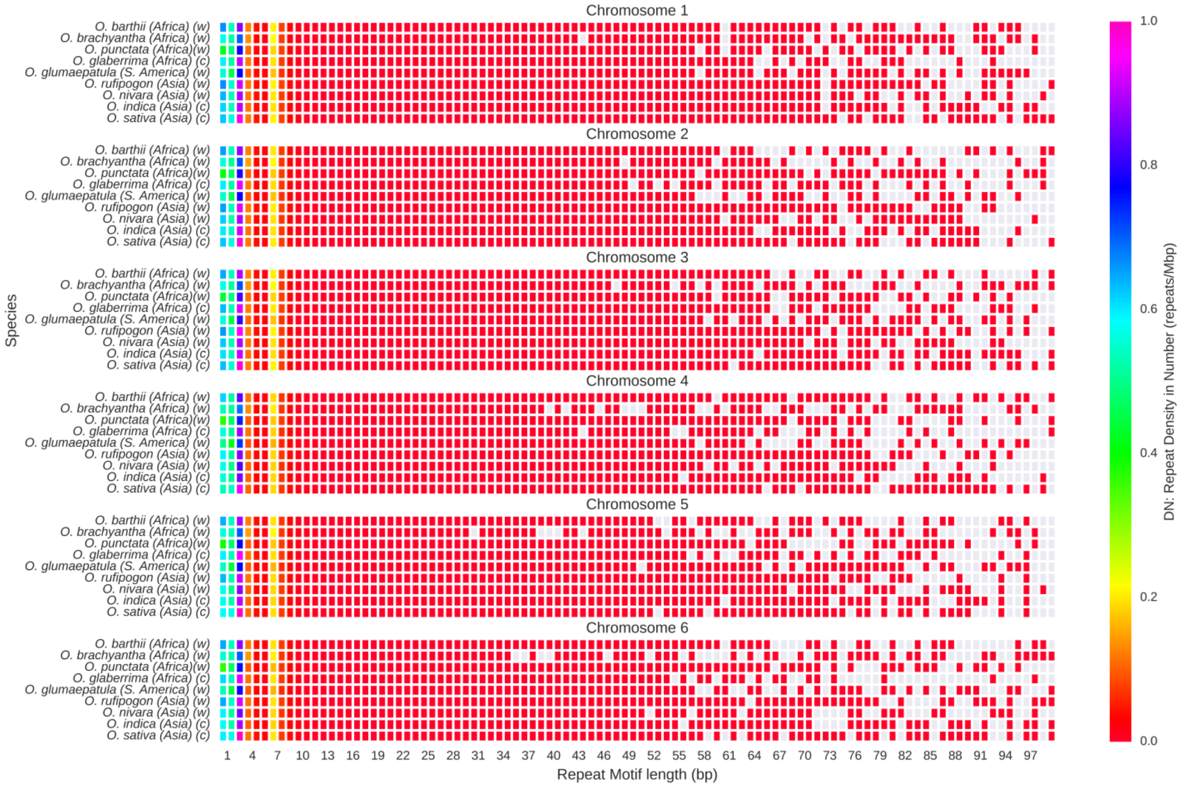


Figure 9: Distribution of perfect repeats with different motif lengths (1-100) across *Oryza* genomes. Chromosomes 1 to 6 have been shown here. Color bar represents repeat density in number (D_N parameter in repeats/Mbp) parameter. From red to magenta denotes lower values to higher values.

Heptameric repeats are more ubiquitous than the other minisatellites ranging from length 8 bp to 100 bp. These observations signify that trimeric repeats have ubiquitously occurred in the genomes compared to other microsatellites [67]. A major portion of the rice genome is not the coding region. So, the presence of trimeric repeats in excess is not only due to their relation with codons but also they must have some definite functional importance. Several earlier reports have shown that both micro- and minisatellites regions are prone to methylation leading to epigenetic regulations of genes and genomes and their crucial roles in various developmental processes such as the development of seeds, embryos and gametophytes, controlling flowering time and stress responses [68].

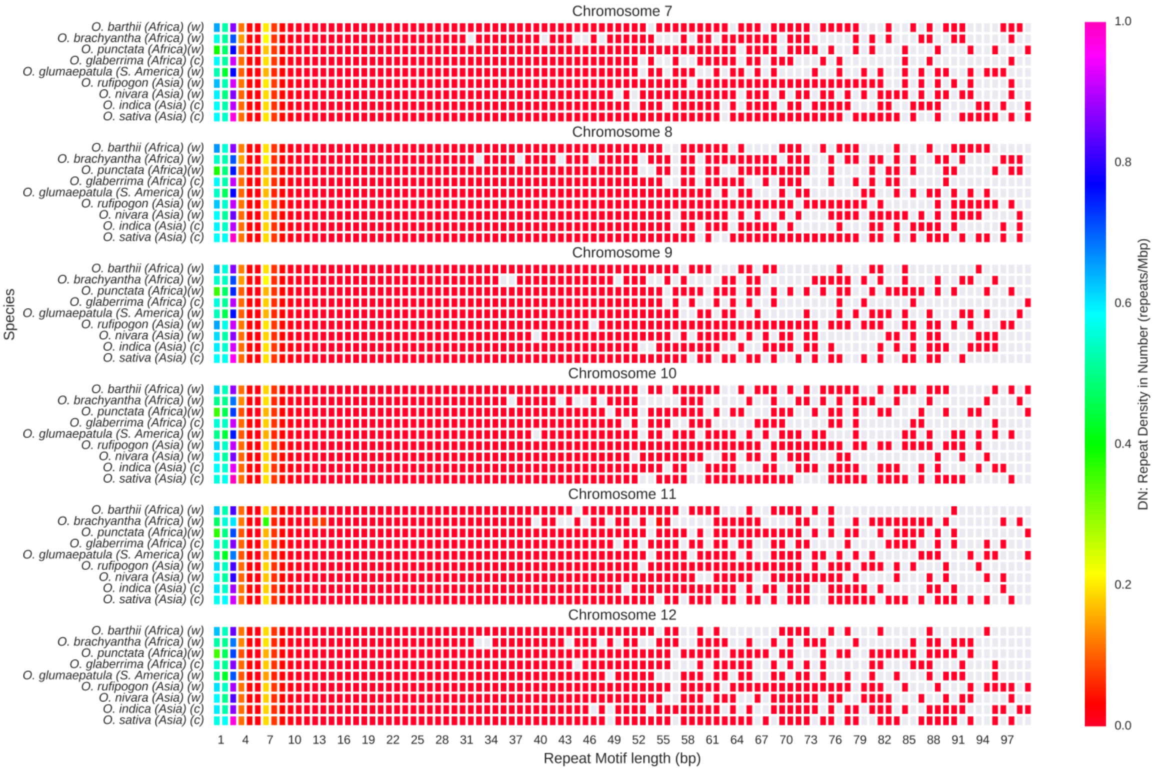


Figure 5: Distribution of perfect repeats with different motif lengths (1-100) across Oryza genomes. Chromosomes 7 to 12 have been shown here. Color bar represents repeat density in number (DN parameter in repeats/Mbp) parameter. From red to magenta denotes lower values to higher values.

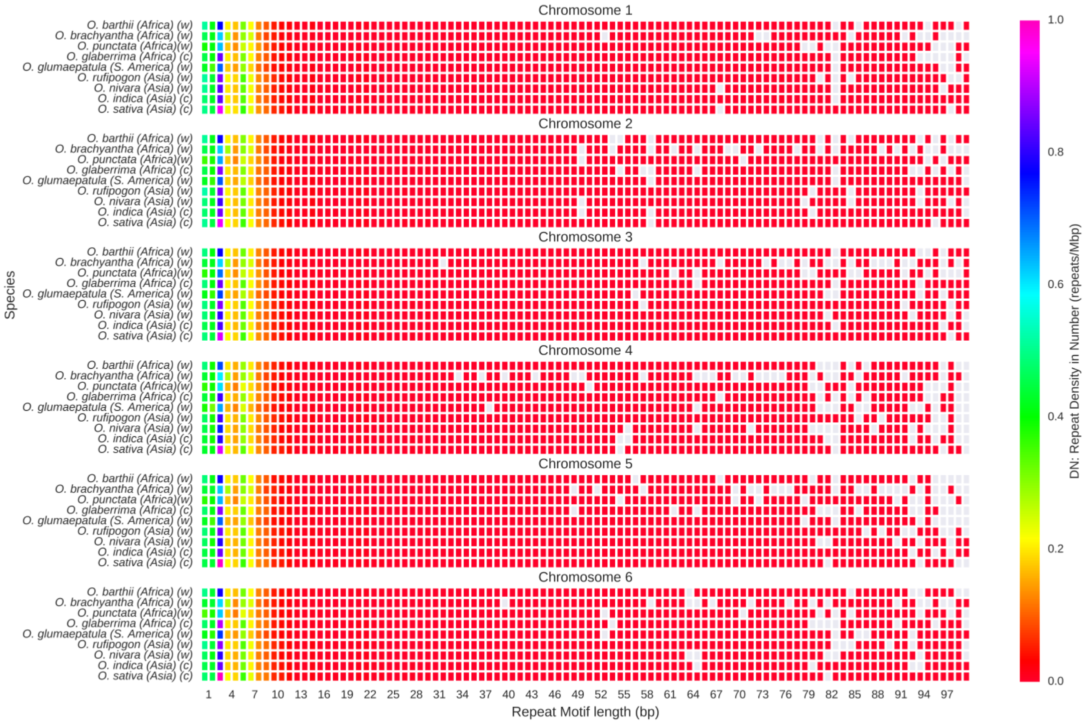


Figure 11: Distribution of imperfect repeats with different motif lengths (1-100) across *Oryza* genomes. Chromosomes 1 to 6 has been shown here. Color bar represents repeat density in number (D_N parameter in repeats/Mbp) parameter. From red to magenta denotes lower values to higher values.

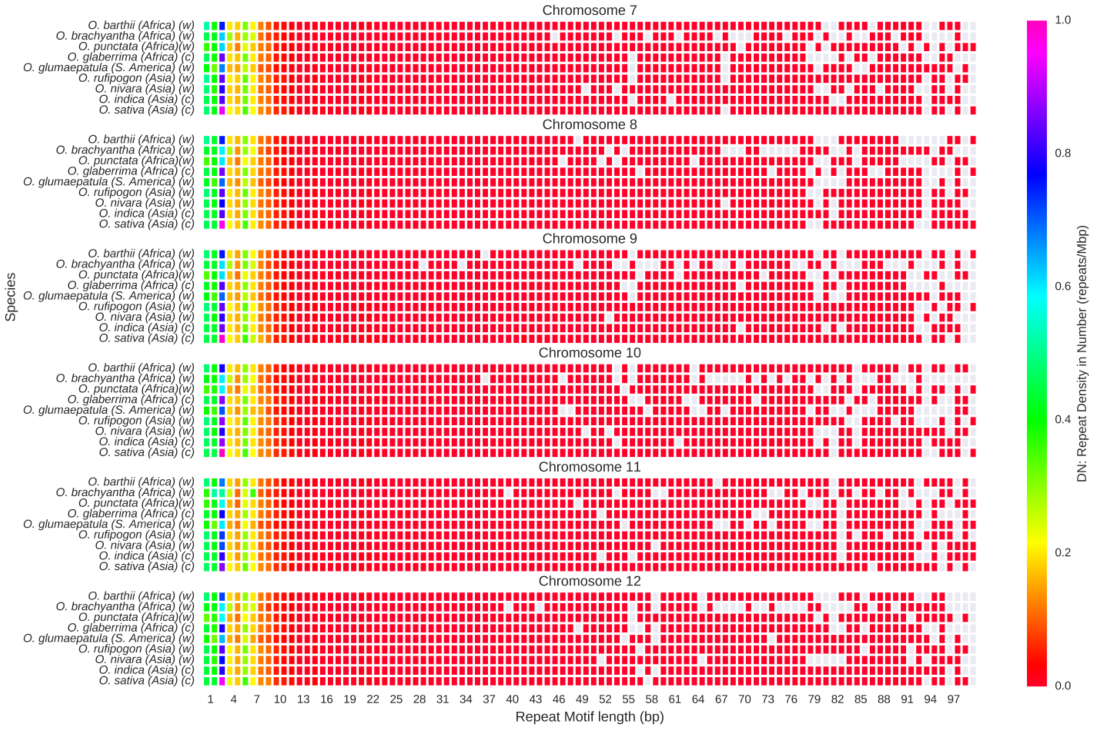


Figure 12: Distribution of imperfect repeats with different motif lengths (1-100) across *Oryza* genomes. Chromosome 7 to 12 has been shown here. Color bar represents repeat density in number (D_N parameter in repeats/Mbp) parameter. From red to magenta denotes lower values to higher values.

Occurrence of repeats in different genomic locations

As shown in series of Figures 13-18, perfect and imperfect tandem repeats are ubiquitous in different genomic locations of rice genomes including genic, intergenic, exons, introns, coding and untranslated regions.

Repeats in genic and intergenic regions

On average, 884 perfect microsatellite repeats/Mbp are occurring in the genic regions which are almost half than that of intergenic regions (median value 1640 repeats) (Figure 13A). *O. rufipogon* has the maximum number of repeats (D_N ~ 1037 repeats/Mbp) followed by *O. nirvara* (D_N ~ 1018). *O. sativa* has the D_N value of 884 repeats/Mbp but contains the maximum number of perfect microsatellites in the intergenic regions (D_N ~ 1909 repeats/Mbp).

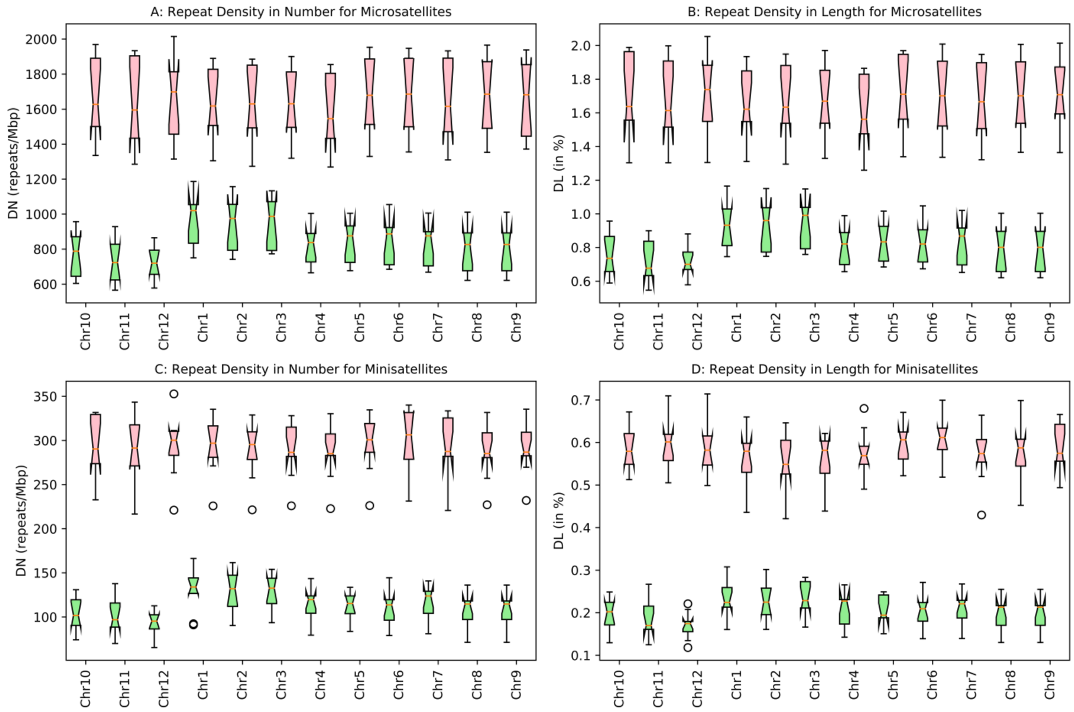


Figure 13: Chromosome-wise distribution of Perfect Tandem Repeats (Microsatellites and Minisatellites) in Genic (Green) and Intergenic regions (pink)

Perfect microsatellites in the genic and intergenic regions have covered approximately 0.8% and 1.6% of the genome which are also uniformly distributed across the chromosomes (variance < 0.006) (Figure 13A-13B). Perfect minisatellites are also frequent in the intergenic regions (median D_N ~ 296 repeats/Mbp) compare to genic regions (median D_N ~ 114 repeats/Mbp) and their numbers are varying across chromosomes too (Figure 13C-13D). In terms of repeat density in length parameter, genic and intergenic repeats cover approximately 0.2% and 0.6% of the genome respectively. Like microsatellites, perfect genic and intergenic minisatellites are also equally distributed in the chromosomes.

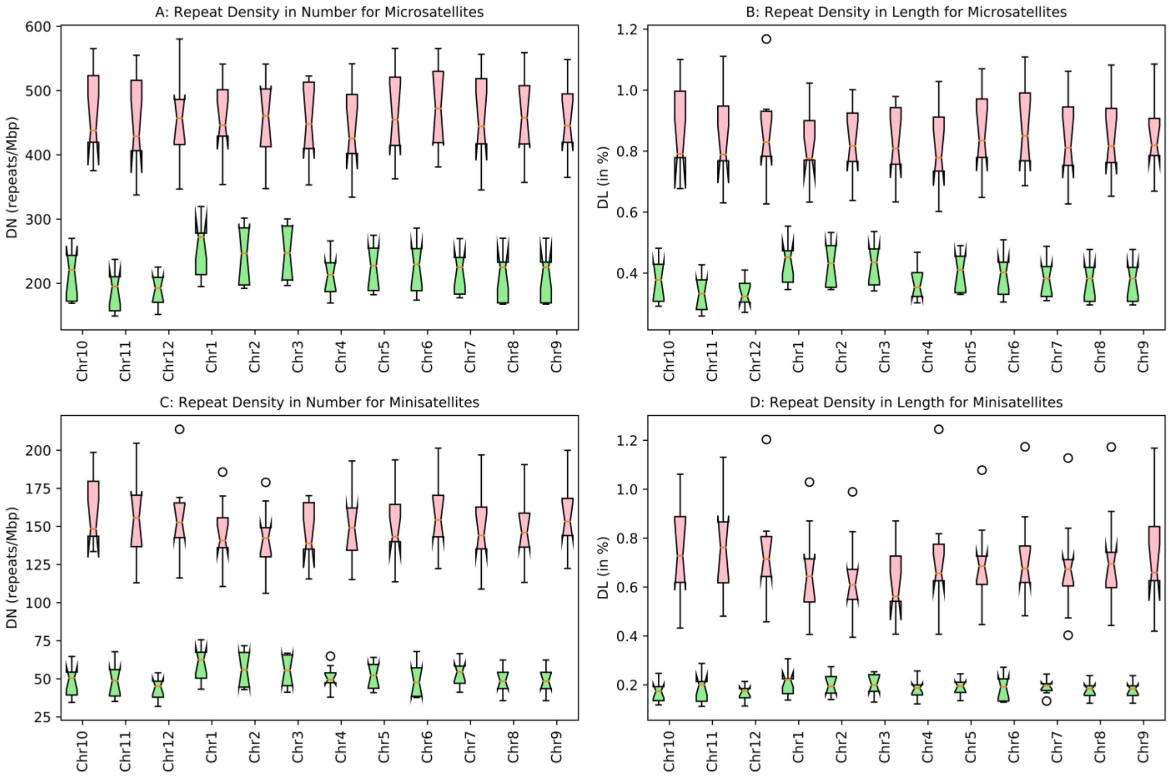


Figure 14: Chromosome-wise distribution of Imperfect Tandem Repeats (Microsatellites and Minisatellites) in Genic (Green) and Intergenic regions (pink)

Unlike perfect microsatellites in the genic and intergenic regions, polymorphic microsatellites are much less frequent in numbers per Mbp (median value of ~232 and ~450 repeats/Mbp respectively) covering ~0.4% and ~0.9% of the genome respectively having nearly equal contributions from each chromosome (Figure 14A-14B). Polymorphic minisatellites are less abundant than microsatellites in genic and intergenic regions (median D_N ~ 53 and 144 repeats/Mbp respectively) covering 0.2% and 0.7% of the genomes respectively (Figure 14C-14D). Almost equal contributions from each chromosome have been observed in this case also.

Repeats in exonic and intronic regions

Inside the genic regions, repeats are ubiquitously found both in the exonic and intronic parts of the genome. D_N values for perfect microsatellites in exons and introns are comparable (median D_N ~ 479 and 422 repeats/Mbp respectively) (Figure 15A) and covering

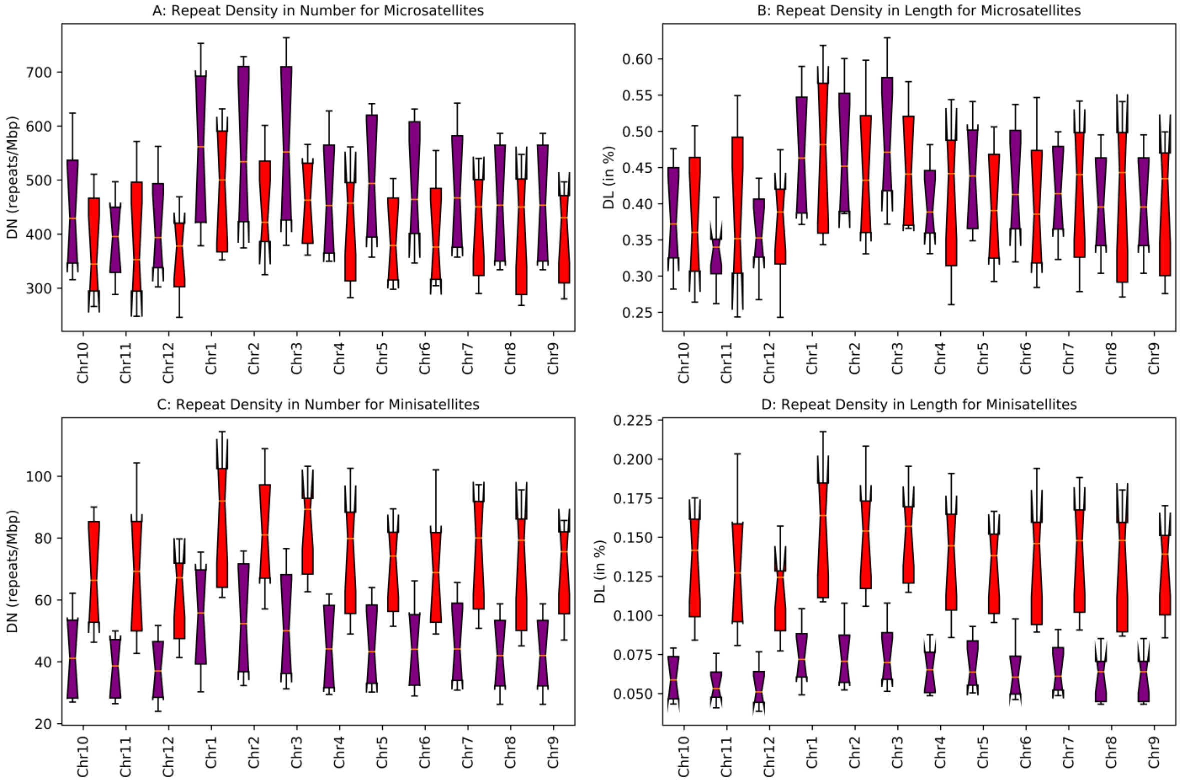


Figure 15: Chromosome-wise distribution of Perfect Tandem Repeats (Microsatellites and Minisatellites) in exons (Purple) and introns regions (Red)

approximately 0.4% of the genomes in both cases (Figure 15B). In the case of imperfect microsatellites, their occurrences are slightly greater in the exonic regions (median D_N ~ 140 repeats/Mbp) than intronic regions (median D_N ~ 100 repeats/Mbp) (Figure 16A). A similar observation has been found in case of repeat density in length parameter also. Imperfect microsatellites in the exonic regions cover only 0.23% of the genome which is little higher than their presence in the intronic region (D_L ~0.16%) (Figure 16B). Perfect minisatellites are more abundant in intronic regions (D_N ~ 80 repeats/Mbp) compare to exons (45 repeats/Mbp) (Figure 15C). Exonic perfect minisatellites cover only 0.06% of the genome whereas intronic repeats cover approximately 0.14% of the genome i.e. more than double than that of perfect ones (Figure 15D). Percentages of occurrences of imperfect minisatellites in the exonic and intronic region are comparable (D_L values ~ 0.07% and 0.11% respectively) (Figure 16D). Similar finding has been observed in case of repeat density in number parameter too. DN values for imperfect minisatellites in exons and introns are 24 and 30 repeats/Mbp respectively. (Figure 16C). Surprisingly, *O. nivara* has maximum number of perfect and imperfect microsatellites than the other species in both regions. *O. sativa* has maximum number of repeats per Mbp in the exonic regions and *O. rufipogon* possess the maximum values in intronic regions.

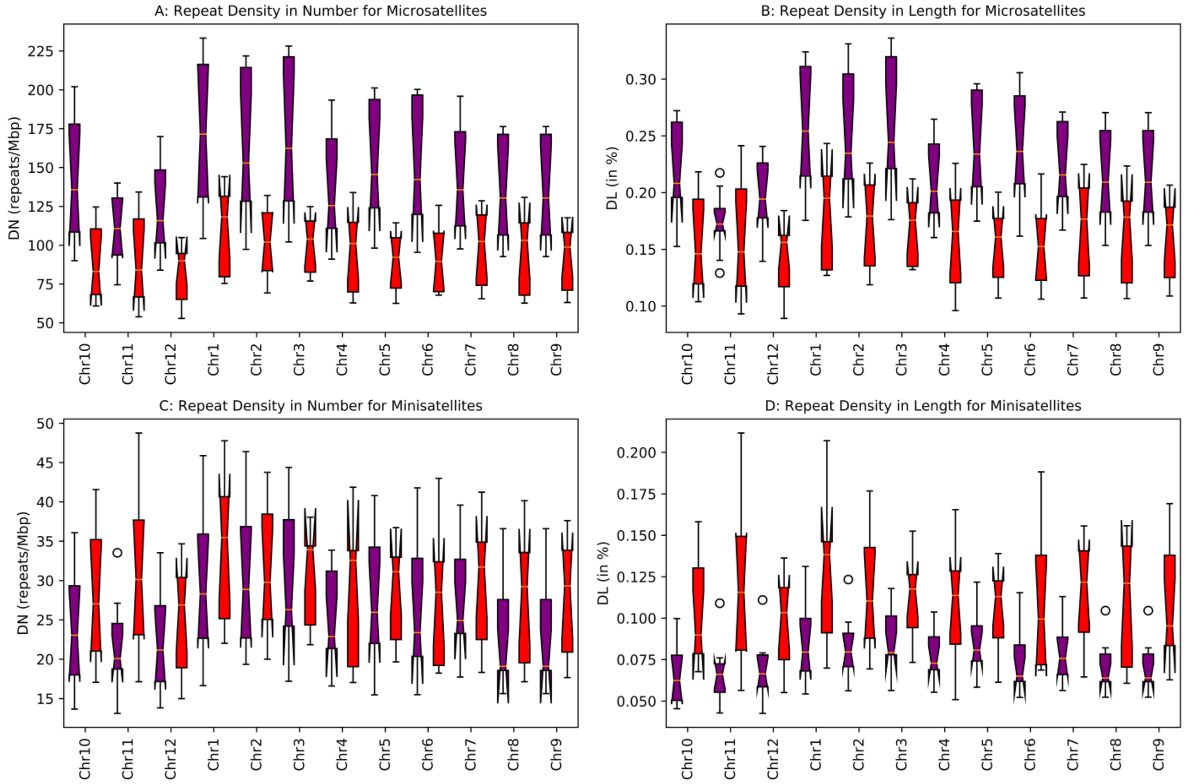


Figure 16: Chromosome-wise distribution of Imperfect Tandem Repeats (Microsatellites and Minisatellites) in exons (Purple) and introns regions (Red)

Repeats in coding regions and untranslated regions

Perfect microsatellite repeats more occur in the coding regions than the untranslated regions. Repeat density in number parameter (D_N) in the coding regions has the value more than 325 repeats/Mbp which is three times greater than the D_N value in untranslated regions (D_N ~ 95 repeats/Mbp) (Figure 17A). A similar observation has noticed in case of imperfect microsatellites also. D_N values of imperfect microsatellites in coding and untranslated regions are 102 and 28 repeats/Mbp respectively (Figure 18A). Like microsatellites, minisatellites are found more in the coding regions than untranslated regions. D_N values for perfect minisatellites in these regions are 24 and 14 repeats/Mbp respectively (Figure 17C) and in case of imperfect minisatellites, the values are 16 and 6 repeats/Mbp respectively (Figure 18C). While checking repeat density in length parameter (D_L), perfect coding region repeats cover more than 0.3% of the genome in contrast to repeats in the untranslated regions which has the D_L value of 0.09% (Figure 17B).

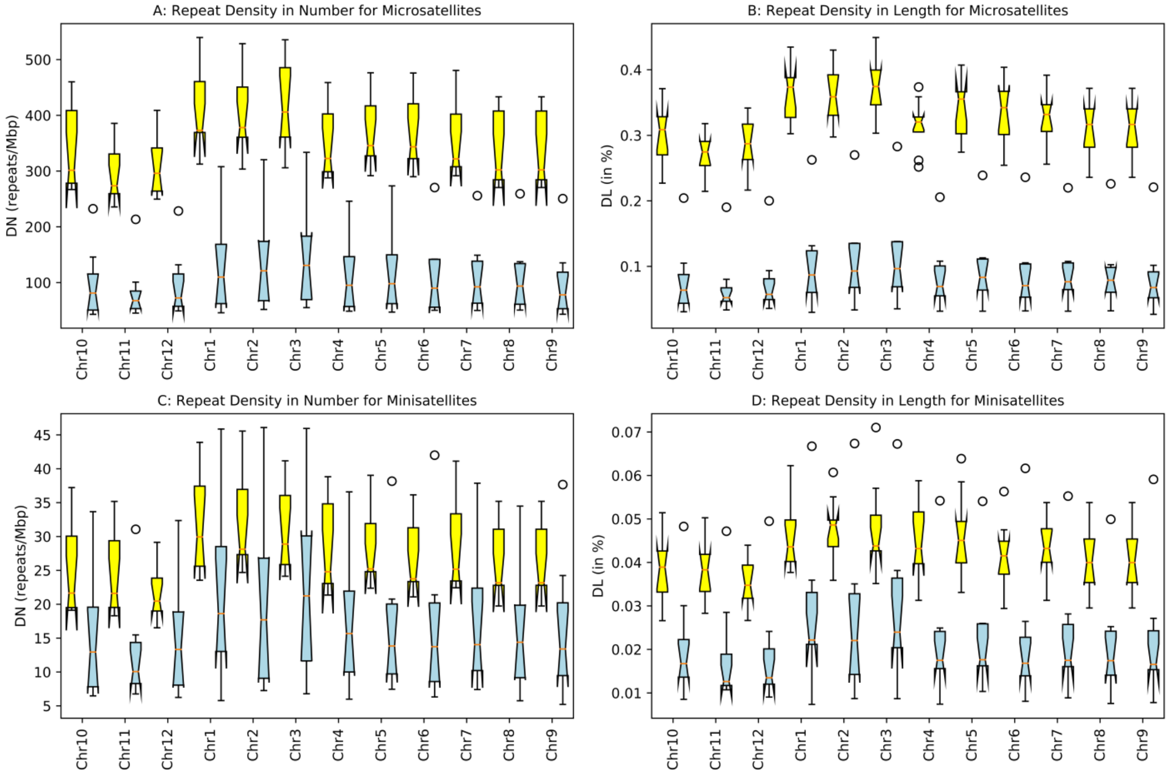


Figure 17: Chromosome-wise distribution of Perfect Tandem Repeats (Microsatellites and Minisatellites) in coding (Yellow) and untranslated regions (Blue).

Imperfect microsatellites occur very less in the untranslated region (D_L ~ 0.04%) compare to coding region (D_L ~ 0.18%) (Figure18B). Perfect minisatellites have been detected slightly more in the coding regions (D_L ~ 0.04) compare to untranslated parts of the genome (D_L ~ 0.02%) (Figure 17D). D_L value for imperfect minisatellites is moderately higher than perfect minisatellites (D_L ~ 0.06%) (Figure 18D). In comparison to coding regions, imperfect minisatellites in the untranslated regions cover only 0.02% of the genome. Regarding tandem repeats in coding and untranslated regions, *O. nivara* and *O. sativa* has the maximum number of perfect microsatellites in coding regions and untranslated regions respectively. One major observation is that the numbers of perfect microsatellites per Mbp in coding and untranslated regions in *O. sativa* genome are close to each other (D_N ~ 321 and 270 repeats/Mbp respectively) compare to the other species. In case of other genomes, as expected coding regions have more perfect microsatellites than untranslated region. Another prime observation is that the number of perfect minisatellites has occurred more in number in the untranslated regions than in the coding regions in case of *O. sativa* genome which might have some functional importance.

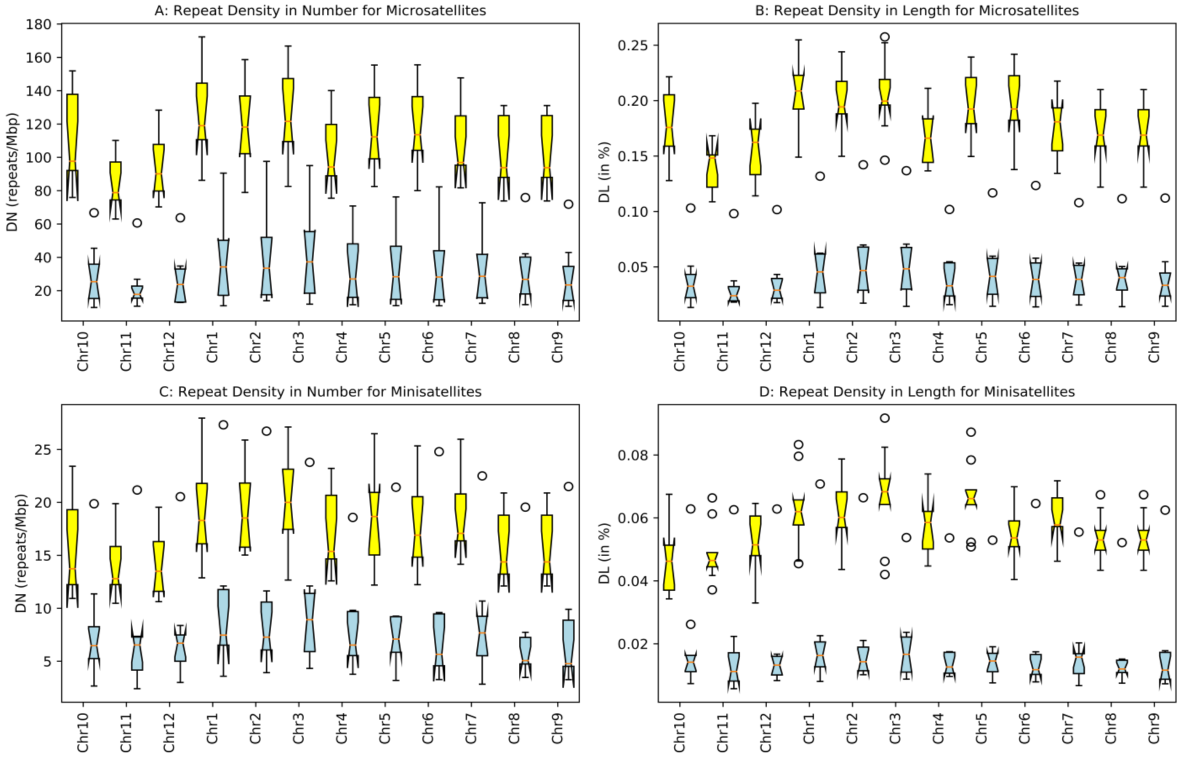


Figure 18: Chromosome-wise distribution of Imperfect Tandem Repeats (Microsatellites and Minisatellites) in coding (Yellow) and untranslated regions (Blue).

Applications: Analysis of Common repeats

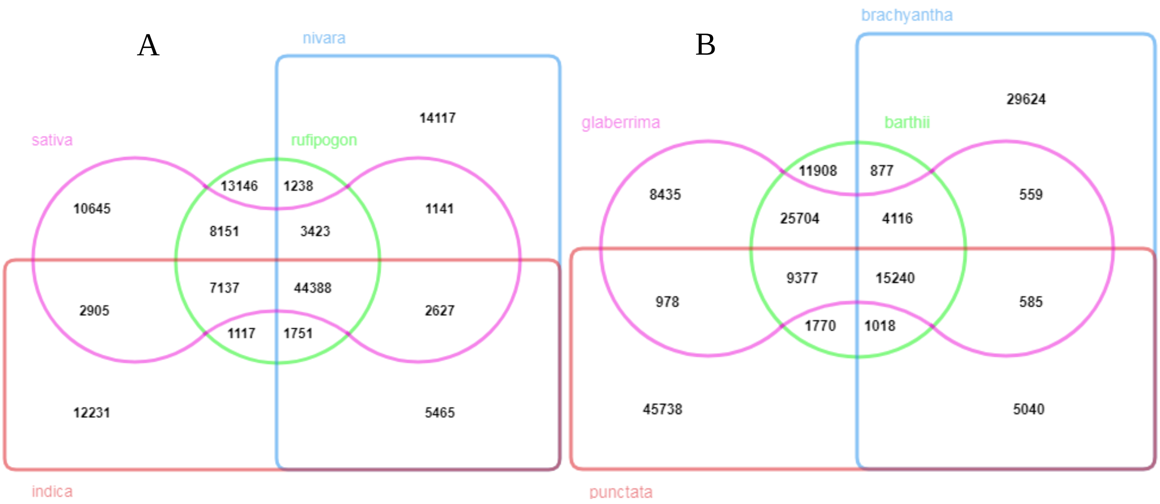


Figure 19: Venn diagram showing common perfect repeats in different species from A: Asia ; B: Africa . Pink: *O. sativa*, *O. glaberrima*; Blue: *O. nivara*, *O. brachyantha*; Green: *O. rufipogon*, *O. barthii*; Red: *O. indica*, *O. punctata*.

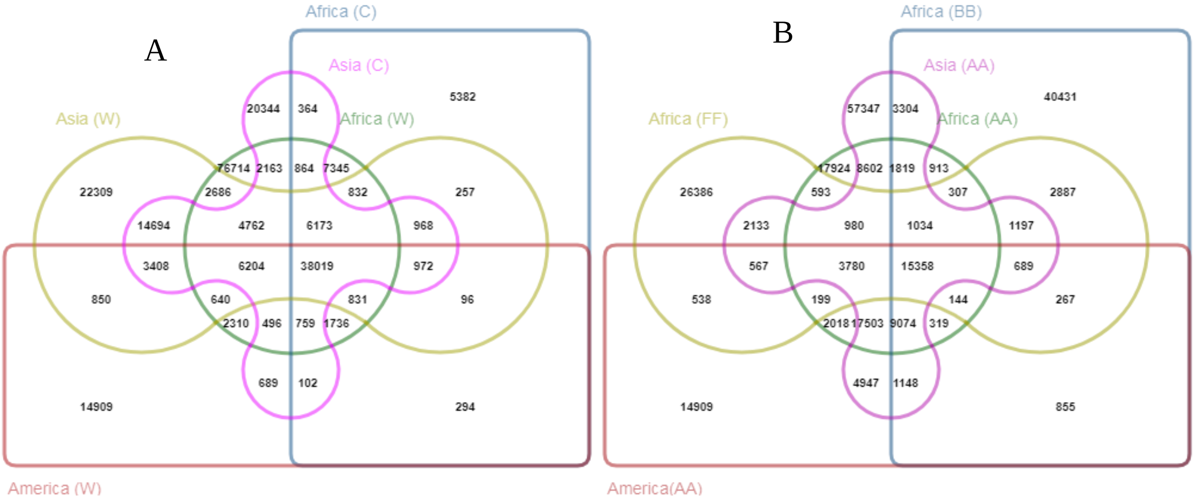


Figure 20: Venn diagram showing common perfect repeats in A: Wild and cultivated species from the different continents; B: Species from the different continents with different genome types. Pink: Asia (Cultivated), Asia (AA); Blue: Africa (Cultivated), Africa (BB);

Green: Africa (Wild), Africa (AA); Red: America (Wild), America (AA); Yellow: Asia (Wild), Africa (FF).

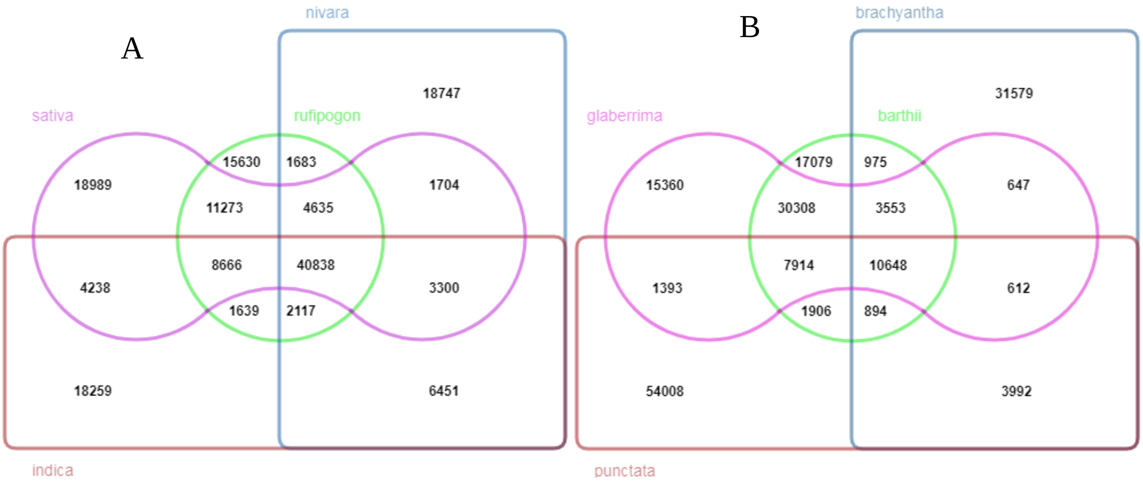


Figure 21: Venn diagram showing common imperfect repeats in different species from A: Asia ; B: Africa. Pink: *O. sativa, O. glaberrima*; Blue: *O. nivara, O. brachyantha*; Green: *O. rufipogon, O. barthii;* Red: *O. indica, O. punctata*.

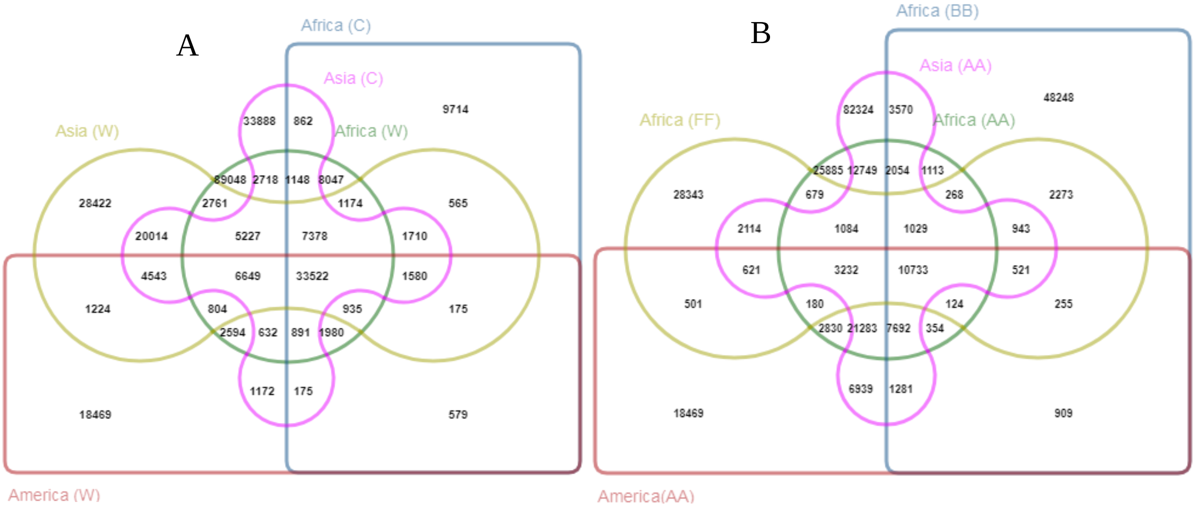


Figure 22: Venn diagram showing common imperfect repeats in A: Wild and cultivated species from the different continents; B: Species from the different continents with different genome types. Pink: Asia (Cultivated), Asia (AA); Blue: Africa (Cultivated), Africa (BB); Green: Africa (Wild), Africa (AA); Red: America (Wild), America (AA); Yellow: Asia (Wild), Africa (FF).

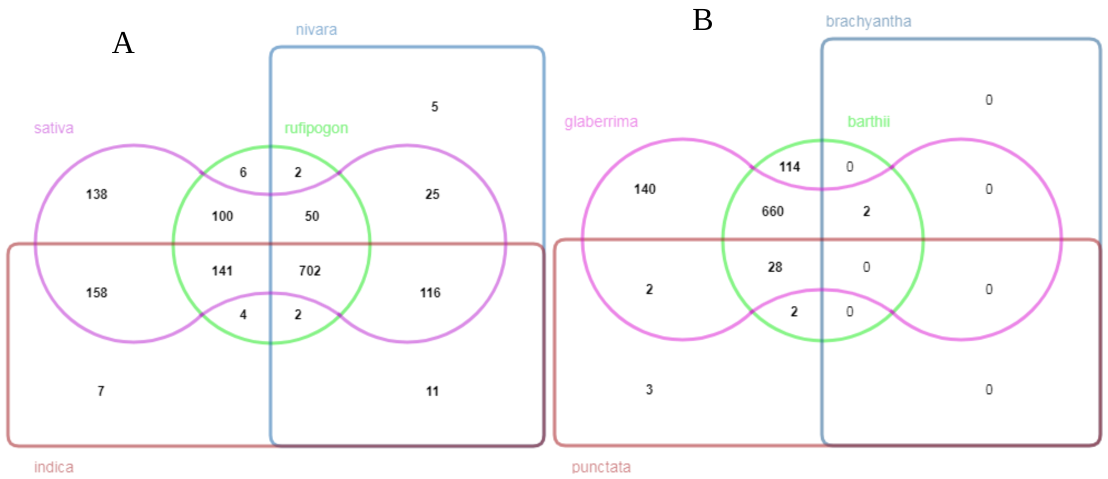


Figure 23: Venn diagram showing common interspersed repeats in different species from A: Asia ; B: Africa. Pink*: O. sativa, O. glaberrima;* Blue*: O. nivara, O. brachyantha;* Green*: O. rufipogon, O. barthii;* Red: *O. indica, O. punctata.*

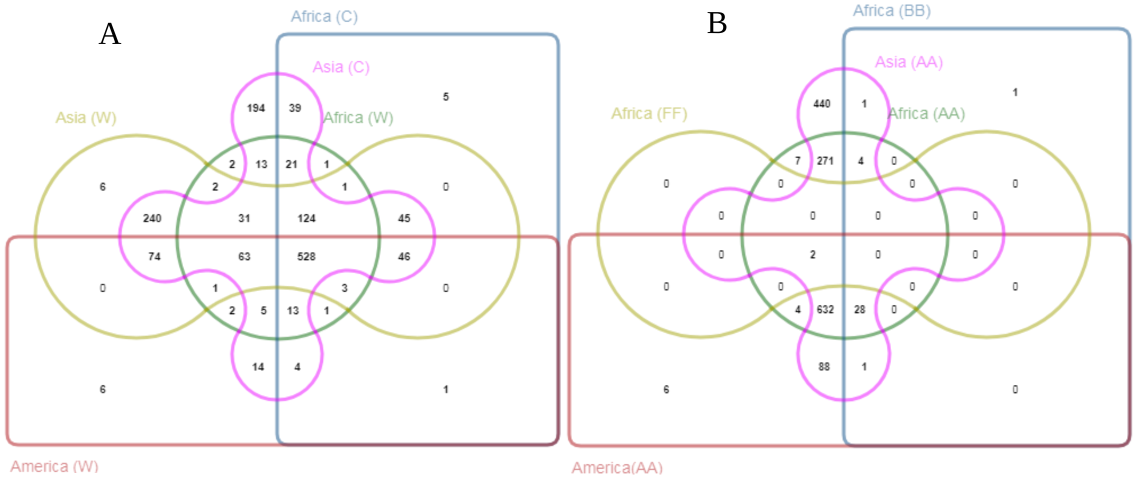
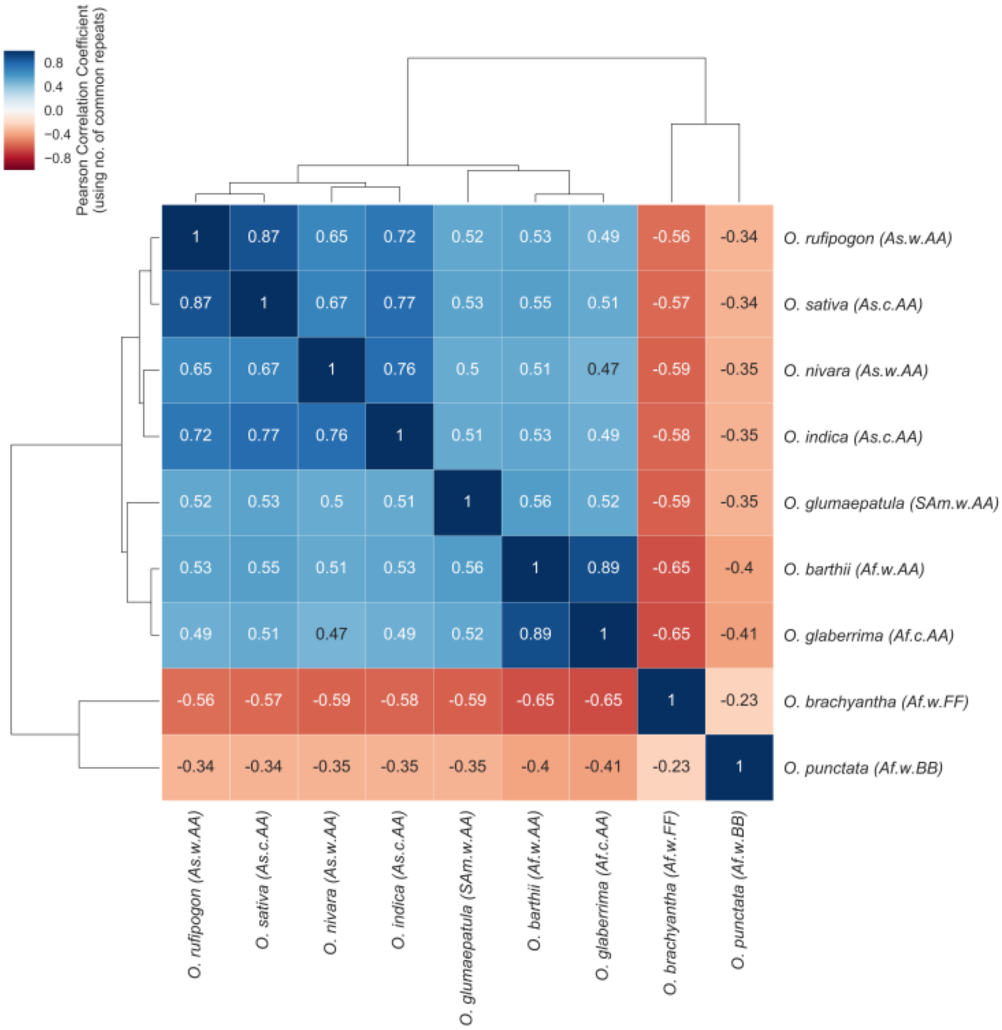


Figure 24: Venn diagram showing common interspersed repeats in A: Wild and cultivated species from the different continents; B: Species from the different continents with different genome types. Pink: Asia (Cultivated), Asia (AA); Blue: Africa (Cultivated), Africa (BB); Green: Africa (Wild), Africa(AA); Red: America (Wild), America (AA); Yellow: Asia (Wild), Africa (FF).

Analysis of common repeats has been performed using several groups. Figure 19A, 21A and 23A has shown clustered the species from Asia and 19B, 21B and 23B for the Africa continent using perfect, imperfect and interspersed repeats respectively. America has only one species hence continent wise grouping cannot be performed. Analysis shows presence of many species specific repeats across both Asian and African varieties. When comparisons between wild and cultivated species from three different continents (Asia, Africa and America has been done using perfect (Figure 20A), imperfect (Figure 22A), and interspersed repeats (Figure 24A), diversification of rice species has been observed. Similar observations have been noticed while comparing different genome types from aforementioned continents using perfect (Figure 5.20B), imperfect (Figure 22B), and interspersed repeats (Figure 22B). All of these results supports the rapid diversification of rice with time [36, 69]. To get more transparency on the results, pair-wise comparisons have been done using perfect (Figure 25) and imperfect repeats (supplementary Figure F2). Result has explained the origins of the cultivated species from their wild ancestors and shows similarity among species with same genome type. Anothet important observation is the grouping of species those belong to same continent. All these results indicates role of the repeats in evolution and their usability in ancestry prediction.



	O. rufipogon (As.w.AA)	O. sativa (As.c.AA)	O. nivara (As.w.AA)	O. indica (As.c.AA)	O. glumaepatula (SAm.w.AA)	O. barthii (Af.w.AA)	O. glaberrima (Af.c.AA)	O. brachyantha (Af.w.FF)	O. punctata (Af.w.BB)
O. rufipogon (As.w.AA)	1	0.87	0.65	0.72	0.52	0.53	0.49	-0.56	-0.34
O. sativa (As.c.AA)	0.87	1	0.67	0.77	0.53	0.55	0.51	-0.57	-0.34
O. nivara (As.w.AA)	0.65	0.67	1	0.76	0.5	0.51	0.47	-0.59	-0.35
O. indica (As.c.AA)	0.72	0.77	0.76	1	0.51	0.53	0.49	-0.58	-0.35
O. glumaepatula (SAm.w.AA)	0.52	0.53	0.5	0.51	1	0.56	0.52	-0.59	-0.35
O. barthii (Af.w.AA)	0.53	0.55	0.51	0.53	0.56	1	0.89	-0.65	-0.4
O. glaberrima (Af.c.AA)	0.49	0.51	0.47	0.49	0.52	0.89	1	-0.65	-0.41
O. brachyantha (Af.w.FF)	-0.56	-0.57	-0.59	-0.58	-0.59	-0.65	-0.65	1	-0.23
O. punctata (Af.w.BB)	-0.34	-0.34	-0.35	-0.35	-0.35	-0.4	-0.41	-0.23	1

Figure 25: Pair-wise comparisons clustering of species using perfect common repeats. (Imperfect repeat: supplementary Figure F2). Red to Blue: Low to High Pearson correlation value; As: Asia, Af: Africa, Sam: South America; w: Wild, c: Cultivated.

Correlation among genomic parameters and repeat densities

As shown in Table 4, repeat densities have no significant correlations with genomic parameters like Size (in Mbp), %GC content, %Genic regions and gene densities per kbp. Similar observations have been found in published literature also [70]. Only significant and strong positive correlation has been found between two repeat density parameters i.e. repeat density in length (D_L) and repeat density in number (D_N) (bolded in Table 4).

Table 4: Correlation Coefficients among Repeat Densities and genomic parameters

Repeat Density	Genomic Parameter	Pearson	P-value*	Spearman	P-value*	Kendall	P-value*
$\boldsymbol{D_L}$	Size(Mbp)	0.0540	1.000	0.0333	1.000	0.0555	1.000
	%GC Content	0.1981	1.000	0.2333	1.000	0.2777	1.000
	Gene Density per kbp	0.8231	0.172	0.7999	0.260	0.6666	0.332
	Genic region (in %)	-0.2917	1.000	-0.1166	1.000	0.0555	1.000
$\boldsymbol{D_N}$	$\boldsymbol{D_L}$	**0.9528**	**0.00193**	**0.8833**	**0.043**	**0.7777**	0.094
	Size(Mbp)	-0.0331	1.000	-0.1499	1.000	-0.0555	1.000
	%GC Content	0.1737	1.000	0.0166	1.000	0.0555	1.000
	Gene Density per kbp	0.7940	0.286	0.7333	0.662	0.5555	1.000
	Genic region (in %)	-0.1441	1.000	0.0333	1.000	0.0555	1.000

*All p-values are adjusted using Bonferroni correction [71]

Comparison of rice with other species:

Arabidopsis thaliana and *Brachypodium distachyon* are the two species that have been chosen here for inter-species comparison. *A. thaliana* is the first model plant species which was sequenced in the year of 2000 due to its small genome size and short life cycle of about 6-8 weeks from its germination to mature seeds [72]. It is well studied and widely used in comparative genomics. On the other hand, *B. distachyon*is the wild monocot grass species that belongs to the same family as rice i.e. Poaceae. Like *Arabidopsis*, it has also small genome, short life cycle and can grow easily. These attributes have made it a suitable new model species for grass family [73]

.

Comparing genomic parameters

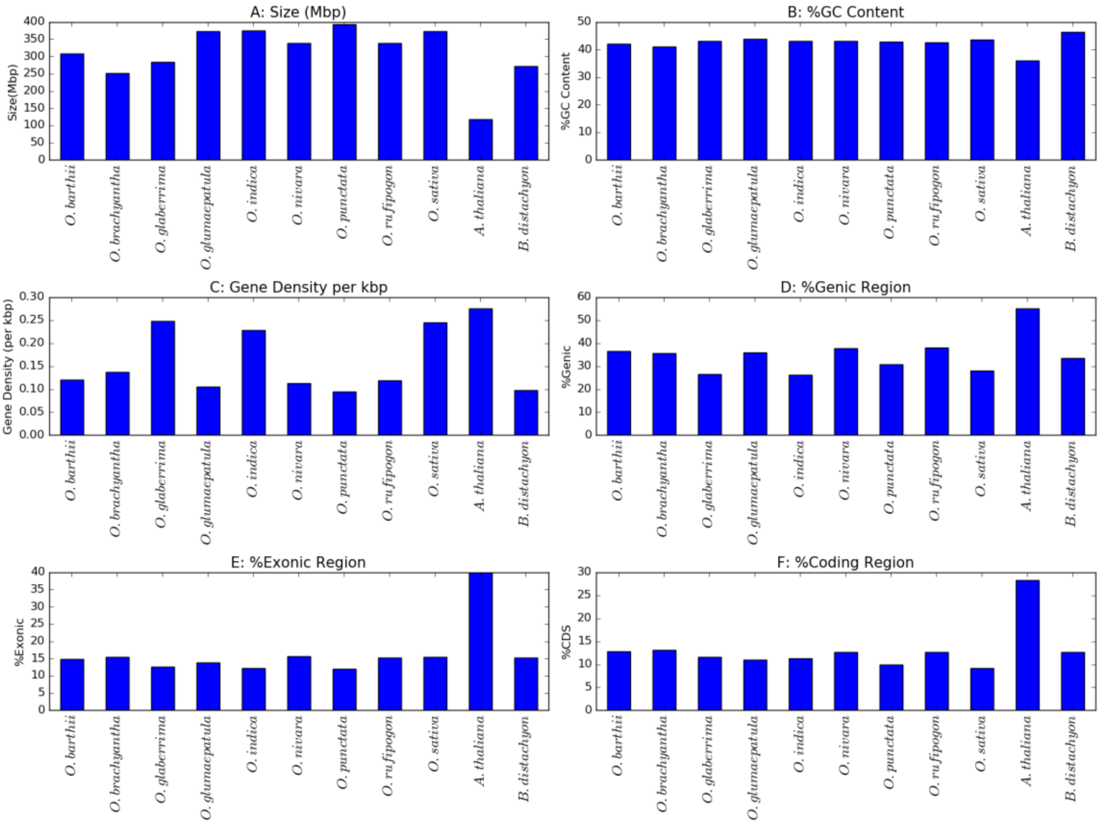


Figure 26: Comparison of genomic parameters across rice and other closely related species. A: Size in Mbp; B: %GC content; C: Gene density per kbp; D: % Genic region; E: % Exonic region; F: %Coding region

As shown in the Figure 26, *Arabidopsis* has the lowest genome size of 119 Mbp followed by wild African rice species *O. brachyantha* and *Brachypodium*. %GC content is almost comparable in all of the species except *Arabidopsis* which has little less value of 36 Mbp (Figure 26A-26B). Though *Arabidopsis* has the smallest genome size among these species, but it has the maximum number of genes per kbp, highest amounts of %genic, %exonic and %coding region (Figure 26C-26F). Three cultivated rice species namely *O. glaberrima*, *O. indica*,and *O. sativa* have comparable gene density which is higher than the other rice species and *Brachypodium* but lower than *Arabidopsis.* Comparisons of the %genic, %exonic and %coding region have revealed that these values are comparable across all of the species except *Arabidopsis*. The observations regarding differences in genomic parameters in rice, *Arabidopsis*, and *Brachypodium* also justified because rice and *Brachypodium* belong to same monocot family whereas *Arabidopsis* which is a dicot belongs to different Brassicaceae (mustard) family.

Comparing repeat distributions

While comparing tandem and interspersed repeats (Figure 27) among the aforementioned plant species, the very first observation is that tandem repeats have been highly varied across all the species. In terms of repeat density in number parameter (D_N), *O. sativa* has the highest value of 4182 repeats/Mbp whereas *B. distachyon* possesses the lowest value of 2770 repeats/Mbp. The median value is 3837 repeats/Mbp. In case of repeat density in length parameter (D_L) also, *B. distachyon* contains the lowest amount of tandem repeats covering 3.7% of the genome in contrast to *O. sativa* genome which has approximately 6% of the genome covered by tandem repeats. The mean value for D_L parameter is 5.07% with variance 0.4 which indicate that distribution of tandem repeats are highly varying across these plants (Figure 27A-27B).

Comparison of interspersed repeats has been performed too (Figure 27C-27D). In the Repbase v20.05 database, no entries have been found for *B. distachyon* species hence not included in the analysis. Comparing repeat density in number parameter has revealed that *O. sativa*has the highest number of interspersed repeats per Mbp (D_N ~75 repeats/Mbp) whereas *O. brachyantha* has the lowest number (< 1 repeat/Mbp). *Arabidopsis* has 29 repeats/Mbp in the genome, which is the median of D_N values also. A similar observation has been found in case of repeat density in length parameter (D_L). *O. sativa* has the highest coverage of 7.75%

followed by *Arabidopsis* (D_L ~ 3.4%). *O. brachyantha* has the lowest value of 0.00072%. The median value for D_L is 1.72%.

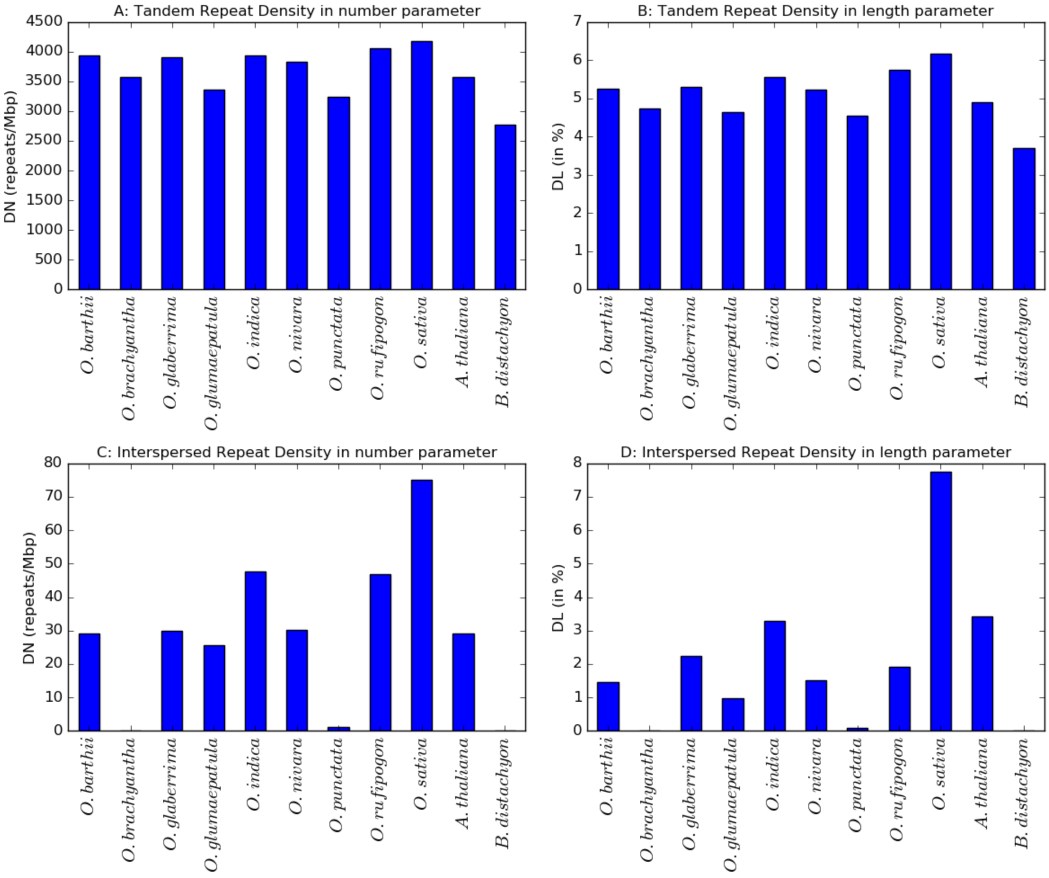


Figure 27: Comparing repeat distribution in *Oryza*, *Brachypodium*, and *Arabidopsis*; A: Tandem repeat density in number (D_N); B: Tandem repeat density in length (D_L); C: Interspersed repeat density in number (D_N); D: Interspersed repeat density in length (D_L);

Occurrences of repeats in stress-related genes

As mentioned earlier, sequencing most of the rice species have not been completed yet and annotations are in progress. Only for model species, *O. sativa* complete annotations have been provided in the RAP database [74]. Consequently, identification of stress-related genes has been performed mainly in the *O. sativa* species. Though, it is possible to detect stress-related genes in the other rice species using homology-based method utilizing *O. sativa* stress-related genes. But this may incorporate unreliable prediction in the analysis. Hence, in the present study of distinguishing stress related repeats, only *O. sativa* genome is considered for further analysis.

Table 5: Repeat Occurrences in stress and housekeeping genomic loci

Repeat Class	Dataset	Genomic loci available	Stress/Negative related loci with repeats	Repeats	Exclusive Repeats
Perfect	A	2692	2688 (99.85%)	11397	7052
	B	952	952 (100%)	4789	2635
	HK	3644	3582 (98.30%)	14277	9280
Imperfect	A	2692	2679 (99.52%)	11260	7484
	B	952	949 (99.68%)	4558	2600
	HK	3644	3571 (98%)	12906	8571
Interspersed	A	2692	794 (29.49%)	1234	1175
	B	952	871 (91.49%)	355	327
	HK*	3644	235 (6.45%)	1065	1015

*HK stands for housekeeping

Tandem repeats have ubiquitously occurred in the rice stress-related genes. Set ‘A’ dataset of experimentally validated stress associated genes comprises of 2692 genomic loci out of which more than 99% of the loci contain perfect repeats (number of loci 2688) (Table 5). Set ‘B’ dataset of predicted stress-related genes, perfect repeats have occurred in all of the genomic loci. Approximately, 98% of the genomic loci that are related to housekeeping genes (the negative set of stress in the present study) contain perfect repeats. Similar observations have been found in case of imperfect repeats also. Imperfect tandem repeats have occurred in more than 99% of the genomic loci in both dataset ‘A’ and ‘B’ whereas 98% of the loci in the negative set have imperfect repeats. Distribution of interspersed repeats in the stress-related genes is quite less than tandem repeats. 29% of Set ‘A’ loci and 91% of Set ‘B’ contains interspersed repeats whereas only 6% of the housekeeping genes contain interspersed repeats.

As shown in Figure 28, short tandem repeats are omnipresent in both the positive sets of stress-related genes but long repeats (motif length > 10 bp) are very rare. Trimeric repeating motifs are more frequent than other motifs followed by mononucleotide repeats and dinucleotide repeats. Imperfect hexanucleotide repeats are also marked their occurrences in the stress-related genes. Minisatellites of motif length 7 bp are found to occur in excess than the other minisatellites. Minisatellites are found to be more polymorphic than microsatellites in the stress-related genes.

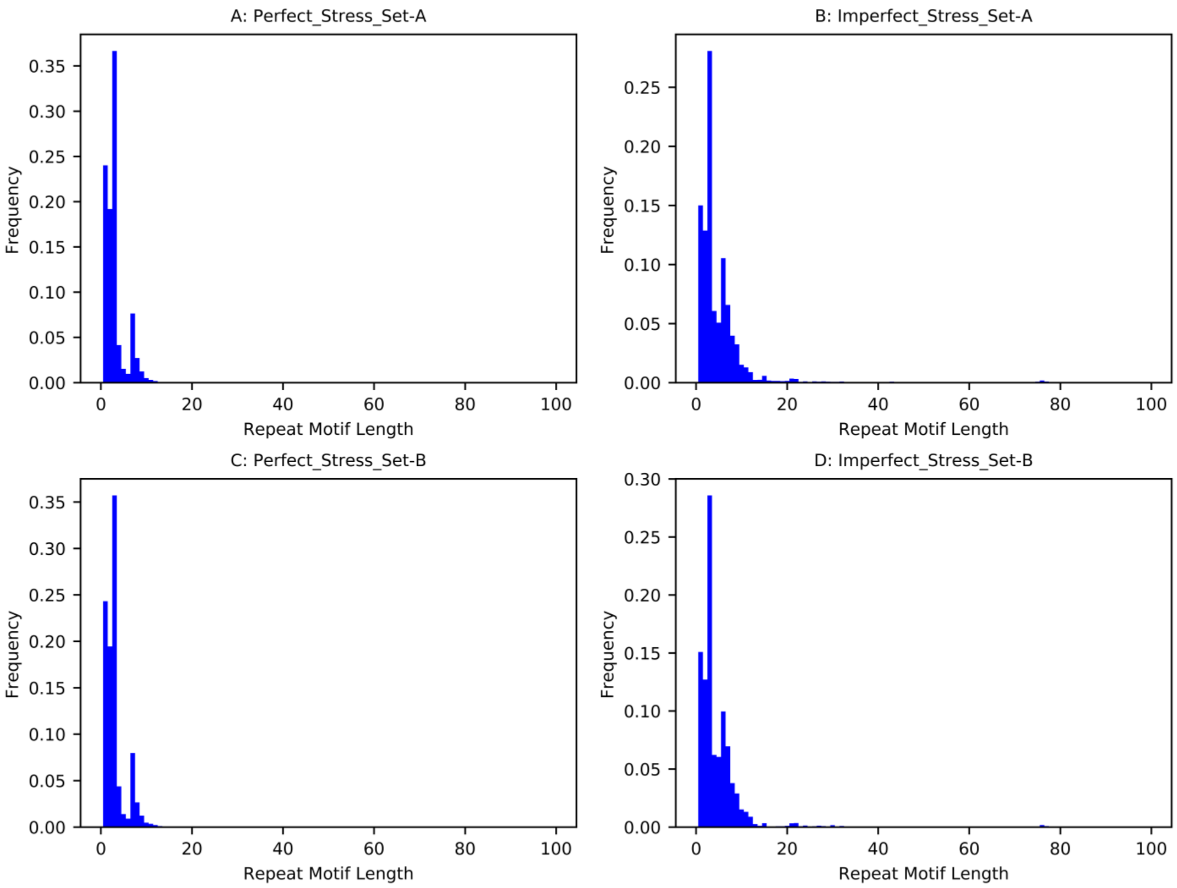


Figure 28: Distribution of repeats of different motif sizes; A: Perfect repeats in Set A gene set; B: Imperfect repeats in Set A gene set; C: Perfect repeats in Set B gene set; D: Imperfect repeats in Set B gene set;

While checking the number of loci that contain any repeat or not, it is found that repeats are not dispersed in the multiple genomic loci rather they are preferred to occur in particular locus (Figure 29). Similar observations have been found in case of both positive sets (A & B) and also valid for both perfect and imperfect repeats. Very few repeats have been found to occur in multiple genomic loci in rice as observed from Figure 29. This is the major difference that has been found between rice and *Salmonella.* In *Salmonella* repeats have occurred both in the single locus and in multiple loci. Present observation suggests that repeats in the rice genomes are highly precise and have some specific functional implications.

Distributions of repeats in stress-related and NSS genes (housekeeping genes) have been compared and whether there is an association between stress and repeat, have been tested

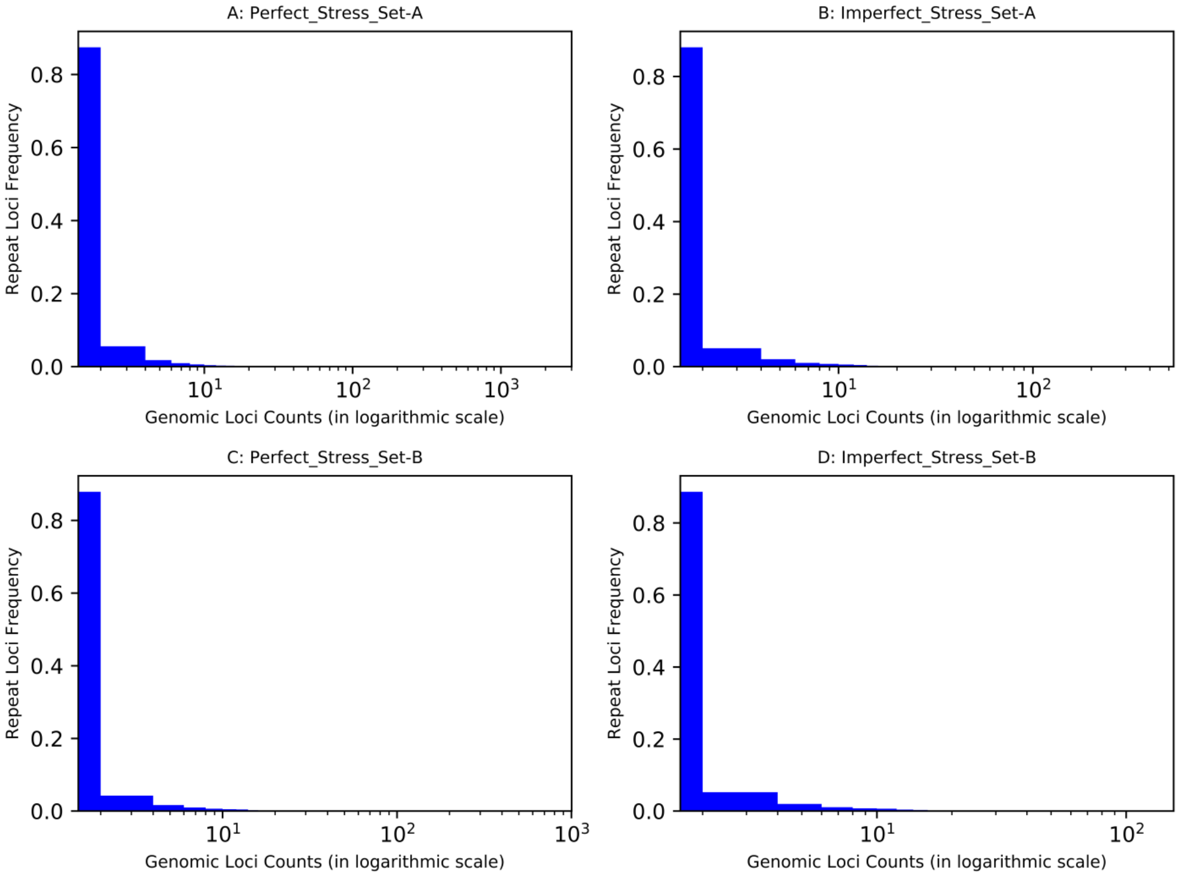


Figure 29: Distribution of repeats with varying genomic loci counts; A: Perfect repeats in stress genes (set A) B: Imperfect repeats in stress genes (set A); C: Perfect repeats in stress genes (set B); D: Imperfect repeats in stress genes (set B).

using null hypothesis stated in Table 6. The comparison suggests that significant positive association has been observed between stress and repeats. Several tests including chi-square test for independence and Fisher exact test have been conducted to compare the occurrences of the repeats in the different genomic locations. Results have shown that repeat's occurring in different stress-related genomic locations are significantly associated (p-value < 0.05) with stress-related genes. Similar observations have been found for all the repeat classes. To reduce Type I error in multiple comparisons of the genome, Bonferroni correction has been applied to adjust the p-values. Repeat classes, genomic locations, test statistics and corresponding p-values have been listed in Table 6.

Table 6: Comparison of repeats in stress and NSS genes

Null Hypothesis: There is no association between repetitive elements and stress related loci						
Repeat Class	Location	Fisher's Exact test		Chi-square Test		Null Hypothesis Significance level 0.05
		Test Statistic	p-value*	Test Statistic	p-value*	
Perfect Microsatellites	Inside gene	15.74	3.30e-13	51.38	1.55e-11	Rejected
	Upstream	15.73	3.32e-13	51.34	1.47e-11	Rejected
	Downstream	15.75	3.32e-13	51.44	1.48e-11	Rejected
	All locations	15.75	3.32e-13	51.44	4.52e-10	Rejected
Perfect Minisatellites	Inside gene	13.99	5.78e-12	44.73	1.20e-11	Rejected
	Upstream	15.9	3.30e-13	51.84	1.07e-11	Rejected
	Downstream	15.97	3.70e-13	52.06	1.50e-11	Rejected
	All locations	15.75	1.68e-13	51.4	2.52e-11	Rejected
Imperfect Tandem	Inside gene	15.49	8.62e-14	50.39	1.41e-11	Rejected
	Upstream	15.79	3.32e-13	51.52	5.22e-12	Rejected
	Downstream	16.33	4.70e-19	53.48	1.49e-11	Rejected
	All locations	15.75	3.94e-13	51.42	2.04e-17	Rejected
Interspersed	Inside gene	27.1	3.34e-15	78.02	2.80e-11	Rejected
	Upstream	16.17	3.90e-15	50.19	2.08e-13	Rejected
	Downstream	19.32	1.72e-13	59.82	2.88e-13	Rejected
	All locations	18.31	3.30e-13	59.18	1.45e-11	Rejected
All Repeats	Inside gene	15.76	1.73e-13	51.47	1.49e-11	Rejected
	Upstream	15.75	3.32e-13	51.42	1.45e-11	Rejected
	Downstream	15.76	3.34e-15	51.47	1.48e-11	Rejected
	All locations	15.75	3.90e-15	51.44	2.88e-13	Rejected

*All p-values are adjusted using Bonferroni correction [75]

Mining associations between stress and repeats

As shown in Table 5, total 16186 perfect and 15818 imperfect tandem repeats (cumulative count of Set A and Set B) have been identified in the stress-related datasets (both Set 'A' and Set 'B'). Total 1589 interspersed repeats have been also detected in the stress-related genes. Extracting exclusive repeats and removing redundancies result 9086 (~ 56%) perfect and

9537 (~ 60%) imperfect tandem repetitive loci. On the other hand, 1502 (~ 94%) exclusive interspersed repeats have been discovered. 493 and 420 exclusive perfect and imperfect tandem repeats have been found to be common between Set ‘A’ and Set ‘B’ stress gene datasets. While finding common exclusive interspersed repeats, only 73 repetitive loci have been distinct. Total 19701 unique exclusive repetitive loci (combining perfect, imperfect and interspersed and removing redundancies) has been identified, out of which more than 99% (~ 19631) have shown positive association (nPMI> 0) with stress (Figure 30). Lists of top 10 potential stress associated perfect, imperfect and interspersed repeats (sorted by nPMI score) have been presented in Table 7, Table 8 and Table 9 respectively and discussed below along with its product information. These repeats can be used as the markers for identifying novel stress-related genes in *Oryza* species.

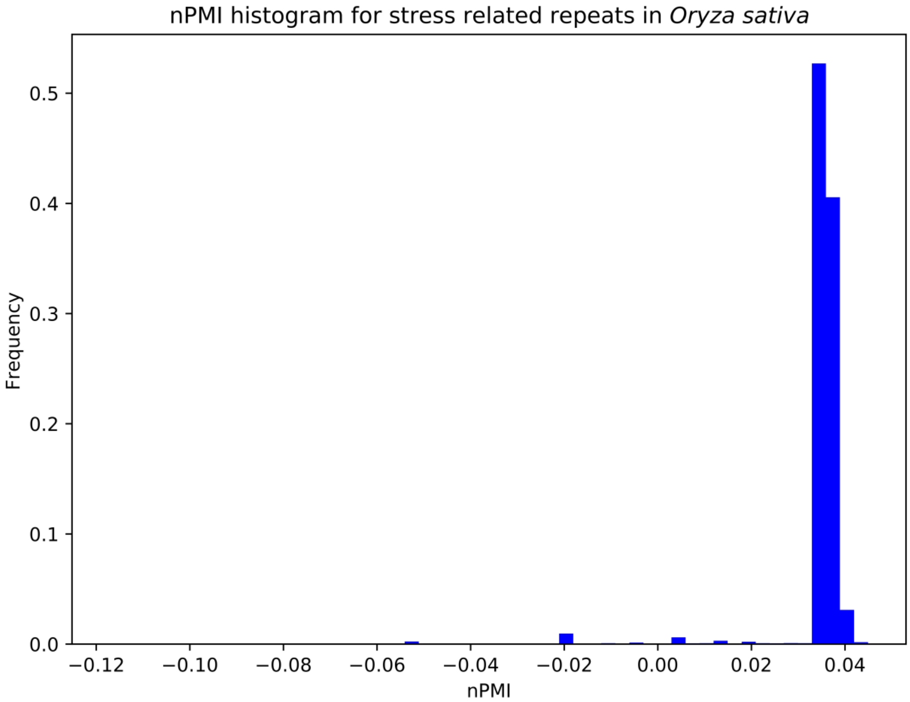


Figure 30: Distribution of nPMI values of exclusive stress related repeats in *O. sativa* species.

Repeats occurring exclusively in stress-related genes of *Oryza sativa*

According to the present analysis of stress associated perfect repeats, approximately 89% of the repeats are found to be minisatellites those are exclusively present in the Set A (gold standard) stress-related genes. A similar result (~90%) has been found for stress-related genes in the Set B also. Majority of these minisatellites have occurred in the upstream or downstream regions of the genome. Very few of these repeats have been found inside the gene.

As mentioned before, Table 7 shows top 10 (sorted by nPMI values) perfect minisatellites of length greater than 10 nucleotides those are exclusively associated with stress. 14-mer repeating motif 'GCCATTGTCCTGCT' with length 30 solely found in the *O. sativa* species in the locus that encodes for heat shock protein (HSP90). Heat shock proteins are a class of molecular chaperones those are crucial for plant growths and related to several important functions like cell signaling, cell immunity, integrity maintenance and many more [76].

Table 7: Exclusive Perfect Minisatellites (Motif length > 10) sorted by nPMI

No.	Repeating Motif	Repeat loci length	Chr. No.	Repeat Count at Location			nPMI
				G	U	D	
PR1	GCCATTGTCCTGCT	30	4	0	0	1	0.043
PR2	CTGCAGTACTACTCACTGACATGTGAGCCCCACAGCCCCTCTGAATGACATCG	106	9	0	1	2	0.042
PR3	ACAAAAGGCAAGGATTATAATGGTTCAGACCCC	66	8	0	1	1	0.042
PR4	TTTTTGGACGGA	24	8	1	1	1	0.042
PR5	AGGTCAAGATTTT	26	5	0	0	1	0.042
PR6	GCGATAGATCACCAAAACGACAGATTT	54	2	0	0	1	0.042
PR7	TTAAATTTTATGAATTTTTTATATAATTGT	67	1	0	0	2	0.042
PR8	TGGAGACAGCCTTTCTTGGTTTTGTGAAATACTTGC	82	12	0	0	1	0.042
PR9	ACATCTCTACAGTAATATGAAAT	47	12	0	0	1	0.042
PR10	TAACATTTTTTT	24	12	0	0	1	0.042

G: Genic; U: Upstream to the gene; D: Downstream to the gene

Another repeat of motif length 53 bp is related to the DNA binding protein coding gene possibly a transcription factor (TF) (locus found in Grass TFDB [77]) of basic helix-loop-helix (bHLH) family. This family of TFs has role in phytochrome reaction under photoreceptor signaling pathway. Other occurrences of this motif are in the downstream region of carbonate dehydratase encoding gene and upstream of an unknown/hypothetical protein-coding gene. This particular repeat is also present in the *O. rufipogon* species which is thought to be the ancestor of *O. sativa* [36]. A 33-mer repeating motif "ACAAAAGGCAAGGATTATAATGGTTCAGACCCC" of length 66 bp is present in the upstream region of the gene having an annotation of CCT domain containing protein which controls the flowering time also has been reported to be involved in abiotic stress response [78].

Table 8: Exclusive Imperfect minisatellites (Motif length > 10) sorted by nPMI

ID	Repeating Motif	Repeat loci length	Chr. No.	Repeat Count at Location			nPMI
				G	U	D	
IR1	CCGCCGCCGCCT	44	8	0	0	1	0.043
IR2	TGTAACTTGCA	59	3	0	0	2	0.043
IR3	TCTGAATGACATCGTCTGCAGTACTACTCACTGACATGTGAGCCCCACAGCCCC	162	9	0	1	2	0.042
IR4	GCGGCCGGGCGGGCGAGGGGCCGGCG	132	2	0	1	1	0.042
IR5	CCCGATACGTAT	25	10	0	1	1	0.042
IR6	TTATCTCTAGGATATATC	54	9	0	0	1	0.042
IR7	GCGGCGCGACG	41	8	0	0	1	0.042
IR8	TAATTGAATCTCACTA	39	5	0	0	1	0.042
IR9	TGTAATTACAGTGTAACTTGTA	70	3	0	0	2	0.042
IR10	AGAGTCCATATAGAAATACAATTTAGAAATAACTGAAATTCGGAATTAAAAATAAGGAATATTAGAAGTAGAGTAT	650	3	0	0	1	0.042

IR stands for imperfect repeat; G: Genic; U: Upstream to the gene; D: Downstream to the gene

The same repeat has been also located in the downstream region of the gene encoding Succinate dehydrogenase iron protein beta subunit (SDBH) which serves as the direct source of reactive oxygen species (ROS) production in rice are also accompanied with up-regulation of many stress-related genes [79]. A complete list of all the exclusive stress associated perfect repeats (both micro- and minisatellites) has been listed in the Supplementary Table file.

Polymorphic stress related minisatellites which are only present in stress genes (Set A and Set B) have been detected too. Top 10 have been shown in Table 8. Like perfect repeats, major proportions (~ 65%) of the imperfect repeats are in the upstream and downstream regions of the stress-related genes. A 12-mer polymorphic repeat of length 44 bp has occurred in the downstream region of the ROS stress controlling gene SDB [79]. Another repeating motif "TGTAACTTGCA" with repeat length 59 has occurred in the downstream regions of two important proteins. One is the IQ calmodulin-binding region domain containing protein which is very much essential in elevation of controlling Ca^{2+} concentration during the cross-talk of various stressors like cold, heat, drought, salt etc. [80]. The other protein is LSTK-1 kinase is responsive to drought and osmosis stress [81].

Table 9: Interspersed repeats exclusively present inside or in vicinity of stress genes

No.	Transposon/TE	Repeat loci length	Chr. No.	Repeat Count at Location			nPMI
				G	U	D	
INR1	RIRE5-I_OS	4471	1,5,7	2	0	2	0.041
INR2	SPMLIKE	4006	1,2	1	2	0	0.040
INR3	Harbinger-N87_OS	957	11	1	1	2	0.040
INR4	Gypsy-16_OS-LTR	2240	10	1	2	1	0.040
INR5	Harbinger-N32_OS	289	9	0	0	1	0.039
INR6	ENSPM4_OS	100	8	0	0	1	0.039
INR7	Gypsy-8_OS-LTR	718	7	1	0	1	0.039
INR8	Helitron-N150_OS	3893	7	1	1	1	0.039
INR9	RIREXC_I	35	7	1	0	0	0.039
INR10	EnSpm-N24_OS	3365	7	1	1	1	0.039

INR: Interspersed Repeats; G: Genic; U: Upstream to the gene; D: Downstream to the gene

“TCTGAATGACATCGTCTGCAGTACTACTCACTGACATGTGAGCCCCACAGCCCC” is another repeating motif of length 54 bp have been exclusively associated with bHLH family of TFs and carbonic dehydratase like proteins whose roles in stress response have been discussed earlier. A complete list of stress associated imperfect repeats has been presented in the Supplementary Table file.

Top 10 transposons like interspersed repeats predicted to be associated with stress response have been listed in Table 9. LTR retrotransposon RIRE5-I_OS have been found inside and downstream region of the two proteins namely non-cyanogenic beta-glucosidase and serine carboxypeptidase2 which are involved in plant defense against biotic factors and oxidative stress [82-83]. A DNA transposon SPMLIKE has been found in the upstream regions of glycoside hydrolase and nitric oxide synthase. The glycoside hydrolase family proteins has a functional role in forming cell wall architecture and often shows their responses under stressful conditions [84]. Nitric oxide synthase is responsible for the production of nitric oxide and reported to be associated with salt stress response in *Arabidopsis* [85]. List of all exclusively stress related interspersed repeats are given in the Supplementary Table file.

Analysis of false positives

Many repetitive sequences are found to occur in both stress and NSS or negative sets of genes (housekeeping genes) and marked as the false positives. Total 4345 perfect (9% of the total repeats associated with stress and negative gene sets), 3776 imperfect (8%) and 689 interspersed (1%) repeats have been detected as false positives in gold standard gene set (Set A). Similarly, 2154 perfect (~5%), 1958 imperfect (~ 3.5%) and 175 interspersed (~ 0.3%) repeats have been identified as false positives in Set B (silver standard) stress-related genes. Out of 4345 perfect false positives in Set A, 61% are found to be minisatellites and 24% of the 3776 imperfect minisatellites have been marked as false positives. Similarly, in the Set B gene list, 49% of 2154 perfect, 14% of 1958 imperfect minisatellites have been identified as false positives. As mentioned earlier, housekeeping genes have been used as the negative set of for comparing repeat distributions but authentic housekeeping genes are few in number. Here, the list of housekeeping genes used has been predicted from expression data sets. Hence, there exists a high possibility that the repeats marked as false positives might be related to stress response but due to data insufficiency, they are predicted as false positives (Supplementary Table file).

Prediction of novel candidate stress associated loci in *Oryza* genomes

Those repeats which are exclusively occurred in stress-related genes of *O. sativa* genome can be used as a marker to identify new candidate stress associated loci from other rice species. Some of the repeating motifs those are extracted earlier (Table 7) have been used to mark stress related loci in the genomes of rice species other than *O. sativa* as shown in Table 10. A 53-mer perfect minisatellite is present in the vicinity of the gene that encodes for carbonate dehydratase-like protein and located on chromosome 9 in *O. sativa*. The same repeat has been also found in the vicinity of "ORUFI09G14140" locus in *O. rufipogon* located in the same chromosome no. 9. The length of the repetitive sequence is 106 bp which is very much unlikely to be a random event.

Table 10: Prediction of novel candidate stress associated genes using repetitive markers

No.	Repeating Motif	Repeat loci length	Chr. No.	Genomic Loci	Product
1	CTGCAGTACTACTCACTGACATGTGAGCCCCACAGCCCCTCTGAATGACATCG	106	9	*O. sativa:* Os09g0464000	Carbonate dehydratase-like protein
				O. rufipogon: ORUFI09G14140	Not Available
2	ACAAAAGGCAAGGATTATAATGGTTCAGACCCC	66	8	*O. sativa:* Os08g0120000	Succinate dehydrogenase iron-protein subunit (SDHB)
				O. barthii: OBART08G10210	Not Available
				O. glaberrima ORGLA08G0007500	Uncharacterized protein
				O. indica: BGIOSGA027744	Uncharacterized protein
				O. rufipogon: ORUFI08G10490	Not Available
3	TTTTTGGACGGA	24	8	*O. sativa*: Os08g0498400	Caffeoyl-CoA O-methyltransferase
				O. indica: BGIOSGA026737	Putative Uncharacterized protein

Conclusions

Global temperature has been rising day by day due to the increment of the pollution level and emission of the greenhouse gases. As a result of this global warming accompanied with other stress factors, the severe adverse effect has been noticed on the rice yield in past few years. The exponential increment of the world's population has magnified this negative impact up to several folds. To understand the stress response mechanisms for adaptation under unfavorable conditions and to increase the annual yield of rice to secure the upcoming food crisis, extensive studies are required. In the present study, a comparative genomics-based approach has been developed to detect and characterize repetitive sequences that might play an important role in adaptation under stressful conditions. Significant non-random distributions of repeats have been observed in rice. In case of tandem repeats, trimeric are found to be present in excess followed by monomeric and dimeric repeats. In comparison to short tandem repeats or microsatellites, long repeats (motif size > 10 bp) are less frequent but have been marked their occurrences in the genomes. Repeats are ubiquitously found in all genomic regions which include genic, intergenic, exons, introns and UTRs. No significant correlation has been observed among repeat densities and genomic parameters. Analysis of common repeats across 9 rice species has shown that each of species has some species-specific repeats. Most importantly, pairwise comparisons have clearly distinguished clusters of rice species from same geographical locations and of a particular genome type (Figure 23). Comparisons with other species like *Arabidopsis* and *Brachypodium* have strengthened this as similar results found from earlier published works.

Presence of repeats in known stress-related genes has been observed in the analysis. Repeat's significant association with stress clearly indicates its functional role in adaptation under stressful environments. Though, experimental data insufficiency is a major obstacle to using it as a biomarker and to identify candidate stress related loci which are still un-annotated or hypothesized in rice species other than *O. sativa*. But, it can be a starting point to use repeats for novel gene identification of functional importance as done in the present work. However, it may not be used with certainty as a generic predictive method as repeats are occurring in almost every important known locations of the stress associated genes and genomes, so experimental validation is required along with the understanding its causal relation with stress response and adaptation.

References

1. Yuan, Shen, Bruce A. Linquist, Lloyd T. Wilson, Kenneth G. Cassman, Alexander M. Stuart, Valerien Pede, Berta Miro et al. "Sustainable intensification for a larger global rice bowl." *Nature Communications* 12, no. 1 (2021): 1-11.
2. Peng, Shaobing, Jianliang Huang, John E. Sheehy, Rebecca C. Laza, Romeo M. Visperas, Xuhua Zhong, Grace S. Centeno, Gurdev S. Khush, and Kenneth G. Cassman. "Rice yields decline with higher night temperature from global warming." *Proceedings of the National Academy of Sciences* 101, no. 27 (2004): 9971-9975.
3. Panda, Debabrata, Swati Sakambari Mishra, and Prafulla Kumar Behera. "Drought Tolerance in Rice: Focus on Recent Mechanisms and Approaches." *Rice Science* 28, no. 2 (2021): 119-132.
4. Mickelbart, Michael V., Paul M. Hasegawa, and Julia Bailey-Serres. "Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability." *Nature Reviews Genetics* 16, no. 4 (2015): 237-251.
5. Ray, Deepak K., James S. Gerber, Graham K. MacDonald, and Paul C. West. "Climate variation explains a third of global crop yield variability." *Nature communications* 6, no. 1 (2015): 1-9.
6. Vo, Kieu Thi Xuan, Md Mizanor Rahman, Md Mustafizur Rahman, Kieu Thi Thuy Trinh, Sun Tae Kim, and Jong-Seong Jeon. "Proteomics and Metabolomics Studies on the Biotic Stress Responses of Rice: an Update." *Rice* 14, no. 1 (2021): 1-16.
7. Temnykh, Svetlana, Genevieve DeClerck, Angelika Lukashova, Leonard Lipovich, Samuel Cartinhour, and Susan McCouch. "Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential." *Genome research* 11, no. 8 (2001): 1441-1452.
8. McCouch, Susan R., Leonid Teytelman, Yunbi Xu, Katarzyna B. Lobos, Karen Clare, Mark Walton, Binying Fu et al. "Development and mapping of 2240 new SSR markers for rice (Oryza sativa L.)." *DNA research* 9, no. 6 (2002): 199-207.
9. Legendre, Matthieu, Nathalie Pochet, Theodore Pak, and Kevin J. Verstrepen. "Sequence-based estimation of minisatellite and microsatellite repeat variability." *Genome research* 17, no. 12 (2007): 1787-1796.

10. Winberg, B. C., Z. Zhou, J. F. Dallas, C. L. McIntyre, and J. P. Gustafson. "Characterization of minisatellite sequences from Oryza sativa." *Genome* 36, no. 5 (1993): 978-983.
11. Mao, Long, Todd C. Wood, Yeisoo Yu, Muhammad A. Budiman, Jeff Tomkins, Sung-sick Woo, Maciek Sasinowski et al. "Rice transposable elements: a survey of 73,000 sequence-tagged-connectors." *Genome Research* 10, no. 7 (2000): 982-990.
12. Juretic, Nikoleta, Thomas E. Bureau, and Richard M. Bruskiewich. "Transposable element annotation of the rice genome." *Bioinformatics* 20, no. 2 (2004): 155-160.
13. Turcotte, Kime, Sujatha Srinivasan, and Thomas Bureau. "Survey of transposable elements from rice genomic sequences." *The Plant Journal* 25, no. 2 (2001): 169-179.
14. McKinley, Justin D., Jeffrey T. LaFrance, and Valerien O. Pede. "Climate Change Adaptation Strategies Vary With Climatic Stress: Evidence From Three Regions of Vietnam." *Frontiers in Sustainable Food Systems* 5 (2021): 762650.
15. Pandey, Renu, Krishnapriya Vengavasi, and Malcolm J. Hawkesford. "Plant adaptation to nutrient stress." *Plant Physiology Reports* (2021): 1-4.
16. Singh, Sangeeta, Suresh Chand, N. K. Singh, and Tilak Raj Sharma. "Genome-wide distribution, organisation and functional characterization of disease resistance and defence response genes across rice species." *Plos one* 10, no. 4 (2015): e0125964.
17. Molla, Kutubuddin Ali, Ananda Bhusan Debnath, Showkat Ahmad Ganie, and Tapan Kumar Mondal. "Identification and analysis of novel salt responsive candidate gene based SSRs (cgSSRs) from rice (Oryza sativa L.)." *BMC plant Biology* 15, no. 1 (2015): 1-11.
18. Cooper, Bret, Joseph D. Clarke, Paul Budworth, Joel Kreps, Don Hutchison, Sylvia Park, Sonia Guimil et al. "A network of rice genes associated with stress response and seed development." *Proceedings of the National Academy of Sciences* 100, no. 8 (2003): 4945-4950.
19. Das, Gitishree, and G. J. N. Rao. "Molecular marker assisted gene stacking for biotic and abiotic stress resistance genes in an elite rice cultivar." *Frontiers in plant science* 6 (2015): 698.
20. Ali, Jauhar, Jian-Long Xu, Yong-Ming Gao, Xiu-Fang Ma, Li-Jun Meng, Ying Wang, Yun-Long Pang et al. "Harnessing the hidden genetic diversity for improving multiple abiotic stress tolerance in rice (Oryza sativa L.)." *PLoS One* 12, no. 3 (2017): e0172515.

21. Sandhu, Maninder, V. Sureshkumar, Chandra Prakash, Rekha Dixit, Amolkumar U. Solanke, Tilak Raj Sharma, Trilochan Mohapatra, and Amitha Mithra SV. "RiceMetaSys for salt and drought stress responsive genes in rice: a web interface for crop improvement." *BMC bioinformatics* 18, no. 1 (2017): 1-11.
22. Akakpo, Roland, Marie-Christine Carpentier, Yue Ie Hsing, and Olivier Panaud. "The impact of transposable elements on the structure, evolution and function of the rice genome." *New Phytologist* 226, no. 1 (2020): 44-49.
23. Wessler, Susan R. "Plant retrotransposons: turned on by stress." *Current Biology* 6, no. 8 (1996): 959-961.
24. Rodrigues, Jessica A., Ping-Hung Hsieh, Deling Ruan, Toshiro Nishimura, Manoj K. Sharma, Rita Sharma, XinYi Ye et al. "Divergence among rice cultivars reveals roles for transposition and epimutation in ongoing evolution of genomic imprinting." *Proceedings of the National Academy of Sciences* 118, no. 29 (2021).
25. Finatto, Taciane, Antonio Costa de Oliveira, Cristian Chaparro, Luciano C. Da Maia, Daniel R. Farias, Leomar G. Woyann, Claudete C. Mistura et al. "Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice." *Rice* 8, no. 1 (2015): 1-18.
26. Barrera-Figueroa, Blanca E., Lei Gao, Zhigang Wu, Xuefeng Zhou, Jianhua Zhu, Hailing Jin, Renyi Liu, and Jian-Kang Zhu. "High throughput sequencing reveals novel and abiotic stress-regulated microRNAs in the inflorescences of rice." *BMC plant biology* 12, no. 1 (2012): 1-11.
27. Wang, Dong, Zhipeng Qu, Lan Yang, Qingzhu Zhang, Zhi-Hong Liu, Trung Do, David L. Adelson, Zhen-Yu Wang, Iain Searle, and Jian-Kang Zhu. "Transposable elements (TE s) contribute to stress-related long intergenic noncoding RNA s in plants." *The Plant Journal* 90, no. 1 (2017): 133-146.
28. Zou, Xin-Hui, Yu-Su Du, Liang Tang, Xin-Wei Xu, Jeff J. Doyle, Tao Sang, and Song Ge. "Multiple origins of BBCC allopolyploid species in the rice genus (Oryza)." *Scientific reports* 5, no. 1 (2015): 1-10.
29. Yu, Jun, Jun Wang, Wei Lin, Songgang Li, Heng Li, Jun Zhou, Peixiang Ni et al. "The genomes of Oryza sativa: a history of duplications." *PLoS biology* 3, no. 2 (2005): e38.
30. Khush, Gurdev S. "Origin, dispersal, cultivation and variation of rice." *Plant molecular biology* 35, no. 1 (1997): 25-34.

31. Callaway, Ewen. "Domestication: The birth of rice." *Nature* 514, no. 7524 (2014): S58-S59.

32. Yamanaka, Shinsuke, Ikuo Nakamura, Hirokazu Nakai, and Yo-Ichiro Sato. "Dual origin of the cultivated rice based on molecular markers of newly collected annual and perennial strains of wild rice species, Oryza nivara and O. rufipogon." *Genetic Resources and Crop Evolution* 50, no. 5 (2003): 529-538.

33. Linares, Olga F. "African rice (Oryza glaberrima): history and future potential." *Proceedings of the National Academy of Sciences* 99, no. 25 (2002): 16360-16365.

34. Fuchs, Eric J., Allan Meneses Martínez, Amanda Calvo, Melania Muñoz, and Griselda Arrieta-Espinoza. "Genetic diversity in Oryza glumaepatula wild rice populations in Costa Rica and possible gene flow from O. sativa." *PeerJ* 4 (2016): e1875.

35. Ammiraju, Jetty SS, Fei Lu, Abhijit Sanyal, Yeisoo Yu, Xiang Song, Ning Jiang, Ana Clara Pontaroli et al. "Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set." *The Plant Cell* 20, no. 12 (2008): 3191-3209.

36. Zhang, Qun-Jie, Ting Zhu, En-Hua Xia, Chao Shi, Yun-Long Liu, Yun Zhang, Yuan Liu et al. "Rapid diversification of five Oryza AA genomes associated with rice adaptation. " *Proceedings of the National Academy of Sciences* 111, no. 46 (2014): E4954-E4962.

37. Yates, Andrew D., James Allen, Ridwan M. Amode, Andrey G. Azov, Matthieu Barba, Andrés Becerra, Jyothish Bhai et al. "Ensembl Genomes 2022: an expanding genome resource for non-vertebrates." *Nucleic acids research* 50, no. D1 (2022): D996-D1003.

38. Tello-Ruiz, Marcela K., Sushma Naithani, Parul Gupta, Andrew Olson, Sharon Wei, Justin Preece, Yinping Jiao et al. "Gramene 2021: harnessing the power of comparative genomics and pathways for plant research." *Nucleic Acids Research* 49, no. D1 (2021): D1452-D1463.

39. Sasaki, Takuji. "The map-based sequence of the rice genome." *Nature* 436, no. 7052 (2005): 793-800.

40. Du, Huilong, Ying Yu, Yanfei Ma, Qiang Gao, Yinghao Cao, Zhuo Chen, Bin Ma et al. "Sequencing and de novo assembly of a near complete indica rice genome." *Nature communications* 8, no. 1 (2017): 1-12.

41. Kapitonov, Vladimir V., and Jerzy Jurka. "A universal classification of eukaryotic transposable elements implemented in Repbase." *Nature Reviews Genetics* 9, no. 5 (2008): 411-412.
42. Liao, Xingyu, Kang Hu, Adil Salhi, You Zou, Jianxin Wang, and Xin Gao. "msRepDB: a comprehensive repetitive sequence database of over 80 000 species." *Nucleic acids research* 50, no. D1 (2022): D236-D245.
43. Luo, Xizhi, Shiyu Chen, and Yu Zhang. "PlantRep: a database of plant repetitive elements." *Plant Cell Reports* (2022): 1-4.
44. Castelo, Adalberto T., Wellington Martins, and Guang R. Gao. "TROLL—tandem repeat occurrence locator." *Bioinformatics* 18, no. 4 (2002): 634-636.
45. Mayer, Christoph, Florian Leese, and Ralph Tollrian. "Genome-wide analysis of tandem repeats in Daphnia pulex-a comparative approach." *BMC genomics* 11, no. 1 (2010): 1-28.
46. Benson, Gary. "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic acids research* 27, no. 2 (1999): 573-580.
47. Moreno-Hagelsieb, Gabriel, and Kristen Latimer. "Choosing BLAST options for better detection of orthologs as reciprocal best hits." *Bioinformatics* 24, no. 3 (2008): 319-324.
48. Sakai, Hiroaki, Sung Shin Lee, Tsuyoshi Tanaka, Hisataka Numa, Jungsok Kim, Yoshihiro Kawahara, Hironobu Wakimoto et al. "Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics." *Plant and Cell Physiology* 54, no. 2 (2013): e6-e6.
49. Lakra, Nita, Charanpreet Kaur, Khalid Anwar, Sneh Lata Singla-Pareek, and Ashwani Pareek. "Proteomics of contrasting rice genotypes: identification of potential targets for raising crops for saline environment." *Plant, cell & environment* 41, no. 5 (2018): 947-969.
50. Alter, Svenja, Kai C. Bader, Manuel Spannagl, Yu Wang, Eva Bauer, Chris-Carolin Schön, and Klaus FX Mayer. "DroughtDB: an expert-curated compilation of plant drought stress genes and their homologs in nine species." *Database* 2015 (2015).
51. Naika, Mahantesha, Khader Shameer, Oommen K. Mathew, Ramanjini Gowda, and Ramanathan Sowdhamini. "STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in Arabidopsis and rice." *Plant and Cell Physiology* 54, no. 2 (2013): e8-e8.

52. Prabha, Ratna, Indira Ghosh, and Dhananjaya P. Singh. "Plant Stress Gene Database: a collection of plant genes responding to stress condition." *ARPN J. Sci. Technol* 1 (2011): 28-31.
53. Zhang, Xianwen, Jiaping Li, Ailing Liu, Jie Zou, Xiaoyun Zhou, Jianhua Xiang, Wirat Rerksiri, Yan Peng, Xingyao Xiong, and Xinbo Chen. "Expression profile in rice panicle: insights into heat response mechanism at reproductive stage." *Plos one* 7, no. 11 (2012): e49652.
54. Priya, Pushp, and Mukesh Jain. "RiceSRTFDB: a database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis." *Database* 2013 (2013).
55. Das, Gourab, Surojit Das, Shanta Dutta, and Indira Ghosh. "In silico identification and characterization of stress and virulence associated repeats in *Salmonella*." *Genomics* 110, no. 1 (2018): 23-34.
56. Bardou, Philippe, Jérôme Mariette, Frédéric Escudié, Christophe Djemiel, and Christophe Klopp. "jvenn: an interactive Venn diagram viewer." *BMC bioinformatics* 15, no. 1 (2014): 1-7.
57. Smita, Shuchi, Amit Katiyar, Sangram Keshari Lenka, Monika Dalal, Amish Kumar, Sanjeet Kumar Mahtha, Gitanjali Yadav, Viswanathan Chinnusamy, Dev Mani Pandey, and Kailash Chander Bansal. "Gene network modules associated with abiotic stress response in tolerant rice genotypes identified by transcriptome meta-analysis." *Functional & integrative genomics* 20, no. 1 (2020): 29-49.
58. Wan, Quan, Hayley Dingerdissen, Yu Fan, Naila Gulzar, Yang Pan, Tsung-Jung Wu, Cheng Yan, Haichen Zhang, and Raja Mazumder. "BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis." *Database* 2015 (2015).
59. Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even et al. "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification." *Bioinformatics* 21, no. 5 (2005): 650-659.
60. Sasaki, Takuji, and Benjamin Burr. "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome." *Current opinion in plant biology* 3, no. 2 (2000): 138-142.
61. Khandagale, Kiran S., Rahul L. Zanan, Sarika V. Mathure, and Altafhusain B. Nadaf. "Haplotype variation of Badh2 gene, unearthing of a new fragrance allele and marker

development for non-basmati fragrant rice 'Velchi'(Oryza sativa L.)." *Agri Gene* 6 (2017): 40-46.

62. Vemireddy, Lakshminarayana R., Bhaben Tanti, Lipika Lahkar, and Zina M. Shandilya. "Aromatic Rices: Evolution, Genetics and Improvement through Conventional Breeding and Biotechnological Methods." *Molecular Breeding for Rice Abiotic Stress Tolerance and Nutritional Quality* (2021): 341-357.
63. Jackson, Scott A. "Rice: the first crop genome." *Rice* 9, no. 1 (2016): 1-3.
64. Lowe, Todd M., and Sean R. Eddy. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." *Nucleic acids research* 25, no. 5 (1997): 955-964.
65. Šmarda, Petr, Petr Bureš, Lucie Horová, Ilia J. Leitch, Ladislav Mucina, Ettore Pacini, Lubomír Tichý, Vít Grulich, and Olga Rotreklová. "Ecological and evolutionary significance of genomic GC content diversity in monocots." *Proceedings of the National Academy of Sciences* 111, no. 39 (2014): E4096-E4102.
66. Li, Xukai, Kai Guo, Xiaobo Zhu, Peng Chen, Ying Li, Guosheng Xie, Lingqiang Wang, Yanting Wang, Staffan Persson, and Liangcai Peng. "Domestication of rice has reduced the occurrence of transposable elements within gene coding regions." *BMC genomics* 18, no. 1 (2017): 1-12.
67. Lawson, Mark J., and Liqing Zhang. "Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes." *Genome biology* 7, no. 2 (2006): 1-11.
68. Lanciano, Sophie, and Marie Mirouze. "DNA methylation in rice and relevance for breeding." *Epigenomes* 1, no. 2 (2017): 10.
69. Zou, Xin-Hui, Fu-Min Zhang, Jian-Guo Zhang, Li-Li Zang, Liang Tang, Jun Wang, Tao Sang, and Song Ge. "Analysis of 142 genes resolves the rapid diversification of the rice genus." *Genome biology* 9, no. 3 (2008): 1-13.
70. Zhao, Zhixin, Cheng Guo, Sreeskandarajan Sutharzan, Pei Li, Craig S. Echt, Jie Zhang, and Chun Liang. "Genome-wide analysis of tandem repeats in plants and green algae." *G3: Genes, Genomes, Genetics* 4, no. 1 (2014): 67-78.
71. Armstrong, Richard A. "When to use the B onferroni correction." *Ophthalmic and Physiological Optics* 34, no. 5 (2014): 502-508.
72. Bevan, Michael, and Sean Walsh. "The Arabidopsis genome: a foundation for plant research." *Genome Research* 15, no. 12 (2005): 1632-1642.
73. Girin, Thomas, Laure C. David, Camille Chardin, Richard Sibout, Anne Krapp, Sylvie Ferrario-Méry, and Françoise Daniel-Vedele. "Brachypodium: a promising hub

between model species and cereals." *Journal of experimental botany* 65, no. 19 (2014): 5683-5696.

74. Sakai, Hiroaki, Sung Shin Lee, Tsuyoshi Tanaka, Hisataka Numa, Jungsok Kim, Yoshihiro Kawahara, Hironobu Wakimoto et al. "Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics." *Plant and Cell Physiology* 54, no. 2 (2013): e6-e6.
75. Armstrong, Richard A. "When to use the B onferroni correction." *Ophthalmic and Physiological Optics* 34, no. 5 (2014): 502-508.
76. Sato, Yutaka, and Sakiko Yokoya. "Enhanced tolerance to drought stress in transgenic rice plants overexpressing a small heat-shock protein, sHSP17. 7." *Plant cell reports* 27, no. 2 (2008): 329-334.
77. Yilmaz, Alper, Milton Y. Nishiyama Jr, Bernardo Garcia Fuentes, Glaucia Mendes Souza, Daniel Janies, John Gray, and Erich Grotewold. "GRASSIUS: a platform for comparative regulatory genomics across the grasses." *Plant physiology* 149, no. 1 (2009): 171-180.
78. Weng, Xiaoyu, Lei Wang, Jia Wang, Yong Hu, Hao Du, Caiguo Xu, Yongzhong Xing, Xianghua Li, Jinghua Xiao, and Qifa Zhang. "Grain number, plant height, and heading date7 is a central regulator of growth, development, and stress response." *Plant physiology* 164, no. 2 (2014): 735-747.
79. Jardim-Messeder, Douglas, Andréia Caverzan, Rafael Rauber, Eduardo de Souza Ferreira, Márcia Margis-Pinheiro, and Antonio Galina. "Succinate dehydrogenase (mitochondrial complex II) is a source of reactive oxygen species in plants and regulates development and stress responses." *New Phytologist* 208, no. 3 (2015): 776-789.
80. Zeng, Houqing, Luqin Xu, Amarjeet Singh, Huizhong Wang, Liqun Du, and B. W. Poovaiah. "Involvement of calmodulin and calmodulin-like proteins in plant responses to abiotic stresses." *Frontiers in plant science* 6 (2015): 600.
81. Campo, Sonia, Patricia Baldrich, Joaquima Messeguer, Eric Lalanne, María Coca, and Blanca San Segundo. "Overexpression of a calcium-dependent protein kinase confers salt and drought tolerance in rice by preventing membrane lipid peroxidation." *Plant physiology* 165, no. 2 (2014): 688-704.
82. Liu, Huizhi, Xiaoe Wang, Huijuan Zhang, Yayun Yang, Xiuchun Ge, and Fengming Song. "A rice serine carboxypeptidase-like gene OsBISCPL1 is involved in regulation

of defense responses against biotic and oxidative stress." *Gene* 420, no. 1 (2008): 57-65.

83. Rouyi, Chen, Supaporn Baiya, Sang-Kyu Lee, Bancha Mahong, Jong-Seong Jeon, James R. Ketudat-Cairns, and Mariena Ketudat-Cairns. "Recombinant expression and characterization of the cytoplasmic rice β-glucosidase Os1BGlu4." *PloS one* 9, no. 5 (2014): e96712.
84. Sharma, Rita, Peijian Cao, Ki-Hong Jung, Manoj K. Sharma, and Pamela C. Ronald. "Construction of a rice glycoside hydrolase phylogenomic database and identification of targets for biofuel research." *Frontiers in plant science* 4 (2013): 330.
85. Zhao, Min-Gui, Qiu-Ying Tian, and Wen-Hao Zhang. "Nitric oxide synthase-dependent nitric oxide production is associated with salt tolerance in Arabidopsis." *Plant physiology* 144, no. 1 (2007): 206-217.