ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(36), DOI: 10.17485/ijst/2019/v12i36/112312, September 2019*

Genome-wide Landscaping of Repetitive Elements in the Genome of *Neurospora crassa*

Kunal Zaveri[1], K. Vijaya Rachel[1] and P. Kiranmayi[2]*

[1]Department of Biochemistry and Bioinformatics, Institute of Science, GITAM (Deemed to be University), Visakhapatnam - 530045, Andhra Pradesh, India; kunal.zaveri22@gmail.com, rachelr68@gmail.com
[2]Department of Biotechnology, Institute of Science, GITAM (Deemed to be University), Visakhapatnam - 530045, Andhra Pradesh, India; kiranmayi.patnala@gmail.com

Abstract

Objectives: Repetitive elements are ubiquitous in eukaryotes and have contributed in expression of genes and maintaining the architecture of genome. Fungi although being a eukaryote, repeat sequences are typically limited due to small genome size and typical genome defense mechanisms which fights against expansion of repeats and maintains streamlined genome. **Methods/Findings:** Despite, the presence of such mechanisms *Tad* element was found to be active in Adiopodoume strain of *N. crassa*. This is one of the most enigmatic discoveries made in *Neurospora.* From this perspective, we have analyzed the genome of *N. crassa* OR74A (NC12) by *in silico* tools for the identification of repeats. The comprehensive analysis of these elements suggests that about six (*Punt*^{RIP}) DNA transposons and 2 *Tad1* (Non-Long Terminal Repeat) elements share highest homology with known repeats from RepBase. The presence of DNA binding and transposase domain in the protein sequence of NCU02991 indicates that the element identified is a DNA transposon. **Application:** The results also suggest that among the identified six *Punt*^{RIP} sequences four of them share homology, which indicates that similar sequence is repeated four times within the genome. The promoters identified in the dataset of identified repeats in the upstream and downstream regions indicate that these repeats may play certain role in the regulation of genes.

Keywords: DNA Transposons, Non-LTR, Repetitive Elements, *Tad1*

1. Introduction

Non-coding DNA, rather being useless was thought to be materialistic remnants from the variations caused in the evolutionary process. But, the current developments suggest that non-protein-coding DNA are helpful for genome in directing various other information such as their role in genome structure, function, gene regulation, rapid speciation and genome defense systems. These non-coding regions consist of repetitive elements which are the copies of nucleotides that repeat throughout the genome. Classically, repeats are classified into three main types: Terminal repeats, Tandem repeats and interspersed repeats, of which interspersed repeats are very well studied. Accomplishment of human genome project and other 150 eukaryotic genome sequencing projects have opened a new era in the dawn of 21st century. This new era emphasizes the study related to function and importance of repeats. According to the selfish or parasitic DNA theory, non-coding DNA persists only because of its power to replicate itself, or it has mutated into a form advantageous to the cell. Eukaryotic genome constitutes millions of copies of repeats[1] that might play a crucial role in various metabolic processes. Understanding and analyzing such a huge content of uncharacterized repeats

*Author for correspondence

in the genome opens up new mechanisms that play a crucial role in the regulation of genes in eukaryotes[2].

Blazing advances in the field of genetics, genomics, and molecular biology with the support of Bioinformatics studies, allowed scientists and researchers to fathom in detail about the numerous features of repetitive elements. The pervasive idea proposed by the Nobel laureate Barbara McClintock on Transposable elements[3] has completely changed ideology of researchers. Since 1950's, much research has taken place on repetitive elements *viz.,* recombination of Alu elements that have played an important role in the evolution of human glycophorein gene family[4,5], presence of Alu element in the third intron of ornithine aminotransferase[6] and presence of SINE in coding region of mRNA of bovine prostaglandin E2 receptor etc[7]. Similarly ENCODE (Encyclopedia of DNA Elements) project has deduced that non-coding regions of DNA play an important role in regulatory mechanisms[8]. All this research insinuates that repetitive elements are no longer said to be junk elements.

Repetitive elements are capable of self-replicating, but certain genomes tend to fight against their expansion by three known pathways of gene silencing mechanisms *viz.,* Quelling, RIP (Repeat Induced Point Mutation)[9] and MSUD (Meiotic Silencing by Unpaired DNA)[10]. These mechanisms are very well opted by all types of filamentous fungi, of which, *Neurospora* provided the first example for eukaryotic genome defense system. In *Neurospora*, Quelling co-operates with RIP and MSUD in controlling the expansion of transposons within the genome. Quelling is a Post-Transcriptional Gene Silencing mechanism (PTGS) observed in *Neurospora,* which is similar to co-suppression in plants and RNA interference (RNAi) in animals[11].

Despite having such strong defense mechanism, an active non-LTR known as *Tad* element was identified from Adiopodoume strain of *Neurospora*[12], which is capable of missing the RIP pathway[13]. An inactive non-LTR known as *Punt*^{RIP} element was also identified in *N. crassa* in the methylated pseudogene which is inactivated by RIP[14]. Based on this, the present work was carried out to identify the presence of repetitive elements in *N. crassa* OR74A (NC12) strain with the help of *in silico* tools. This study will shed a light on the presence of non-coding elements in *N. crassa* and further it will be helpful in understanding the role of non-coding DNA and its capability to skip-out from the strong genome defense mechanisms in *N. crassa.*

2. Materials and Methods

Genome sequence of *N. crassa* was obtained from the sequencing consortia: Fungal Genome Initiative (BROAD Institute)[15], now the sequence information is available at FungiDB[16]. Retrieval of all the sequences was performed before August 26th, 2013 which was publicly available. To detect the sequences corresponding to repetitive elements two programs were applied on genome sequence of *N. crassa.* The combination of two tools is applied, since there are no specific tools to identify repetitive elements in fungal genomes. The complete genome of *N. crassa* was analyzed for repetitive elements using RepeatMasker[17] and Censor[18] with RepBase library as reference dataset, a manually curated repetitive sequence library. The size and homology in the repeat identification were low cut-off thresholds for Censor (uses DASHER3 algorithm) (4.5), low ratio of mismatches to transitions (2:1) and a relatively high LOCAL alignment score (30.0). In both the programs sequence source was set to fungi, RepeatMasker uses HMMER whereas Censor uses WU-BLAST (Washington University -Basic Local Alignment Search Tool) for prediction of repetitive elements.

Results from two programs were merged, yielding a set of repetitive elements. Nucleotide sequences harboring least score and low similarity with that of reference sets were discarded for further analysis. The nucleotide sequences were subjected to BLAST to predict the function and were then translated to protein sequences. For each translated sequence, the respective domains were identified by ProDom[19]. Figure 1 represents the step by step process for identification of the repetitive elements in *N. crassa.*

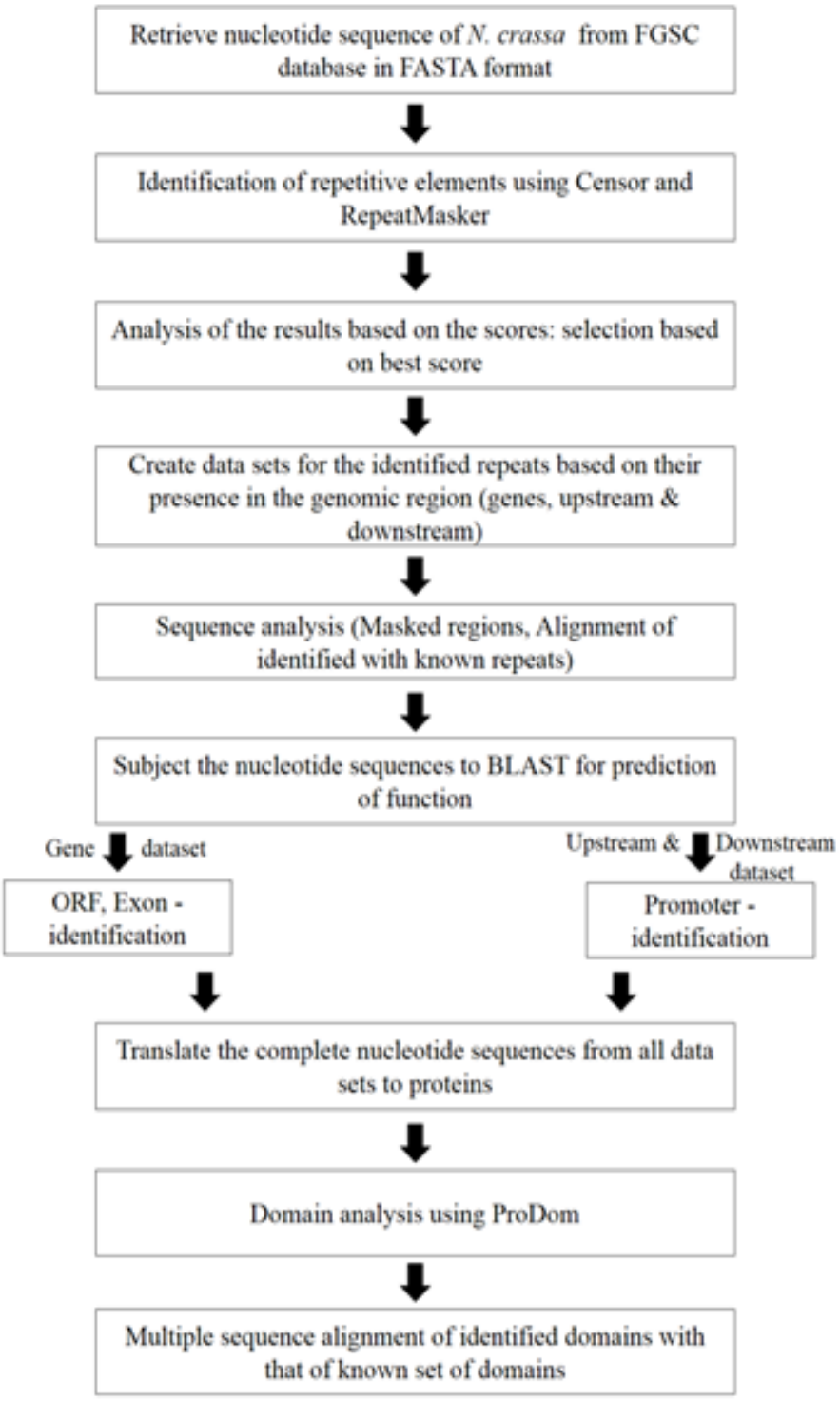

Figure 1. Methodology for prediction of repetitive elements in *Neurospora crassa.*

Figure shows the step by step process for identification of repetitive elements using in silico

3. Results and Discussion

In the arena of DNA sequencing, the faster and cheaper technologies for sequencing have led to the challenge of annotating the sequencing data, which includes both coding and non-coding regions. The availability of complete genome sequence of *N. crassa* and identification of *Tad* element and *Punt*^{RIP} element sensitive to RIP pathway has driven the present study for identification of repetitive elements in *N. crassa*.

3.1 General Genome Features of *Neurospora crassa*

The *N. crassa* genome was assembled into 21 scaffolds with N50 approximately of 6.07 Mb encompassing 39Mb. Approximately 9730 genes were predicted, including 9334 orthologous genes and 16 pseudogenes Table 1.

3.2 Identification of Repetitive Elements

In the present study identification of repetitive elements in *N. crassa* was carried out through series of steps. A total of 9730 genes constituting a nucleotide sequence of 23,369,813bp were retrieved from BROAD institute. Similarly, for each gene their upstream and downstream sequence lengths of 1,000bp each were also retrieved. Since there are no precise tools for predicting the fungal specific repetitive elements, a combined approach using Censor and RepeatMasker were employed to identify the repeats in the genome of *N. crassa*. The resultant data was mined based upon their occurrence in different regions (genes, upstream and downstream) of the genome. The *in silico* predictions of repetitive elements in *N. crassa* have retrieved two different types of elements based on the sequence similarity with the elements present in the RepBase database.

3.3 Annotation of Identified Repeats

The repeat families that were identified are categorized into different datasets. From these datasets, we have manually reviewed them based upon their score of similarity with that of the known repeats. The nucleotide sequences that were found to have least scores with less similarity or no similarities against reference data sets were discarded for further analysis.

Table 1. Main features of the *Neurospora crassa* genome

Feature	*N. crassa*
Total sequence length	41,102,378 bp
Total assembly gap length	40,775 bp
Gaps between scaffolds	0
Number of scaffolds	21
Scaffold N50	6.07 Mb
Number of contigs	412
Total number of chromosomes	7
tRNA genes	415
Pseudogenes	16
GC content	49.3%

Legend: Table represents the main features of *N. crassa* genome (bp: base pairs; Mb: Mega Base pairs)

Table 2. Repetitive elements identified

S.No	Genomic region	FGSC GENE ID	Type of the Repeat Identified	Similar to the known repeat (RepBase)	Score
1.	Gene	NCU02991	*Puntrip*/DNA Transposon	PUNTRIP_NC	2569
2.	Upstream to	NCU05159	*Puntrip*/DNA Transposon	PUNTRIP_NC	3762
3.	Upstream to	NCU07945	*Puntrip*/DNA Transposon	PUNTRIP_NC	3783
4.	Upstream to	NCU16528	*Tad1*/Non-LTR	Tad1-1_NC	2993
5.	Downstream to	NCU02990	*Puntrip*/DNA Transposon	PUNTRIP_NC	2050
6.	Downstream to	NCU07896	*Puntrip*/DNA Transposon	PUNTRIP_NC	7104
7.	Downstream to	NCU09994	*Puntrip*/DNA Transposon	PUNTRIP_NC	3301
8.	Downstream to	NCU03846	*Tad1*/Non-LTR	Tad1	2179

Legend: Table represents the repetitive elements identified in the genome of N. crassa.

The repetitive elements that were identified in the different regions of genome were observed to be similar with known active non-LTR *Tad1* and non-active DNA transposon *Punt*^{RIP} of *N. crassa* Table 2. The nucleotide sequences of gene NCU02991, upstream regions of NCU05159, NCU07945 and downstream region of NCU02990, NCU07896, NCU09994 were found to be similar to the non-active DNA transposon *Punt*^{RIP}, which was originally found in the 5s rRNA pseudogene of *N. crassa*. The only active repetitive element found in *N. crassa* is *Tad1*, a LINE like non-LTR[20]. By *in silico* analysis we could retrieve similar kind of nucleotide sequences in the upstream region of NCU016528 and to the downstream region of NCU03846.

These nucleotide sequences were further analyzed for the sequence similarity with that of known repeats by local alignment and the masked sequences were obtained. The alignment files and the masked regions of the sequences are given in the *additional file 1*. With the sequence similarity, we have even identified the start and end positions of repeats in the nucleotide sequences with that of reference repeats Table 3. The sequence analysis indicates that the identified repeats are about >70% similar to that of known repeats.

The nucleotide sequences of these identified repeats were subjected to functional analysis by BLAST. The nucleotide sequence of NCU02991 being a gene have retrieved the homology with itself, which is annotated to be a hypothetical protein with a function to bind DNA. Interestingly, the nucleotide sequences of upstream to NCU07945 and down streams of NCU02990 and NCU07896 were also observed to share homology with NCU02991. As it is known that DNA binding is one of the integral domain of the DNA transposons[21], it suggests that the above-mentioned sequences may be DNA transposons. The nucleotide sequence upstream to NCU16528 shared homology with pol-like protein, downstream of NCU05159 shared homology with NCU05158, a

Table 3. The start and end regions of repeats identified

Analysis of identified repeats in genome of *N. crassa*				Similar to nucleotide sequences of known repeat elements from RepBase			
Region	**FGSC Gene ID**	**Start**	**End**	**Name**	**Start**	**End**	**Class**
Gene	NCU02991	1	383	PUNTRIP_NC	1348	1731	DNA transposon
	NCU02991	386	975	PUNTRIP_NC	1	587	DNA transposon
Upstream	NCU07945	6	948	PUNTRIP_NC	930	1873	DNA transposon
	NCU09994	5	999	PUNTRIP_NC	65	1048	DNA transposon
	NCU16528	1	448	Tad1-1_NC	1817	2261	NonLTR/Tad1
Downstream	NCU02990	576	996	PUNTRIP_NC	1456	1873	DNA transposon
	NCU03846	43	599	Tad1-1_NC	5932	6490	NonLTR/Tad1
	NCU05159	3	907	PUNTRIP_NC	1	901	DNA transposon
	NCU07896	216	1000	PUNTRIP_NC	1089	1874	DNA transposon

Legend: The start and end positions of identified repeats with that of reference repeats

hypothetical protein whereas no homologies were observed for upstream to NCU09994 and downstream of NCU03846.

The results also suggest that the similar sequence of *Punt*^{RIP} was found to be repeated for four times within the genome. These four regions include, gene NCU02991, upstream to NCU07945 and down streams of NCU02990 and NCU07896, as their sequences have shared homology. Further these sequences were subjected to multiple sequence alignment using T-COFFEE[22]. The overall alignment of the four sequences on bad to good scale was observed to be average with a total score of 310. The multiple alignment results also reveal that most of the

In the gene sequence of NCU02991 the ORF and exon regions were predicted to be spanning in the region of 1-813bp. Furthermore, the ORF and exon

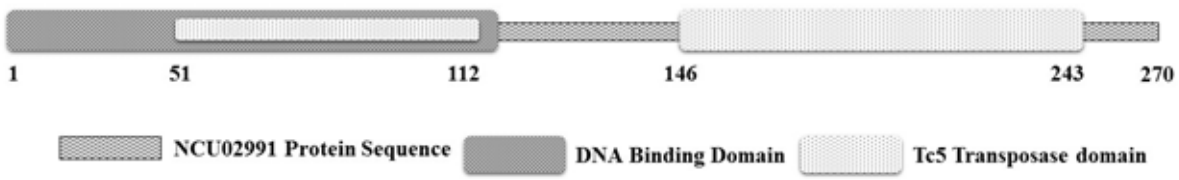

Figure 2. Domains identified in the protein sequence of NCU02991.
A protein sequence of NCU02991 with 270 amino acids is predicted to have two domains: DNA binding (1-115) and Tc5 Transposase (51-122; 146-270)

Table 4. Promoters predicted in the dataset of identified repeats in upstream and downstream regions

Region	FGSC Gene ID	Promoter site (Start-End)	Promoter sequence
Upstream	NCU05159	202 - 252	TTAAGAAGTTTTAAGAAGGGGGGTAAGCTCTAAAA ACCTT**A**AATTTAGGT
		340 - 390	GCTATTACTATTTAAAACGCTACTTAGGGACGATA TATTT**A**AGGCCAATT
		365 - 415	AGGGACGATATATTTAAGGCCAATTATACTTTTAA AATTC**T**TTAGAAGTC
		589 - 639	AAGGAAGGTTAATAAAGAGGGGGGCAATAGTAGAT TTATA**T**ATATAGAGG
		617 - 667	AGTAGATTTATATATATAGAGGAGATTATTTAAGT TTGTG**A**AGCCTTAAA
	NCU07945	612 - 662	TATCCATTTGTATAAGAAGGATTCTTGGGCGGGGG ATAGC**T**ATTGTTGGG
	NCU16528	298 - 348	TTTATTTGCATATATTACCGCCCTTCCCTTCTCCA AAGAA**A**ATTAACCCC
Downstream	NCU02990	-	----------
	NCU03846	167 - 217	TGATAACCTTATAGAAGGCGTTAGGGGAGGCGGGC CTTAC**G**CTGGTCTTC
		276 - 326	CGGGTTAGTATATAAATTAGGAAAAATAGGGGAGA AAATT**A**TACTAGTTC
		544 - 594	AGGAGGATTATATAAATAGAGGAGAAAGTCTTTAT ATTAT**T**ATAGATAAG
	NCU07896	136 - 186	AAAGCGGAGAAAAAAAAGACGGAAGGACCAAAAAA CATAC**A**ACACCAGGG
		602 - 652	GGCTTTATAATAGCGATATTAAAAAGAGGAGGCAA TTATA**A**TAAGTTGGG
		610 - 660	AATAGCGATATTAAAAAGAGGAGGCAATTATAATA AGTTG**G**GGGCGCATT
		628 - 678	AGGAGGCAATTATAATAAGTTGGGGGCGCATTGGG TAGAC**T**GTTTTATAA
		894 - 944	AGAAAGTATATTTAATAACGACCGATATAACAATT TGGGT**A**ACTATTATT
	NCU09994	870 - 920	TCGTCAAATATTAAAAAACGCTAACTACTATTAAT ATAAA**T**CTTAAGAAG

Legend: Start and End sites of promoters in identified repeats, in the sequence bold and big font letter represents transcription start site

regions were observed to be within the repeat region of 1-975bp. This indicates that repeat region is playing a crucial role for the protein coded by this gene. Hence, to understand the function of its translated sequence, the protein sequence of NCU02991 was retrieved and was subjected to ProDom for domain analysis. DNA binding homeodomain and Tc5 transposase domain were the two domains predicted for NCU02991. These domains were schematically represented based on their positions in the protein sequence of NCU02991 Figure 2. The DNA binding domain uses helix-turn-helix structure for DNA recognition[23] and the transposase domain acts as a catalyst for cut and paste mechanism[24]. These are the basic characteristics of DNA transposons and are predicted in the protein of NCU02991 which substantiates that NCU02991 belongs to the class DNA transposon.

Similarly, for the identified repeats in the upstream or downstream of genes, promoters and the transcription start sites were predicted. Except in the nucleotide sequence which is downstream to NCU02990 the promoter and TSS could not be predicted whereas in other six predicted repeats both promoters and TSS have been predicted Table 4. The prediction of promoters and transcription start site indicates that these repetitive elements may have certain role in regulation of the genes.

4. Conclusion

Repetitive elements were initially considered to be a part of junk DNA but after an incessant encroachment in the arena of DNA sequencing and ENCODE projects, suggest that repetitive elements have a role in regulation. A systematic screening for identification of repetitive elements in the genomic sequence of the non-pathogenic fungus *N. crassa* was carried out. The screening strategy predicted 8 repetitive elements in the genome of *N. crassa* of which six are DNA transposons (*Punt*^{RIP} type) and two are non-LTRs (*Tad1* type). The DNA transposon, *Punt*^{RIP} was identified in the gene of [FGSC: NCU02991] which encodes for a hypothetical protein and the domain analysis of this protein revealed the presence of DNA binding and transposase domain, which are the common domains in the transposons. The promoters with transcription start sites were identified in the dataset of repeats that are identified in the upstream and downstream regions, which indicates that the repeats that are identified may play role in the regulations of certain genes.

5. List of Abbreviations

1. FGSC: Fungal Genetics Stock Centre
2. LTR: Long Terminal Repeats
3. Non-LTR: Non-Long Terminal Repeats
4. RIP: Repeat Induced Point Mutation

6. Acknowledgement

This work was supported and funded by University Grants Commission (UGC) with grant number F.No.42-669/2013 to PK. The computational part of this work was executed using the Sun Workstation provided by the Bioinformatics facility from the Department of Biochemistry and Bioinformatics, Institute of Science, GITAM University. Authors would also like to thank Department of Biotechnology, Institute of Science, and GITAM University for providing all the necessity facilities. PK acknowledges the support from a Project operated within the UGC-Major Research Project Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7. References

1. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over Two-Thirds of the human genome. PLoS Genetics. 2011; 7.
2. Kazazian HH. Mobile elements: Drivers of genome evolution. Science. 2004; 303:1626–32.
3. McCLINTOCK B. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences of the U S A. 1950; 36:344–55.
4. Kazazian HH. Mobile elements and disease. Current Opinion in Genetics and Development. 1998; 8:343–50.
5. Deininger PL, Batzer MA. Alu repeats and human disease. Mol Genet Metab. 1999; 67:183–93.

6. Mitchell GA, Labuda D, Fontaine G, Saudubray JM, Bonnefont JP, Lyonnet S. Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proceedings of the National Academy of Sciences of the United States of America. 1991; 88:815–9.
7. Shimamura M, Nikaido M, Ohshima K, Okada N. Letter to the Editor A SINE that acquired a role in signal transduction during evolution; 1996. p. 923–5.
8. Qu H, Fang X. A brief review on the human encyclopedia of DNA elements (ENCODE) Project. Genomics, Proteomics and Bioinformatics. 2013; 11:135–41.
9. Selker EU, Cambareri EB, Jensen BC, Haack KR. Rearrangement of duplicated DNA in specialized cells of Neurospora. Cell. 1987; 51(5):741–52.
10. Shiu PK, Raju NB, Zickler D, Metzenberg RL. Meiotic silencing by unpaired DNA. Cell. 2001; 107(7):905–16.
11. Fulci V, Macino G. Quelling: Post-transcriptional gene silencing guided by small RNAs in Neurospora crassa. Current Opinion in Microbiology. 2007; 10:199–203.
12. Kinsey JA, Helber J. Isolation of a transposable element from Neurospora crassa. Proceedings of the National Academy of Sciences of the United States of America. 1989; 86:1929–33.
13. Cambareri EB, Foss HM, Rountree MR, Selker EU, Kinsey JA. Epigenetic control of a transposon-inactivated gene in Neurospora is dependent on DNA methylation. Genetics. 1996; 143:137–46.
14. Margolin BS, Garrett-Engele PW, Stevens JN, Fritz DY, Garrett-Engele C, Metzenberg RL. A methylated neurospora 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation. Genetics. 1998; 149:1787–97.
15. McCluskey K, Wiest A, Plamann M. The fungal genetics stock center: A repository for 50 years of fungal genetics research. Journal of Biosciences. 2010; 35:119–26.
16. Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, Harb OS. Fungi DB: An integrated functional genomics database for fungi. Nucleic Acids Res. 2012; 40.
17. RepeatMasker Open-3.0 [Internet]. [cited 2019 May 20]. Available from: http://www.repeatmasker.org.
18. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: Rep base Submitter and Censor. BMC Bioinformatics. 2006; 7:474.
19. Corpet F, Servant F, Gouzy J, Kahn D. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. Nucleic Acids Research. 2000; 28:267–9.
20. Cambareri EB, Helber J, Kinsey JA. Tad1-1, an active LINE-like element of Neurospora crassa. Molecular Genetics and Genomics. 1994; 242:658–65.
21. Watkins S, van Pouderoyen G, Sixma TK. Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. Nucleic Acids Research. 2004; 32(14):4306–12.
22. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology. 2000; 302(1):205–17.
23. Siegmund T, Lehmann M. The Drosophila Pipsqueak protein defines a new family of helix-turn-helix DNA-binding proteins. Development Genes and Evolution. 2002; 212:152–7.
24. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proceedings of the National Academy of Sciences of USA. 2011; 108(19):7884–9.