IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Stemformatics And Its Potential Role In Future Translational Applications In CRISPR CAS 9 Gene With GATA-1, 2"

Deepika Pal[1, 2*], Utkarsh Tyagi[1*],Udit Narayan Sharma[1, 2, 3] and Eliza Chakraborty[1, 2 **]

1. Medical Translational Biotechnology Lab, Meerut Institute of Engineering &Technology (MIET), Meerut-250005, U.P, INDIA.

2. DST- Fist Center, Sponsored by Department of Science and Technology, Ministry of Science and Technology. Govt. of India at Meerut Institute of Engineering and Technology Meerut Institute of Engineering & Technology, Meerut- 250005, U.P, INDIA.

3. Department of Polymer and Process Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, Uttarakhand, India.

Abstract

Stemformatics is a newly introduced platform of bioinformatics which deals with stem cell related *in-silico* information and a well-known source of gene expressiondatasets that have been compiled using single cell profiling, RNA sequencing, and microarray technologies.It is also a well known data portal of stem cell gene expressions. It allows for simple to view and compare gene expression profiles in mouse and human across many platforms from various lab sources. Our group also has preliminary omics study of translation applications in CRISPR CAS 9 Gene by exploiting Stemformatics tool viz: phylogeny along with the transcriptional factors and CpG islands analysis.Using "molecular scissors (CRISPR)," genome editing involves adding, and replacing DNA in a living organism's genome. Traditional genome editing for human stem cells using designed nucleases has limitations by its ineffectiveness, Cost effective, and lack of selectivity. The CRISPR system has recently come into prominence as a potent gene-editing method with benefits of high editing efficacy. Despite the enormous potential for gene modification in a variety of species, from prokaryotes to higher mammals. The sample annotations are not adequately standardized, the abundance of primary data in stemformatics creates difficulties for data aggregation and downstream analysis. So, in this article, we focused on the underlying biology and use of the CRISPR/Cas9 system in current human stem cell research, *in-silico* evaluations of its advantages and future applications of human stem cells in regenerative medicine.

Keywords: Stemformatics, CRISPR, CAS 9 Gene, CpGI, Phylogenetic tree, Embryo editing, Stem cell and Regenerative medicines.

1. Introduction

1.1 CRISPR

CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats in accordance to Khan et al. (2018). CRISPR is the name of a family of DNA sequences discovered in bacteria. The sequences contain bits of DNA from viruses that have attacked the bacterium. The bacteria utilize these fragments to recognize and eliminate DNA from related viruses during subsequent assaults. These sequences are crucial components of the bacterial defense mechanism (Zhang et al. 2022).

The initial findings were made in 1987 at Japan's Osaka University. The sequence of the "iap" gene and its relationship to E. coli were the subject of a study by YoshizumiIshino and colleagues, which was then published. With the aid of a method known as metagenomics, technological advancements in the 1990s allowed them to speed up their sequencing and continue their study. They may take samples of dirt or ocean and sequence the DNA in those samples. (Ishino et al 2018).

1.2 CAS 9

Streptococcus pyogenes, like other bacteria, possesses an adaptive defence mechanism called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). RNA-guided DNA endonuclease Cas9 (CRISPR associated protein 9) is an enzyme. S. pyogenes uses Cas9 to recall things (Heler et al. 2015). In addition to its original function in bacterial immunity, the Cas9 protein has been extensively exploited as a genome engineering tool to generate site-directed double strand breaks in DNA. In many laboratory model organisms, these fractures can result in non-homologous end joining that inactivates genes or homologous recombination that introduces heterologous genes. Cas9 is emerging as a key tool for genome editing, alongside TALEN proteins and zinc finger nucleases.

1.3 CRISPR-CAS 9 genome editing and stem cells

Traditional genome editing for human stem cells using designed nucleases is constrained by its ineffectiveness, high expense, and lack of selectivity. The CRISPR system has lately come into prominence as a potent gene-editing tool with benefits of high editing effectiveness and cheap cost (Heler et al 2015). Using CRISPR/Cas9 to manipulate genes, such as gene knockdown, numerous chromosome-related applications, including gene knockin, gene interference or activation, have been used in biological and scientific research (Zhang et al 2017).

Many different stem cell types have received approval for use in medical therapies, and many have had outstanding results in clinical studies. More and more evidence is pointing to the effectiveness of CRISPR/Cas9 genome editing as a significant tool for advancing stem cell research, from fundamental biology to translational investigations (Wang et al 2016).

1.4 Stemformatics

Collaboration between the stem cell and bioinformatics communities led to the development of stemformatics by Professor Christine Wells and her group. The concept of creating stemformatics was inspired by the abundance of fascinating cell models available in both public and private territories, as well as the fact that many biologists do not have easy access to these models. The data on the Stemformatics website has all been manually selected, vetted, examined for experimental repeatability and quality of design, and internally normalized. Additionally, the RohartMesenchymal Stromal Cells (MSC) test provides a case study of how to develop an algorithm to categorise stem cells that behave like MSCs using well selected data and metadata (Rivera et al, 2017). High-quality gene expression and annotation data that are rapid and simple to use and are of interest to stem cell biologists are at the core of stemformatics.

The stemformatics workbench features help people learn the following information:

- The gene's expression pattern (YuGene graph) Additional genes that exhibit the same behaviour as the target gene (Gene Neighbourhood)
- The genes that can distinguish these cell types most effectively (Comparative Marker Selection)
- The samples' similarity (Hierarchical Cluster)
- Gene set-related pathway (Annotate a Gene Set)
- The Rohart MSC test, which uses an analysis of MSC data and a related categorization tool.
- The stem cell community-relevant datasets and analysis will continue to be added to Stemformatics.

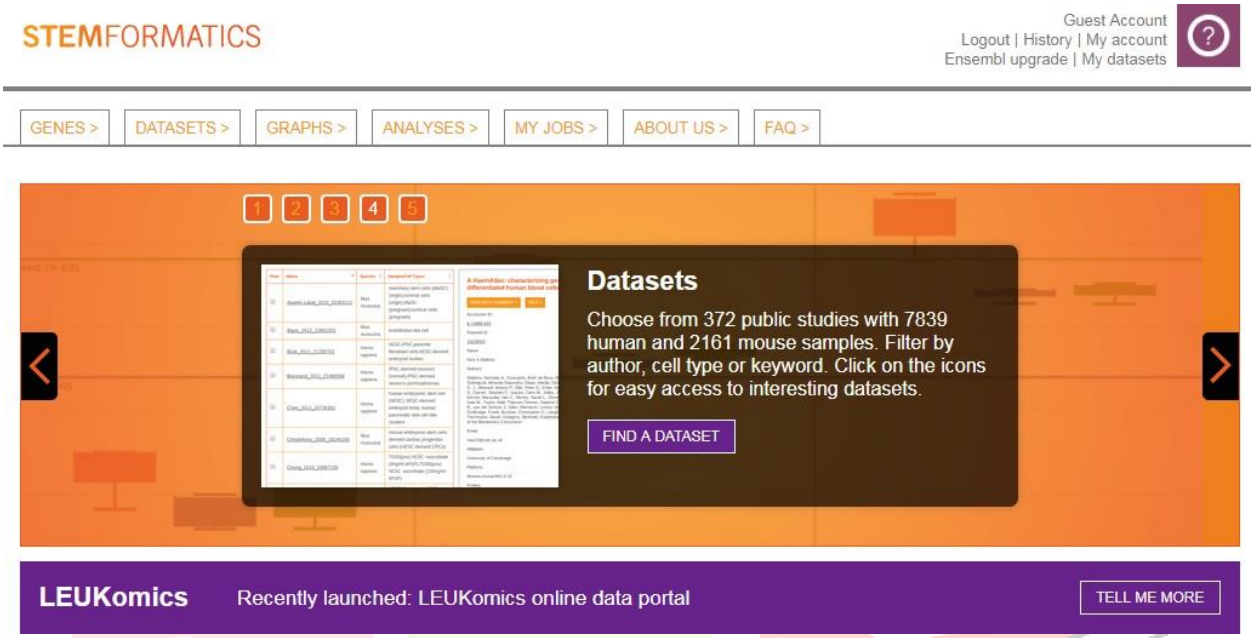


Fig: 1.3.1 Introduction to Stemformatics (www.stemformatics.org)

It is made up of a curated atlas of stem cells with more than 800 stem cell datasets, which gathers transcriptome, proteome, and epigenome information from tens of thousands of stem cell samples. The atlas offers quick and simple views of specific genes, pathways, or model public projects for stem cell researchers. For the computational biology community interested in method development or integrative studies, the atlas also offers a collection of high quality; thoroughly curated datasets (Rajab et al., 2021).The stem cell research community has difficulties as a result of this data, including locating and accessing material that is trustworthy and pertinent to the current biological topic (Choi et al, 2019). The Stemformatics site was developed to provide simple tools for easy visualization and comparison of gene expression output across platforms, laboratories, and cell models. This will help to future-proof current stem cell datasets.

1.5 Transcription Factor

By regulating the rate of transcription, they are factor proteins that bind to DNA to influence gene expression (Latchman, 1997; Karin, 1990). Multipotency of a stem cell can be achieved by some transcription factors such as GATA 1 and GATA 2. Due to their significance as important regulators of hematopoiesis and their dynamic expression, the hematopoietic transcription factors GATA1, GATA2, provide an appealing trio for analysing differentiation. GATA1 is predominantly linked to erythroid cells and megakaryocytes, whereas GATA2 is primarily linked to stem cells and multipotent progenitors.GATA1 and GATA2 are related transcription factors that recognize comparable DNA patterns and function in a coordinated manner to coordinate extensive programmes of gene activation and repression during the formation of the blood. (Doré and Crispino, 2011)

1.6 CpG Island (CpI)

Clusters of CpGdinucleotides called CpG islands (CGIs) are an essential component of the mammalian genome(Bird et al., 1986). Even though about 70–80% of the methylatable sites are really in the methylated state, the dinucleotide CpG carries all of the 5-mC. The capacity of various restriction enzymes to discriminate between the methylation and unmethylation versions of the same sequence has been primarily used in research on the relative distribution of methylated cytosines in DNA (Antequera et al.,1993).In 2002, additional GC-rich genomic sequences, such as Alu repeats, were removed from the criteria for CpG island prediction. A thorough analysis of the complete sequences of human chromosomes 21 and 22 revealed that DNA regions longer than 500 bp had a higher likelihood of being the "true" CpG islands connected to the 5' regions of genes if they had a higher GC content than 55% and a higher observed-to-expected CpG ratio than 65%. (Saxonov et al 2006). The term "CG suppression" refers to the fact that CpG islands have CpG dinucleotide content that is at least 60% more than what would be statistically anticipated (around 4-6%), compared to the remainder of the genome, which has a significantly lower CpG frequency (about 1%). Contrary to CpG sites in a gene's coding region, most CpG sites in a gene's promoter's CpG islands are unmethylated if the gene is expressed. This finding inspired the hypothesis that gene expression may be inhibited by the methylation of CpG sites in a gene's promoter. (Takai at el 2002) The majority of the methylation variations across tissues, or between normal and cancer samples, take place close to the CpG islands (also known as the "CpG island shores") as opposed to the islands themselves(Feil et al 2007).

2. Material and Methods

Data collections of Five species were selected ranges from lower to higher vertebrates including mammals) containing transcription factor GATA-1 and GATA-2 namely:

2.1 GATA-1

1. RattusNorvegicus, Norway rat (Accession number: NM_012764.1)

2. Ovis Aries, Sheep (Accession number: AB612140.1)

3. Musmusculus, house mouse (Accession number: NM_008089.2)

4. Homo sapiens, human (Accession number: NM_002049)

5. Gallus gallus, chicken (Accession number:NM_205464.1)

2.2 GATA-2

1. RattusNorvegicus, Norway rat (Accession number: AY032734.1)

2. Ovis Aries, Sheep (Accession number: AB612141.1)

3. Musmusculus, house mouse (Accession number: NM_008090.5)

4. Homo sapiens, human (Accession number: KU892678.1)

5. Gallus gallus, chicken (Accession number: NM_001003797.1)

2.3 Tool used

MEGA 7.0 is free software that may be used to build phylogenetic trees and undertake statistical analyses of molecular evolution. Additionally, it is used to do manual and automated sequence alignment, mine online databases, determine the pace of molecular evolution, and test evolutionary ideas. In partnership with his graduate student Sudhir Kumar and post-doctoral colleague Koichiro Tamura, Masatoshi Nei, of Pennsylvania State University, spearheaded the initiative for creating this programme. In a monograph, Nei described the software's features and introduced brand-new statistical techniques that were integrated into MEGA. The complete collection of software

was created by Kumar and Tamura. MEGA has undergone several updates and expansions, and all of these versions are now accessible through the MEGA home website (www.megasoftware.net). The approach is frequently employed and referenced by academics and experts.

Primer3 is the foundation of MethPrimer, a tool for building PCR primers for methylation mapping. It begins by looking for potential CpG islands in a DNA sequence that is supplied. Then, primers are chosen around user-specified locations or on anticipated CpG islands. Results of primer selection are displayed in a web browser in text and graphic form. The input sequence is a DNA sequence in any format. No editing is required prior to entry.

2.4 Construction of Phylogenetic Tree

1. The GATA-1 and GATA-2 DNA sequences of different species for phylogenetic tree construction were obtained from the NCBI database and stored in a file(http://www.ncbi.nlm.nih.gov/)

2. The sequences were aligned using MEGA as shown below

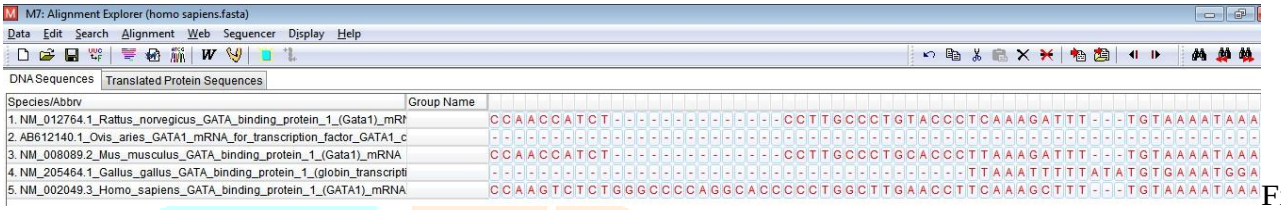


Fig2.1-ScreenshotofsequenceofGATA-1forphylogeneticanalysis

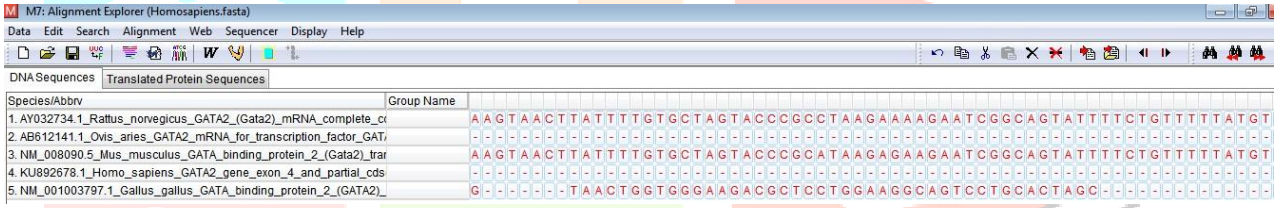


Fig2.2-ScreenshotofsequenceofGATA-2forphylogenetictreeanalysis

Phylogenetictreewascomputedusingneighborjoiningmethodin MEGA

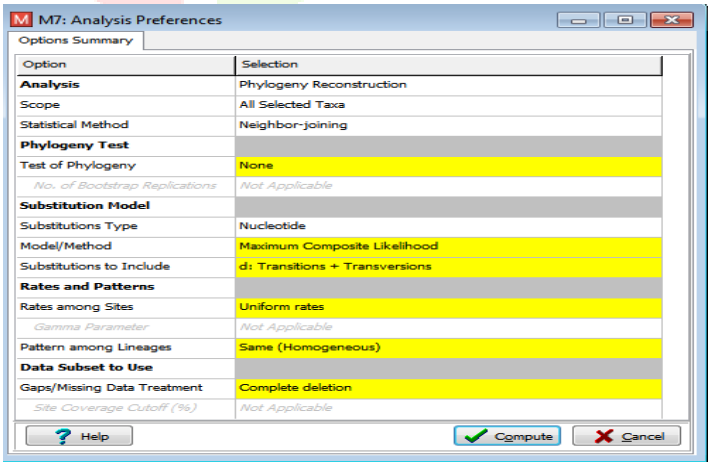


Fig3.1-Screenshotforcomputationofphylogenetictree

2.5 Identification of CpIs

1. TheGATA-1andGATA-2DNAsequencesofdifferentspeciesforthepredictionofCpIswereobtainedfromtheNCBI databaseandstoredinafile(http://www.ncbi.nlm.nih.gov/)
2. Thenucleotidesequenceforeachspecieswasgivenasinputtothe MethPrimertoolandthetoolgeneratedthenumberofCpIs present in theinputsequence.

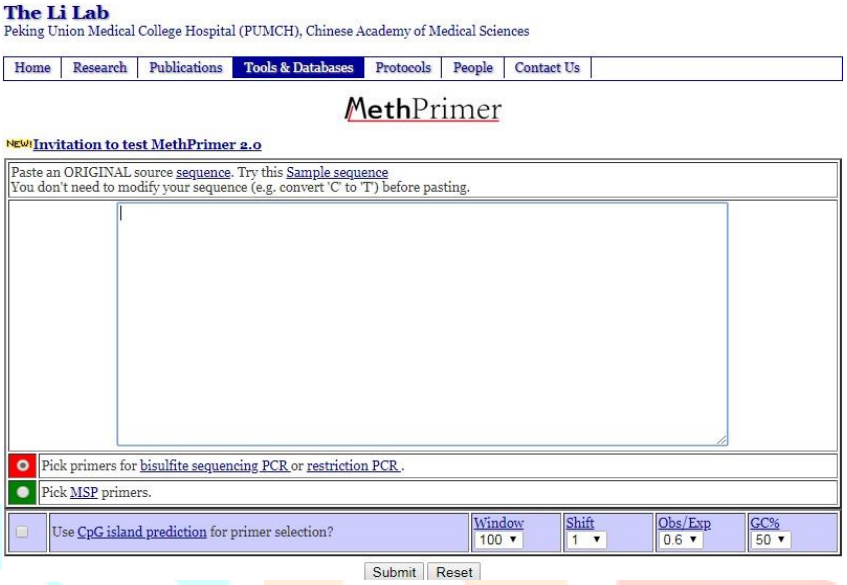


Fig1-ScreenshotofthehomepageofMethPrimerwheretheinputsequenceispastedtofindouttheresult

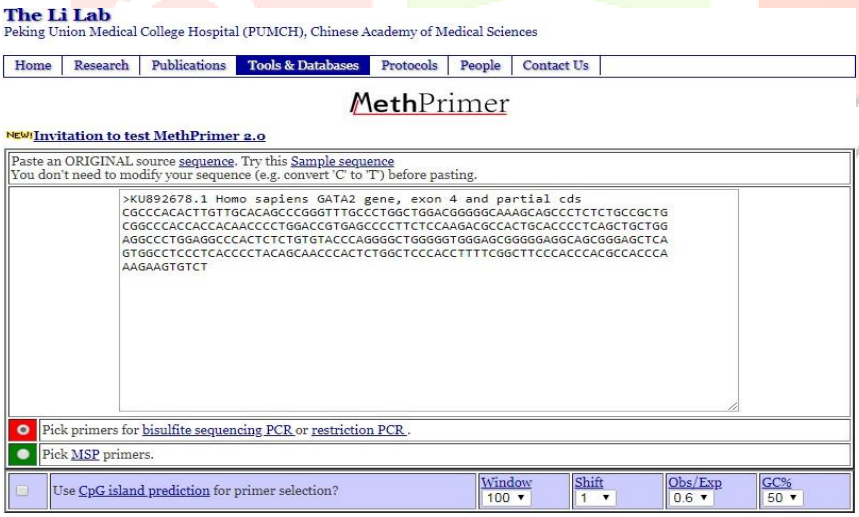
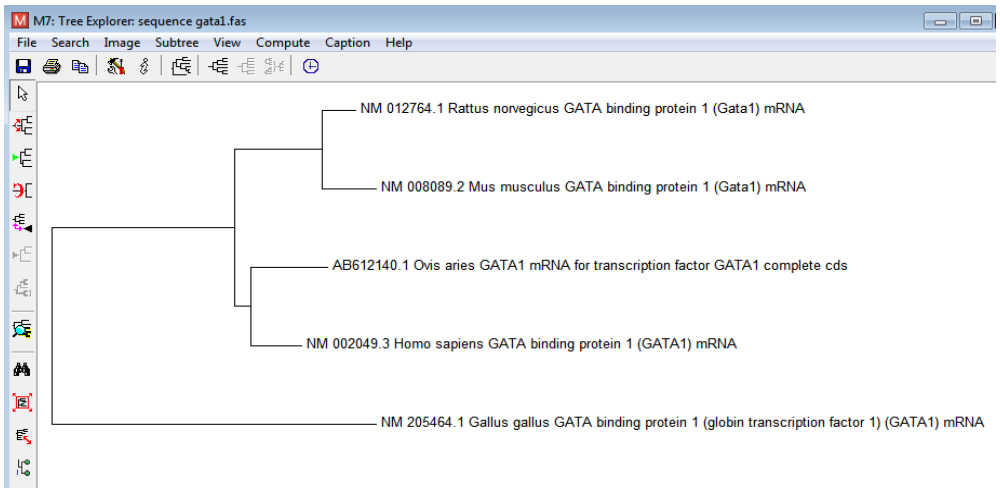


Fig2-Screenshot showingasequencepastedintheMethPrimerdialogboxasinput

3. Results

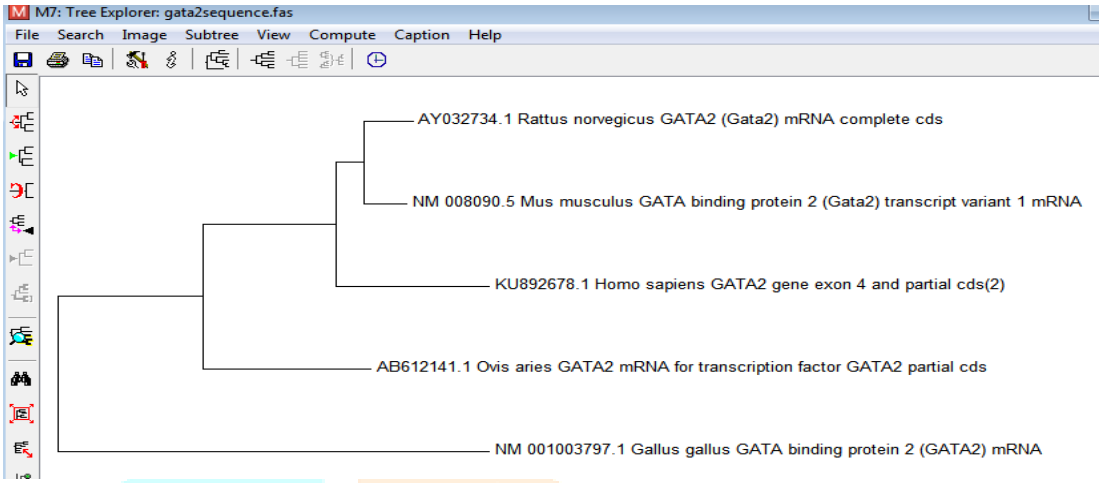
3.1 Construction of Phylogenetic tree

Neighbor joining Tree Construction: A phylogenetic tree of five different species was obtained using MEGA 6.0. The sequences were aligned as explained before and a neighbor joining tree was computed based on the alignment.



FigPhylogeneticTreeof5specieswithGATA-1

PhylogeneticanalysisofGATA-1suggeststhattheminimumdistancesexistamong*Ovisaries*and*Homosapien*;*Rattusnorvegicus*and*Musmusculus*; i.e.thesewerethecloselyrelatedspeciesandremainingspecieshadlessnumberofconserved



sequences.

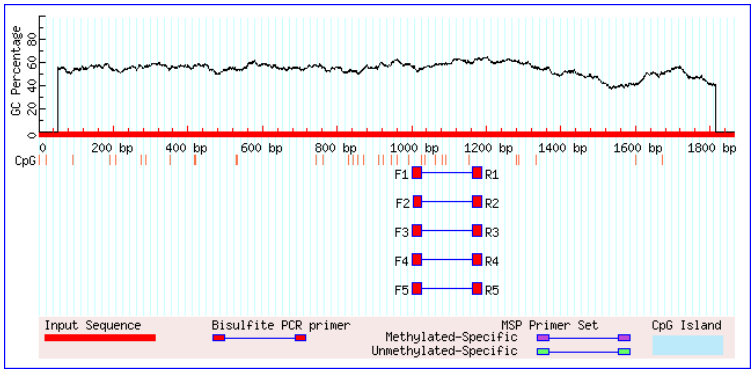
FigPhylogeneticTreeof5specieswithGATA-2

PhylogeneticanalysisofGATA-2suggeststhattheminimumdistancesexistamong*Rattusnorvegicus*and*Musmusculus*;*Homosapiens*and*Ovisaries*;i.e.thesewerethecloselyrelatedspeciesandremainingspecieshadlessnumberofconservedsequences.

3.2 IdentificationofCpIs

AfterinsertingthenucleotidesequencesofdifferentspeciesintheMethPrimerdialogboxthefollowingresultswereobtainedforGATA-1andGATA-2:

3.3 Resultoffivedifferentspecieswith GATA-1

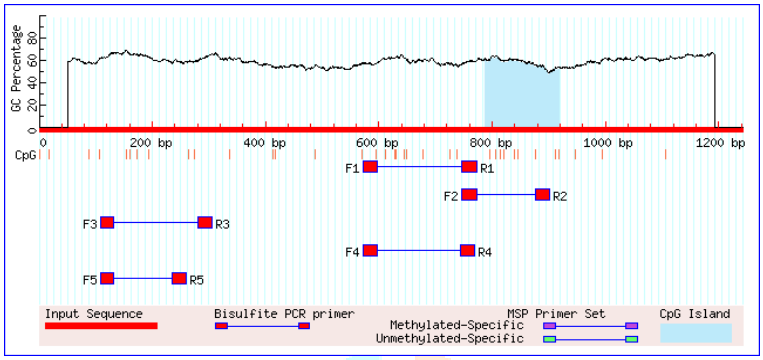


SequenceName:NM_012764.1Rattusnorvegicus

SequenceLength:1863

CpGislandpredictionresultsNoCp

Gislandswerefound

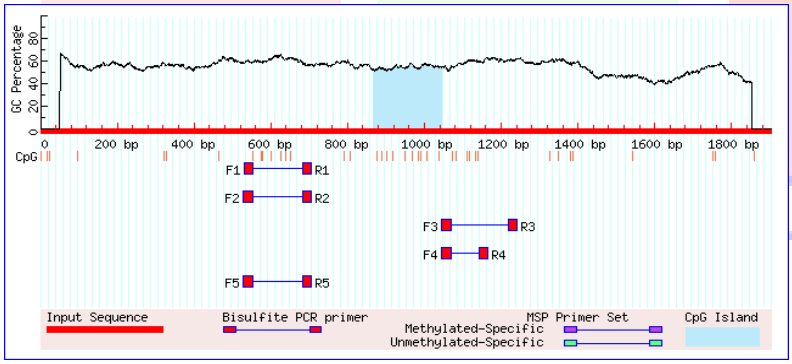


SequenceName:AB612140.1Ovisaries

SequenceLength:1242

CpG island prediction results

1 CpG islands were found

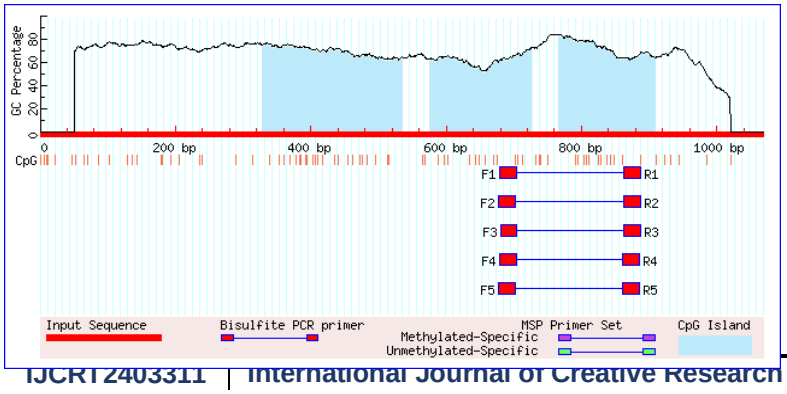


SequenceName:NM_008089.2Musmusculus

SequenceLength:1902

CpGislandpredictionresults

1CpGisland(s)werefound

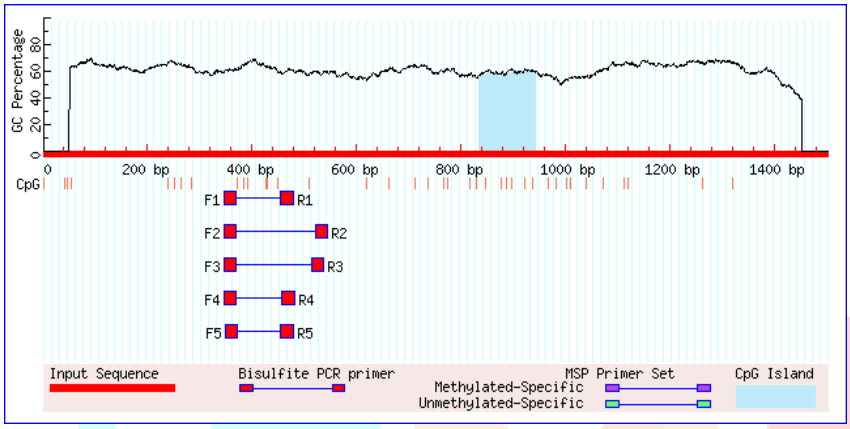


SequenceName:NM_205464.1Gallusgallus

SequenceLength:1068

CpG island prediction results

3 CpG islands were found



SequenceName:NM_002049.3 Homo sapiens

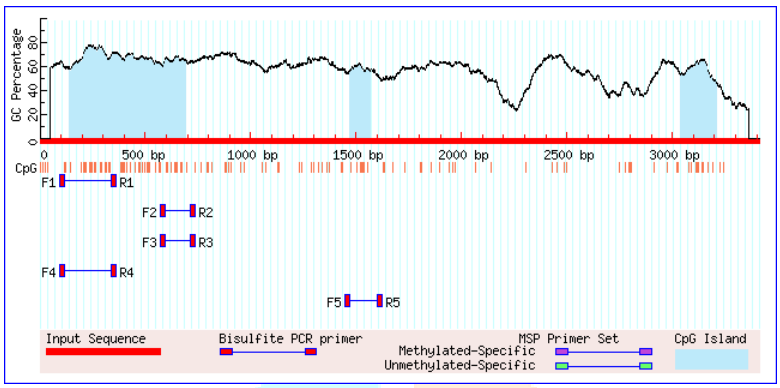
SequenceLength:1501

CpG island prediction results

1 CpG islands were found

IntheabovefiguresarangeofdifferentCpGIshavebeenfoundinDNAsequences of selected GATA-1 gene. It has been observed that *Gallus gallus*containshighernumberofCpIsamongtheotherfourspecies.

3.4 ResultoffivedifferentspecieswithGATA-2

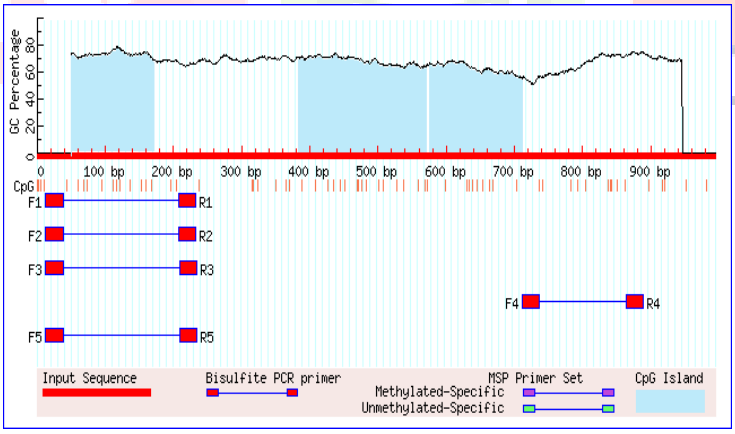


SequenceName:AY032734.1Rattusnorvegicus

SequenceLength:3411

CpG island prediction results

3 CpG islands were found

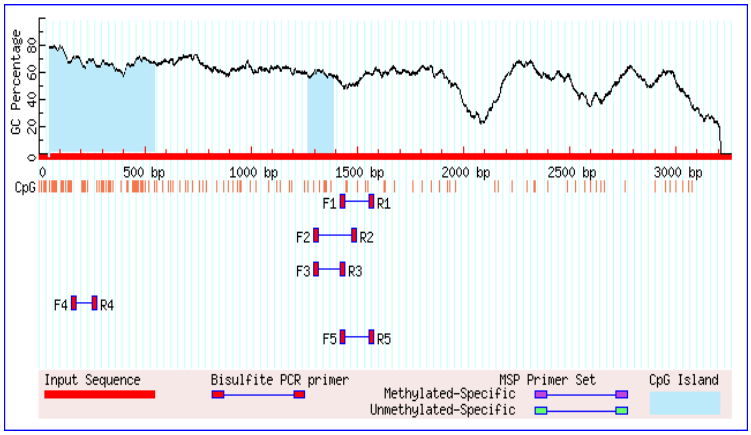


SequenceName:AB612141.1Ovisaries

SequenceLength:994

CpG island prediction results

3 CpG islands were found

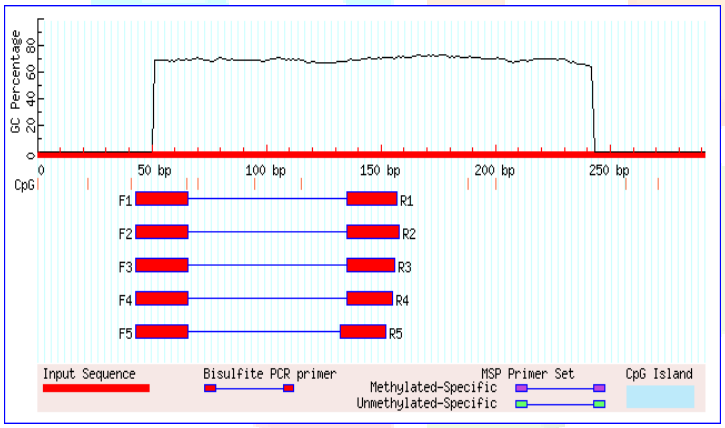


SequenceName:NM_008090.5Musmusculus

SequenceLength:3258

CpG island prediction results

2 CpG islands were found



SequenceName:KU892678.1Homosapiens

SequenceLength:291

CpG island prediction results

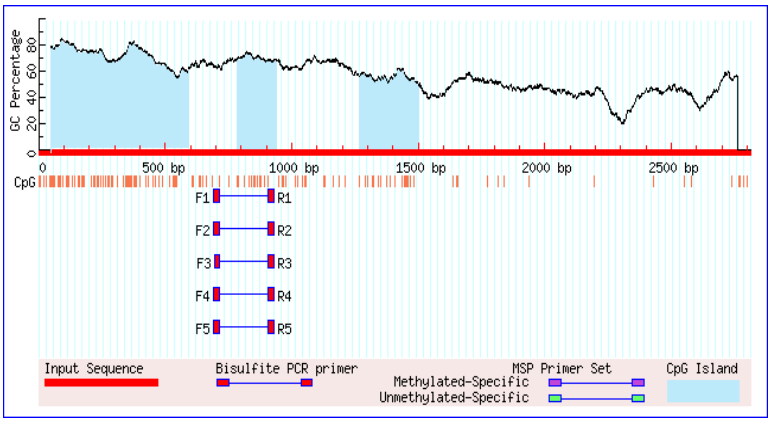
No CpG islands were found

SequenceName:KU892678.1Homosapiens

SequenceLength:291

CpG island prediction results

No CpG islands were found



SequenceName:NM_001003797.1 Gallus gallus

SequenceLength:2812

CpG island prediction results

3 CpG islands were found.

In the above figures a range of different CpIs have been found in DNA sequences of selected GATA-2 gene. It has been observed that *Gallus gallus, Ovisaries* and *Ratusnorvegicus* contains equal number of CpIs among the other two species.

In phylogenetic analysis we investigated the distances existing between different species in GATA-1 and GATA-2 gene to estimate conserved domains, which may be helpful in predicting the region of methylation in DNA sequences by using CpIs in all mentioned species.

4. Conclusions

Present study depicted that interpretation of two transcription factors (GATA-1 and GATA-2) for maintaining stemness in stem cell population by exploiting new set of translational research i.e. Stemformatics. Knowledge of stem cell has given strength to design in-silico primers for methylation involved in stem cell reprogramming. Phylogenetic analysis helped in the identification of distinctly related species whereas the interpretation of CpGIs helped in the interpretation of methylation sites. In this study, we have analyzed these factors in different species to understand the function of these genes. The Phylogenetic analysis helped to identify the conserved domains in different species with GATA-1 and GATA-2 gene which further helped in the study of Epigenetics by predicting regions of methylated sites present in nucleotide sequences by using CpGIs. Combination of different disciplines in Epigenetics studies may help us to better predict the gene regulatory pathways for the fate of future stem cell transplantation and Nobel Drug designing. In this study, DNA sequences of a particular GATA-1 gene have been revealed to have a variety of distinct CpGIs. Compared to the other four species, Gallus gallus has been shown to have a larger amount of CpGIs. It demonstrates that a variety of various CpGIs have been discovered in the DNA sequences of a particular GATA-2 gene. Rattusnorvegicus, Ovisaries, and Gallus gallus have been shown to have a similar quantity of CpGIs as the other two species which is helpful for future translational application of CRISPR CAS 9 gene is highlighted for the research and eradication of certain disease like diabetes, cancer and HIV and future cure for it. Explant tissue culture

and Stemformatics would play essential role and fascinating reservoir for future clinical Regenerative Medicine. Epigenetics plays a vital role in the field of genetic modification, molecular changes, medicine, evolution, cancer and developmental abnormalities, etc. It is also helpful in understanding the role of transcription factors (such as GATA-1 and GATA-2 which have been studied in this project). Stemformatics also allows us to learn about gene neighborhood, comparative marker selection, hierarchical cluster, association of a gene set etc., in a fast and easy way. The combination of stem cell biology and Stemformatics opened few new avenues of modern translational medicines like Regenerative Medicine and customized drug therapy. Epigenetics deals with chromatin remodeling (activating or silencing the gene). Epigenetic factors mediate stem cells self-renewal and differentiation including DNA modification such as methylation (at CpG sites). Epigenetics plays a vital role in the field of genetic modification, molecular changes, medicine, evolution, cancer and developmental abnormalities, etc.CpGlsplay a vital role in methylation and it leads to the gene silencing as well as cell reprogramming during embryogenesis. Our recent observation validly concludes that stemformatics tool to explore omics scenario of respective genes. In addition to it could be a vital platform to edit embryo and utilization of CRISPR CAS9 gene.

References:

1. Antequera, F., Bird, A. (1993). CpG Islands. In: Jost, JP.,Saluz, HP. (eds) DNA Methylation. EXS, vol 64. Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-9118-9_8
2. Bird AP. CpG-rich islands and the function of DNA methylation. *Nature*. 1986;321:209–213. doi: 10.1038/321209a0.
3. Doré, L. C., &Crispino, J. D. (2011). Transcription factor networks in erythroid cell and megakaryocyte development. *Blood, The Journal of the American Society of Hematology*, *118*(2), 231-239.
4. Feil R, Berger F (2007). "Convergent evolution of genomic imprinting in plants and mammals". Trends Genet. 23 (4): 192–199. doi:10.1016/j.tig.2007.02.004. PMID 17316885.
5. Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D., &Marraffini, L. A. (2015). Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature*, *519*(7542), 199-202.
6. Ishino, Y., Krupovic, M., &Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *Journal of bacteriology*, *200*(7), e00580-17. https://doi.org/10.1128/JB.00580-17
7. Jarny Choi and others, Stemformatics: visualize and download curated stem cell data, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D841–D846, https://doi.org/10.1093/nar/gky1064
8. Karin, M. (1990). Too many transcription factors: positive and negative interactions. *The new biologist*, *2*(2), 126-131.
9. Khan, M. H. U., Khan, S. U., Muhammad, A., Hu, L., Yang, Y., & Fan, C. (2018). Induced mutation and epigenetics modification in plants for crop improvement by targeting CRISPR/Cas9 technology. Journal of Cellular Physiology, 233(6), 4578-4594.
10. Rajab, N., Angel, P. W., Deng, Y., Gu, J., Jameson, V., Kurowska-Stolarska, M., ...& Wells, C. A. (2021). An integrated analysis of human myeloid cells identifies gaps in in vitro models of in vivo biology. *Stem cell reports*, *16*(6), 1629-1643.
11. Rivera, C. P., Mosbergen, R., Korn, O., Chen, T., Nagpal, I., & Wells, C. A. (2017). Ontology Challenges for the Stem Cell Community: Towards Integrative Data Mining in the Stemformatics Atlas. In *ICBO*.

12. Saxonov S, Berg P, Brutlag DL (2006). "A genome-wide analysis of CpGdinucleotides in the human genome distinguishes two distinct classes of promoters". ProcNatlAcadSci USA. 103 (5): 1412–1417. doi:10.1073/pnas.0510310103. PMC 1345710 Freely accessible. PMID 16432200.

13. Takai D, Jones PA (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22". ProcNatlAcadSci USA. 99 (6): 3740–5. doi:10.1073/pnas.052410099. PMC 122594 Freely accessible. PMID 11891299.
14. Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in genome editing and beyond. *Annual review of biochemistry*, *85*, 227-264.
15. Zhang, Y., MA, X., & XIE, X. (2022). *CRISPR*. Springer Singapore.
16. Zhang, Z., Zhang, Y., Gao, F., Han, S., Cheah, K. S., Tse, H. F., &Lian, Q. (2017). CRISPR/Cas9 genome-editing system in human stem cells: current status and future prospects. *Molecular Therapy-Nucleic Acids*, *9*, 230-241.