Title: Homology Modelling and *in-silico* analysis of 39bp Insertion-Deletion in Sahiwal Cattle SERPINA14 gene: The first report

Running title: Homology Modelling and *in-silico* analysis

P. B. Nandhini, D. Ravikumar, Oshin, M. R. Vineeth and Anupama Mukherjee

ICAR-National Dairy Research Institute, Karnal.

Corresponding author: Dr. Anupama Mukherjee, Principal Scientist, Animal Genetics and Breeding Division, ICAR-National Dairy Research Institute, Karnal 132001, Haryana, India. writetoanupama@gmail.com

Acknowledgement

The author wishes to thank ICAR- National Dairy Research Institute and ICAR-National Bureau of Animal Genetic Resources for their funds and laboratory facilities to carry out this work.

Abstract

SERPINA14 proteins are progesterone induced and are secreted during pregnancy in large quantities by the endometrial epithelium. Serine proteinase inhibitor being represented only in a limited group of mammals has been associated with higher embryo survival rates, productive life, milk production and health traits and a minisatellite insertion has been reported in Bali cattle. The most variable exons (1&4) of SERPINA14 gene in Sahiwal cattle were sequenced to reveal the 39bp repeats in the coding region of the exon 4. In order to ascertain the changes in this gene that directly affects the protein structure, its structure was deduced using homology modelling with *Bos taurus* as reference, after imputing the missing coding sequence. The comparison of protein structure using SWISS-MODEL, I-TASSER and PHYRE2 showed that PHYRE2 predicted the best model for the proteinswith more than 90% of the residues lying in the most favoured regions in the Ramachandran plot.The impact of the indel with 5 repeats was assessed to be deleterious using PROVEANwith a score of -22.464 while indel with 4 repeats had a score of -10.676 against athreshold of -2.5 comparing with 130 sequences and 30 clusters.However, the association of the indel with reproduction data failed to reveal any significant effect which could be attributed to the data size. Phylogenetic study of the gene with its relatives showed that the sequence with 5 repeats was similar to Yak and Bison while the one with 4 repeats resembled all bovines alike.

Key words: Homology modelling, SERPINA14, INDEL, Ramachandran plot, Sahiwal

Introduction

The studies fixated on genetic evaluation are still restricted to the identification of single nucleotide polymorphisms and its association with the traits involved, either candidate genes or GWAS. The impact of the resultant outcome of this polymorphism on protein structure and function has been minimally explored. Protein structure modelling could be the tool to bridge the huge gap between the protein sequences available and its structures the reason being genome projects producing sequences at a much higher rate than NMR and X-ray laboratories can solve the three-dimensional structures[1,2]. The three-dimensional structure of protein can help in determining the function of protein better than the sequence itself, the reason being that the conservation of structure is far higher than the sequence in the same family[4]. The difficulties start with retrieval of the required amount of protein needed for structural analysis followed by the optimum crystallization. Protein modelling has gained further importance due to the advent of structural genomics which emphasizes the relationship between the evolutionarily related proteins in their structural and functional similarity[2,4]. No ventures have been attempted to model SERPINA14 gene in cattle, which is responsible for the immunomodulation of the uterine environment during the implantation of the foetus playing a major role in early embryonic growth. The protein structure may reveal the existence of motifs that can throw some light over the hazed information on the function of the protein. The study can lead to rationalization of the studies conducted[5] that resulted in significant association of single nucleotide polymorphisms with higher embryo survival rates in cattle, but the overall fertility of buffaloes remained consistent despite showing 8 single nucleotide polymorphisms [6,7,8,9].

Indels less than 40 bp are easily identified by sequencing and have been in the dark and have not received as much attention as the SNPs[10]. Insertions/deletions can also be identified by PCR fragment size analysis[11]. A common source of structural variation, indels in protein super families are the indels, occurring commonly in loops and turns, since indels in these positions are less likely to disrupt folding than in the core of the protein. Insertions often provide novel structural elements that contribute to catalysis, substrate binding, or protein–protein interactions and confer novel characteristics to a diverging family. Indels have the potential to be used as an important genetic marker owing to the huge amount of sequenced data generated from non-model organisms, for the study of natural population[12]. Even in-frame mutations have been reported to have antagonistic roles[13,14]. Novel genetic marker systems have been developed for parentage testing, molecular traceability, breed certification

and identity test for which indels, SNPs and Microsatellite markers within gene that is responsible for a specific genotype specific to a breed can be used[15]. The most studied indels in cattle were the ones in PRNP, IGF2, CAPN1, ADD1/SREBP1c, SMAD3, Pax 7, Visfatin genes. Prior reports on Buffalo and Bali cattle have revealed a 13 amino acid residues insertion in SERPINA14 gene, MNAKEVPVVVKVP and VPMKAKEVPAVVK respectively[16,9] while it remains unexplored in Indian Sahiwal cattle.

Hence, the intension of this study was to anticipate the most credible structure of SERPINA14 gene of Sahiwal, its stability and its probable effect in the biological system and the association of the indel with the reproduction traits like age at first calving, age at first service, service period, calving interval and calving to first service.

Materials and methods

Sample collection

Blood samples were collected from 70 animals (35 low yielders and 35 high yielders based on herd average) in the Sahiwal herd maintained at Livestock Research Centre, ICAR-National Dairy Research Institute, abiding by the rules laid down by the Institutional Animal Ethics Committee and approved in 43rd meeting held on 13.10.2018 (43-IAEC-18-8) held at ICAR-NDRI, Karnal. The samples were used for isolation of DNA using the Wizard Genomic DNA Purification Kit (cat no # A1620, Promega, USA) as per the manufacturer's instructions. The DNA was subjected to quantity (Nanodrop) and quality (Agarose gel electrophoresis) check before further processing.

PCR amplification

The highly variable coding regions of SERPINA14 gene was amplified using two primers designed using primer blast of NCBI; primer 1 (5'-GATTGCCGCAGAAATGTCCC-3', 3'-CACATGGTGGCTGATGGTCT-5') targeting exon 1 and primer 2 (5'-CTGCCTCTCGATCTTGCCAT-3', 3'- CCACTCCATTCCCAGACCAC-5') targeting exon 4, were used to amplify 329 and 514 bp long regions, at 55℃ and 60℃ melting temperatures respectively. The PCR was optimized for 30µL reactions with 15µL of GoTaq® DNA Polymerase (2X), 11µL of Nuclease free water, 1µL each of the primers and 2µL of DNA containing 100ng of DNA/µL of DNA. The amplified sequences were aligned with the reference gene with ID 286871 of *Bos taurus*, trimmed of the intronic regions, translated using ExPASy Bioinformatics resource portal, and the longest open reading frame

(5’3’Frame) was imputed into the protein sequence obtained from NCBI (Protein Id=NP_777222.1).

Structure modelling

Primary, secondary and tertiary structure of the protein was modelled using three different interfaces. The input sequence was used as the supported input in SWISS-MODEL[17,18,19]. The five templates that were the top hits in SWISS-MODEL template search was used to predict the protein structure. Global Model Quality Estimate, QMEAN[20] (Z score) and Ramachandran plot scores have been used to evaluate the structures predicted[21].

I-TASSER is a fold recognition/threading tool for modelling proteins[22,23] with less than 30% sequence identity[4,24,25]. Each of the 5 models generated by I-TASSER is graded using the C-score, TM-score and RMSD[26].

Phyre2 (**P**rotein **H**omology/analog**Y** **R**ecognition **E**ngine V 2.0) is again a threading tool that aims to improve the structure prediction of proteins along with function and mutations. Phyre2 uses remote homology modelling methods to predict the three-dimensional structure of the target protein[27].

The impact of indel on the biological function of the protein, related to the change in structure was deduced using PROVEAN protein (**Pro**tein **V**ariation **E**ffect **An**alyzer)[28]. Phylogeny was constructed for the sequenced samples of Exon 4 of SERPINA14 with Mr.Bayes software[29,30]. Burnin parameter was kept at 10,000 and excluding first 250 trees.

Phenotype

The age at first calving, age at first service, first service period, first calving interval and first calving to first service were recorded. The association of the indel with the phenotypic reproductive traits has been carried out using least square means technique and their significance was tested. The traits were corrected for environmental factors by Mixed Model Least-Squares and Maximum Likelihood Computer Program (LSMLMW)[31]. Age at first service and age at first Calving were corrected using the model, $Y_{ijm}= \mu + SB_i + PB_j + e_{ijm}$, where,

Y_{ijm} = Observation on m^{th} animal that was bornin i^{th} season and j^{th} period.

μ = Overall mean

SB_i = Effect of k^{th} season of birth

PB_j = Effect of l^{th} period of birth

e_{ijm} = Random error NID (0, $\sigma^2 e$)

and the other reproduction traits were corrected using the model, $Y_{ijkm} = \mu + S_i + P_j + PA_k + e_{ijkm}$,

where,

Y_{ijkm} = Observation on m^{th} animal that calved in i^{th}season and j^{th} period in k^{th} parity

μ = Overall mean

S_i = Effect of i^{th}season of calving

P_j = Effect of j^{th} period of calving

PA_k = Effect of k^{th}parity

e_{ijkm} = Random error NID (0, $\sigma^2 e$)

To estimate association of INDEL with reproduction traits, a regression analysis was done on adjusted records using the model, $Y_{ij} = a + b_i INDEL_i + e_{ij}$,

where,

Y_{ij} = Adjusted observation on j^{th} animal having i^{th}INDEL

A = Intercept

b_i = Partial regression coefficient for the $INDEL_i$

$INDEL_i$ = Effect of i^{th} INDEL as independent variable

e_{ij} = Random error NID (0, $\sigma^2 e$)

Results and discussion

A complete sequence alignment using the reference sequence, *Bos taurus,* resulted in the discovery of 13 SNPs (2 SNPs in exon1 and 11 in exon 4) and (39bp)n repeats (exon 4). The PCR products visualized using gel documentation showed that there were 4 types of sequenced products based on their movement in agarose gel electrophoresis. They were confirmed to be of 4 different sizes in terms of their sequences by alignment using MEGA7[32] software. These 4 types of sequences were categorized into 2 classes namely, homozygous individuals showing single band and heterozygotes showing double bands. The single bands were found to be 514bp (3 Repeats) and 592bp (5 Repeats) in length while the double bands had a longer fragment of 553bp and a shorter 514bp (4 Repeats) but the band with smaller size could not be concluded in both the cases due to the fact that the insertion caused the chromatograms to be overlapped making it impossible to read them (Fig 1). Analysis using T-REKS algorithm using the translated amino acid sequence of exon 4, which aims at *de novo* detection and alignment of repeats based on K-means algorithm[33], showed that there were in total 3 kinds of repeats; (39bp)5, (39bp)4 and (39bp)3. The sequences with 3 repeats

were found in homozygous and heterozygous condition. One representative sequence with different number of repeats was carefully chosen to model the probable types of protein produced by the gene.

The sequence alignment of sequences showing 3, 4 and 5 repeats denoted as 3R, 4R and 5R respectively using CLUSTALW[34].

```
4R   GCCCTACATCAAGCTGAGATAGAGCTGAGCGAGCACGCCTTAACCGTGGACACAGCCATT  60
5R   GCCCTACATCAAGCTGAGATAGAGCTGAGCGAGCACGCCTTAACCGTGGACACAGCCATT  60
3R   GCCCTACATCAAGCTGAGATAGAGCTGAGCGAGCACGCCTTAACCGTGGACACAGCCATT  60
Ref  GCCCTACATCAAGCTGAGATAGAGCTGAGCGAGCACGCCTTAACCGTGGACACAGCCATT  60
     ************************************************************

4R   CACACAGATAATCTGTTGAAAGTCCCAGTGAAGGCAAAGGAGGTCCCGGCGGTCGTGAAA  120
5R   CACACAGATAATCTGTTGAAAGTCCCAGTGAAGGCAAAGGAGGTCCCGGCGGTCGTGAAA  120
3R   CACACAGATAATCTGTTGAAAGTCCCAATGAAGGCAAA----------------------  98
Ref  CACACAGATAATCTGTTGAAAGTCCCAATGAAGGCAAA----------------------  98
     *************************** **********

4R   GTCCCAATGAAGGCAAAGGAGGTCCCGGCGGTCATGAAAGTCCCAATGAACACAAAGGAG  180
5R   GTCCCAGTGAAGGCAAAGGAGGTCCCGGCGGTCGTGAAAGTCCCAATGAACACAAAGGAG  180
3R   ------------------------------------------------------------GGAG  102
Ref  ------------------------------------------------------------GGAG  102
                                                                 ****

4R   G---------------------------------------TCCCGGTGGTCGTGAAAGTC  201
5R   GTCCCGGTGGTCGTGAAAGTCCCAATGAAGGCAAAGGAGGTCCCGGTGGTCGTGAAAGTC  240
3R   GTCCCGGCGGTCGTGAAAGTCCCAATGAACACAAAGGAGGTCCCAGTGGTTGTGAAAGTC  162
Ref  GTCCCGGCGGTCGTGAAAGTCCCAATGAACACAAAGGAGGTCCCAGTGGTTGTGAAAGTC  162
     *                                      **** ***** *********

4R   CCAATGAACGCAAAGGAGGTCCCGGTGGTCGTGAAGGTCAACATACCCTTCTTGCTGTTT  261
5R   CCAATGAACACAAAGGAGGTCCCGGTGGTCGTGAAGGTCAACAGACCCTTCTTGCTGTTT  300
3R   CCAATGAACACAAAGGAGGTCCCGGTGGTCGTGAAGGTCAACAGACCCTTCTTGCTGTTT  222
Ref  CCAATGAACACAAAGGAGGTCCCGGTGGTCGTGAAGGTCAACAGACCCTTCTTGCTGTTT  222
     ********* ********************************* ****************

4R   GTGGAGGATGAGAAGACTCAAAGACACCTCTTTGTGGGCAAAGTCCTCAACCCCACAGTT  321
5R   GTGGAGGATGAGAAGACTCAAAGAGACCTCTTTGTGGGCAAAGTCCTCAACCCCCAAGTT  360
3R   GTGGAGGATGAGAAGACTCAAAGAGACCTCTTTGTGGGCAAAGTCCTCAACCCCCAAGTT  282
Ref  GTGGAGGATGAGAAGACTCAAAGAGACCTCTTTGTGGGCAAAGTCCTCAACCCCCAAGTT  282
     ************************ ****************************  ****

4R   GAATAC  327
5R   GAGTAG  366
3R   GAGTAG  288
Ref  GAGTAG  288
     ** **
```

This result is similar to the indel reported in Bali cattle which caused a minisatellite VPMKAKEVPAVVK insertion which was the result of 114bp insertion[35].

The repeats found in translated sequence of 5R using T-REK (identification of Tandem REpeats in sequences with a K-meanS based algorithm)

KVPVKAKEVPAVV---

KVPVKAKEVPAVV---
KVPMNTKEVPVVV---
KVPMKAKEVPVVV---
KVPMNTKEVPVVV---. This shows five 13amino acid repeats in 5R which is in acceptance with the five 39bp repeats in the DNA sequence.

Homology modelling of the three sequences using SWISS-MODEL with Conserpin in the latent state (SMTL ID: 5cdz.1.A) (Sequence similarity-36%; coverage-76%)was carried out and the quality parameters were analysed(Table1).

GMQE reflects the accuracy of a model built with the given template, alignment and the coverage and values close to 1 indicates the best model. The QMEAN score is also accounted for while judging the model to increase the reliability of estimating the quality of the estimation. The three models using SWISS-MODEL have GMQE less than 0.6 and this must be attributed to the sequence similarity which is less than 40%[36,37]. QMEAN is a composite estimator based on different geometrical properties and provides both global (i.e. for the entire structure) and local (i.e. per residue) absolute quality estimates based on one single model[20]. The QMEAN Z-score provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. It indicates whether the QMEAN score of the model is comparable to what one would expect from experimental structures of similar size. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size. Scores of -4.0 or below is an indication of models with low quality. The models built have Z-score less than -4 and hence fold recognition which uses the structure of the known protein structure for modelling is used since it is more effective in many cases where the sequence similarity is less than 25-30%. Moreover, the Ramachandran plot values have not crossed the optimum of 90% of residues lying in the most favoured region.

I-TASSER predicts protein structure and function using the sequence-to-structure-function rule[22,23]. Five models were built for each sequence using the top ten templates from LOMETS threading program. Confidence score (C-score) is the benchmark of accuracy used to score the model. Large numbers of decoys, structural conformations, are simulated and the SPICKER program selects the final model based on the resulting cluster of decoys based on pair-wise similarity of structure. C-score measures the confidence of each model quantitatively by calculations of threading template alignments significance and structure assembly simulations convergence parameters. The range of C-score is [-5, 2][26]; higher C-score of signifies a model with a higher confidence and vice-versa. C-score is the basis of

estimation of TM-score and RMSD and this is justified by observing the correlation observed between these qualities. The models with higher C-scores usually have better TM-score and RMSD.A TM-score >0.5 indicates a model of correct topology and a TM-score<0.17 means a random similarity[26]. The models built had C-scores in the optimum range and a C-score of >-1.5 indicates model of correct global topology. TM score of >0.5 indicates that the model is valid. The Ramachandran plot was generated using SAVES v5.0 and the percentage of the residues in the favoured region lies below 80% (Table 2) for all the models even after refinement using ModRefiner[38] and this led to the use of PHYRE2 for further probing.

The best model identified by Phyre2 was modelled based on heparin cofactor ii-s195a thrombin complex as the template; 388 residues (82% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template. Our target is a member of the SERPIN family and hence the model that was modelled with SERPIN family protein as a template has been chosen to be studied, of which 367 were aligned, though the template was ranked 5th in the list of suggested 20 templates with 100.0% coverage and 31% sequence identity. The Ramachandran plot values after modification shows that these are by far the best models developed with more than 90% of the residues lying in the most favoured region. Hence the further analysis was carried out using the structures predicted using PHYRE2 (Fig: 3) considering them to be the most appropriate structure. The structures were viewed using PyMOL[39].

The PROVEAN protein prediction for the variant P393_M394insVKAKEVPAVVKVPMNTKEVPVVVKVP (five 13 amino acid repeats for 5R) showed a score of -22.464 and the variant P393_M394insMKAKEVPAVMKVP (four 13 amino acid repeats for 4R) showed a score of -10.676 compared to the reference sequence or the wild type with three 13 amino acid repeats (3R). The default threshold is -2.5 and variants with a score equal to or below -2.5 are considered deleterious while variants with a score above -2.5 are considered neutral[28]. The threshold indicates that the repeats, both 4 and 5, are predicted to be deleterious with respect to the trait that they govern.

Association of reproduction traits

The phenotypic observations of the reproductive traits were associated to the indel variants. The mean of AFC agrees with the results[40] showing 322.2 ± 6.82 days, while the AFS has not yet been studied in Sahiwal cattle. The CI and SP agreed with the studies[41] showing 494.45 ± 5.05 days and 223.00 ± 6.12 days respectively, while CS has not yet been studied. In this

study, CI and SP were significantly affected by parity and period of calving at $p<0.05$ level with animals in 4th parity with minimum CI and SP of 357.96 ± 56.96 days and 88.64 ± 44.35 days respectively, while animals born before 2010 had the least CI and SP of 394.71 ± 64.27 days and 103.34 ± 56.87 days respectively. The results match with the studies[41,42] reported the significant effect of period of calving on SP while one[43] reported only season of calving to be significant and other[44] reported period and season of calving to significantly influence SP. The CI was found to be significantly affected by period[41,44]. CI was influenced by period of calving[41] while other reports show that season and period of birth influenced CI[43,44].

The indel did not significantly influence any of the trait considered, since it failed to show association with them which could be attributed to the sample size of the study.

The phylogenetic analysis of the three sequences namely 3R, 4R and 5Rwas carried out with *Odocoileus virginianus* (White-tailed deer), *Bubalusbubalis* (Water Buffalo), *Bos indicus* x *Bos Taurus* (Crossbred cattle), *Bos indicus* (Zebu cattle), *Bos taurus* (Exotic cattle), *Bos mutus* (Wild Yak), *Bos bison* (Bison), *Ovisaries* (Sheep), *Capra hircus* (Goat) and *Tursiops truncatus* (Common bottlenose dolphin) using Mr. Bayes software with 10,000 generations taking one tree every 10 generations with a burnin of 25%. The results were interesting. 5R was related to Yak and Bison, while 3Rwas related to indicine cattle and 4R was related to all bovines alike (Fig 4).

Conclusion

The genetic variants of SERPINA14 gene was discovered and the structure of the probable protein produced by the variants caused by the INDEL was predicted for the first time in Sahiwal cattle. Association of the indel with the reproductive traits showed no significance and hence further studies are recommended in a diverse population. The study of INDELS is gaining importance while it is still lacking behind in Indigenous cattle populations in India for which this study will be a forerunner.

References

1. Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G. Homology modeling, model and software evaluation: three related resources. Bioinformatics. 1998;14(6):523-528. doi:10.1093/bioinformatics/14.6.523
2. Xiang Z. Advances in homology protein structure modeling. Curr Protein Pept Sci. 2006;7(3):217-227. doi:10.2174/138920306777452312

3. Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 1993;2(11):1811-1826. doi:10.1002/pro.5560021104

4. Goldsmith-Fischman S, Honig B. Structural genomics: Computational methods for structure analysis. *Protein Sci*. 2003;12(9):1813-1821.

5. Khatib H, Schutzkus V, Chang YM, Rosa GJM. Pattern of expression of the uterine milk protein gene and its association with productive life in dairy cattle. J Dairy Sci. 2007;90(5):2427-2433. doi:10.3168/jds.2006-722

6. Khatib H, Huang W, Wang X, et al. Single gene and gene interaction effects on fertilization and embryonic survival rates in cattle. J Dairy Sci. 2009a;92(5):2238-2247. doi:10.3168/jds.2008-1767

7. Khatib H, Maltecca C, Monson RL, Schutzkus V, Rutledge JJ. Monoallelic maternal expression of STAT5A affects embryonic survival in cattle. BMC Genet. 2009b;10:13. doi:10.1186/1471-2156-10-13

8. Khatib H, Monson RL, Huang W, et al. Short communication: Validation of in vitro fertility genes in a Holstein bull population. Journal of Dairy Science. 2010;93(5):2244-2249. doi:10.3168/jds.2009-2805

9. Jerome A, Pandey AK, Sarkar SK. Homology modeling of single nuleotide polymorphisms in candidate genes controlling embryonic growth of buffalo. J. Anim. Sci. 2015; 85 (6): 578-583. http://krishi.icar.gov.in/jspui/handle/123456789/8024. Accessed April 15, 2020.

10. Sanders SJ, Mason CE. The Newly Emerging View of the Genome. Genomics, Circuits, and Pathways in Clinical Neuropsychiatry. 2016:3-26. doi:10.1016/b978-0-12-800105-9.00001-9.

11. Joseph L. Setting Up a Laboratory. Genetic Diagnosis of Endocrine Disorders. 2010:303-314. doi:10.1016/B978-0-12-374430-2.00027-4

12. Väli Ü, Brandström M, Johansson M, Ellegren H. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genetics. 2008;9(1):8. doi:10.1186/1471-2156-9-8

13. Venhoranta H, Pausch H, Flisikowski K, et al. In frame exon skipping in UBE3B is associated with developmental disorders and increased mortality in cattle. BMC Genomics. 2014;15(1):890. doi:10.1186/1471-2164-15-890

14. Philipp U, Lupp B, Mömke S, et al. A MITF Mutation Associated with a Dominant White Phenotype and Bilateral Deafness in German Fleckvieh Cattle. PLOS ONE. 2011;6(12):1-6. doi:10.1371/journal.pone.0028857

15. Han SH, Cho SR, Cho IC, Cho WM, Kim SG, et al. A Parentage Test using Indel, Microsatellite Markers and Genotypes of MC1R in the Jeju Black Cattle Population.

Journal of Embryo Transfer [Internet]. 2013; 28 (3): 207–13. doi:10.12750/JET.2013.28.3.207

16. Kandasamy S, Jain A, Kumar R, Agarwal SK, Joshi P, Mitra A. Molecular characterization and expression profile of uterine serpin (SERPINA14) during different reproductive phases in water buffalo (Bubalus bubalis). Anim Reprod Sci. 2010;122(1-2):133-141. doi:10.1016/j.anireprosci.2010.08.005

17. Bienert S, Waterhouse A, de Beer TAP, et al. The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. 2017;45(D1):D313-D319. doi:10.1093/nar/gkw1132

18. Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis. 2009;30 Suppl 1:S162-173. doi:10.1002/elps.200900140

19. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296-W303. doi:10.1093/nar/gky427

20. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 27(3):343-350.

21. Hooft RW, Sander C, Vriend G. Objectively judging the quality of a protein structure from a Ramachandran plot. Comput Appl Biosci. 1997;13(4):425-430. doi:10.1093/bioinformatics/13.4.425

22. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725-738. doi:10.1038/nprot.2010.5

23. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12(1):7-8. doi:10.1038/nmeth.3213

24. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. *Nature Structural Biology*. 2001;8(6):559-566. doi:10.1038/88640

25. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. 2017;7. doi:10.1038/s41598-017-09654-8

26. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9(1):40. doi:10.1186/1471-2105-9-40

27. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845-858. doi:10.1038/nprot.2015.053

28. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745-2747. doi:10.1093/bioinformatics/btv195

29. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754-755. doi:10.1093/bioinformatics/17.8.754

30. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003;19(12):1572-1574. doi:10.1093/bioinformatics/btg180

31. Harvey WR. Mixed model least-squares and maximum likelihood computer program. PC-2 version. 1990;4255.

32. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33(7):1870-1874. doi:10.1093/molbev/msw054

33. Jorda J, Kajava AV. T-REKS. Bioinformatics. 2009;25(20):2632–2638. doi:10.1093/bioinformatics/btp482

34. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-4680.

35. Jakaria J, Saputra F, Paramitasari K, Partogi Agung P, Maskur M. Identification of uterin milk protein (utmt) gene in bali cattle using direct sequencing. Journal of the Indonesian Tropical Animal Agriculture. 2016;41. doi:10.14710/jitaa.41.1.1-6

36. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006;22(2):195-201. doi:10.1093/bioinformatics/bti770

37. Guex N, Diemand A, Peitsch MC. Protein modelling for all. *Trends Biochem Sci*. 1999;24(9):364-367. doi:10.1016/s0968-0004(99)01427-9

38. Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. Biophys J. 2011;101(10):2525-2534. doi:10.1016/j.bpj.2011.10.024

39. Schrodinger LLC.The PyMOL molecular graphics system. 2010;*Version*, *1*(5), p.0.

40. Kathiravan P, Sachdeva GK, Gandhi RS, Raja TV, Singh, Singh A. Genetic evaluation of first lactation production and reproduction traits in Sahiwal cattle. Journal of Livestock Biodiversity. 2009;1: 51-55.

41. Dubey P, Singh C. Estimates of genetic and phenotypic parameters considering first lactation and lifetime performance traits in Sahiwal and crossbred cattle. The Indian journal of animal sciences. 2005;75:1289-1294.

42. Raja T, Gandhi R. Factors influencing productive and reproductive performance of Sahiwal cattle maintained at organized farm conditions. The Indian Journal of Animal Sciences. 2015;85(6).

43. Narwaria US, Mehla RK, Verma KK, Lathwal SS, Yadav R, Verma AK. Study of short lactation in Sahiwal cattle at organized farm. Vet World. 2015;8(5):690-694. doi:10.14202/vetworld.2015.690-694.

44. Kumar A, Gandhi RS. Evaluation of pooled lactation production and reproduction traits in Sahiwal cattle. Indian J Anim Sci. 2011;81:600-604.

Table 1: The Quality estimates of the models developed using SWISS-MODEL

Estimate	**5R**	**4R**	**3R**
Sequence similarity with template	0.36	0.36	0.34
GMQE	0.55	0.44	0.48
QMEAN (Z score)	-5.77	-3.01	-8.17
Ligands	None	None	None
Oligo-State	Monomer	Monomer	Monomer
Ramachandran favored region	88.84	88.84	85.31
Ramachandran outliers	4.28	4.28	7.34

3R-(39)3 repeats; 5R-(39)5 repeats; 4R-(39)4 repeats

Table 2: The quality parameters built using I-TASSER

Model	**5R**	**4R**	**3R**
C-score	-1.67	-1.70	-1.16
TM-score	0.51+-0.15	0.51+-0.15	0.57+-0.15
RMSD (Å)	11.1+-4.6	11.1+-4.6	9.7+-4.6
Ramachandran plot values before modification			
Most favored region	63.1 (274)	60.8 (258)	69.9 (288)
Additional allowed region	28.1 (122)	28.1 (119)	21.8 (90)
Generously allowed region	6.2 (27)	7.3 (31)	6.1 (25)
Disallowed region	2.5 (11)	3.8 (16)	2.2 (9)
Total	100 (434)	100 (424)	100 (412)
Ramachandran plot values after modification			
Most favored region	76.5 (332)	76.9 (326)	78.6 (324)
Additional allowed region	18.7 (81)	17.9 (76)	17.5 (72)
Generously allowed region	2.3 (10)	2.1 (9)	2.4 (10)
Disallowed region	2.5 (11)	3.1 (13)	1.5 (6)
Total	100 (434)	100 (424)	100 (412)

3R-(39)3 repeats; 5R-(39)5 repeats; 4R-(39)4 repeats; the figure in brackets is the number of residues

Table 3: The Ramachandran plot values for models built using PHYRE2

Model	**5R**	**4R**	**3R**
Ramachandran plot values before modification			
Most favored region	88.1 (297)	87.6 (298)	87.9 (298)
Additional allowed region	10.4 (35)	11.2 (38)	10.3 (35)
Generously allowed region	1.5 (5)	1.2 (4)	1.8 (6)
Disallowed region	0.0 (0)	0.0 (0)	0.0 (0)
Total	100 (337)	100 (340)	100 (339)
Ramachandran plot values after modification			

Most favored region	92.3 (311)	90.3 (307)	91.4 (310)
Additional allowed region	6.5 (22)	8.8 (30)	6.8 (23)
Generously allowed region	0.6 (2)	0.9 (3)	1.8 (6)
Disallowed region	0.6 (2)	0.0 (0)	0.0 (0)
Total	100 (337)	100 (340)	100 (339)

3R-(39)3 repeats; 5R-(39)5 repeats; 4R-(39)4 repeats; the figure in brackets is the number of residues

Table 4: The proportion of the residues under each category in the secondary structure prediction in PHYRE2

Secondary structure	**Proportion**
Disordered	32%
Alpha helix	32%
Beta strand	27%
TM helix	3%

Table 5: Simple mean and Least square mean with standard error

Trait	**Simple mean ± SE**	**Least square mean ± SE**
AFC	1233.28 ± 194.86	1220.59 ±38.36
AFS	865.65 ± 138.85	854.73±23.94
CI	498.40 ± 156.17	444.98 ± 24.53
SP	205.30 ± 138.26	159.46 + 21.20
CS	136.22 ± 92.67	117.04 ± 13.94

SE-Standard error

Table 6: Least square mean and their standard error for AFC and AFS

Class	**Number of observations**	**AFC**	**AFS**
Mean	68	1220.59 ± 38.3	854.73 ± 23.94
	Season of birth		
Winter	26	1245.013 ± 47.87	848.05 ± 29.88
Summer	21	1203.28 ± 53.36	855.25 ± 33.30
Rainy	15	1292.36 ± 59.96	879.20 ± 37.42
Autumn	6	1141.69 ± 94.57	836.42 ± 59.02
	Period of birth		
2003-2009	6	1187.97 ± 86.57	856.58 ± 54.02
2010-2011	7	1270.74 ± 51.45	823.75 ± 32.11
2012-2014	19	1203.05 ± 38.32	883.86 ± 23.91

AFC-Age at First Calving; AFS-Age at First Service

Table 7: Least square mean and their standard error for CI, SP and CS

Class	N.O	CI	N.O	SP	N.O	CS
Mean		444.98 ± 24.53		159.46 ± 21.20		117.04 ± 13.94
	Parity					
1	69	516.99^{d} ± 28.20	69	222.14^{d} ± 24.78	71	143.99 ± 16.43
2	40	477.61^{cd} ± 33.18	51	180.96^{c} ± 27.27	60	131.60 ± 17.59
3	16	419.80^{b} ± 43.10	20	137.54^{b} ± 35.30	30	112.86 ± 20.75
4	8	357.96^{a} ± 56.96	11	88.64^{a} ± 44.35	15	90.58 ± 26.45
5	18	452.55^{bc} ± 41.83	19	167.99^{c} ± 36.35	21	106.16 ± 23.30
	Season of calving					
Winter	67	434.11 ± 31.31	70	147.32 ± 27.20	77	98.44 ± 17.63
Summer	43	479.84 ± 33.65	51	195.52 ± 27.96	62	130.13 ± 17.79
Rainy	24	449.62 ± 36.91	28	141.29 ± 31.20	34	132.03 ± 19.57
Autumn	17	416.36 ± 41.57	21	153.70 ± 33.77	24	107.56 ± 21.46
	Period of calving					
<2010	6	394.71^{a} ± 64.27	6	103.34^{a} ± 56.87	6	87.64 ± 38.08
2011-2015	59	506.52^{c} ± 23.37	62	217.48^{c} ± 19.75	62	139.97 ± 12.96
2016-2019	86	433.72^{b} ± 21.85	102	157.55^{b} ± 16.73	129	123.51 ± 9.71

N.O = Number of observations; ^{a,b,c,d}- represents the significant difference; CI-Calving Interval; SP-Service Period; CS-days from Calving to first Service

Table 8: The association of indel with reproduction traits

Repeats	AFC	AFS	CI	SP	CS
3R	1342.91 ± 45.83	877.65 ± 33.33	490.05 ± 44.41	256.35 ± 41.06	161.82 ± 26.70
5R	1260.41 ± 77.15	838.67 ± 56.10	580.76 ± 74.76	282.83 ± 69.11	182.33 ± 44.94
4R	1334.98 ± 66.81	922.87 ± 48.59	504.86 ± 64.74	241.50 ± 59.85	193.75 ± 38.92
4R(H)	1296.23 ± 31.94	847.83 ± 23.23	517.83 ± 30.95	235.43 ± 28.62	140.94 ± 18.61
p Value	0.741	0.513	0.761	0.602	0.637

Repeats	FCI	FSP	FCS	ACI	ASP	ACS
3R	490.05 ± 44.41	256.35 ± 41.06	161.82 ± 26.70	514.45 ± 63.33	259.32 ± 33.07	144.89 ± 15.31
5R	580.76 ± 74.76	282.83 ± 69.11	182.33 ± 44.94	402.52 ± 106.61	230.17 ± 55.67	169.94 ± 25.78
4R	504.86 ± 64.74	241.50 ± 59.85	193.75 ± 38.92	466.94 ± 92.32	212.65 ± 48.21	135.32 ± 22.32
4R(H)	517.83 ± 30.95	235.43 ± 28.62	140.94 ± 18.61	517.17 ± 44.14	205.04 ± 23.05	134.95 ± 10.67
p value	0.761	0.602	0.637	0.772	0.921	0.579

3R-(39)3 repeats; 5R-(39)5 repeats; 4R-(39)4 repeats; 4R (H)-(39)4 repeats (Heterozygous); AFC-Age at First Calving; AFS-Age at First Service; CI-Calving Interval; SP-Service Period; CS-days from Calving to first Service; ACI-Average Calving interval; ASP-Average Service Period; ACS-Average days from Calving to first Service.

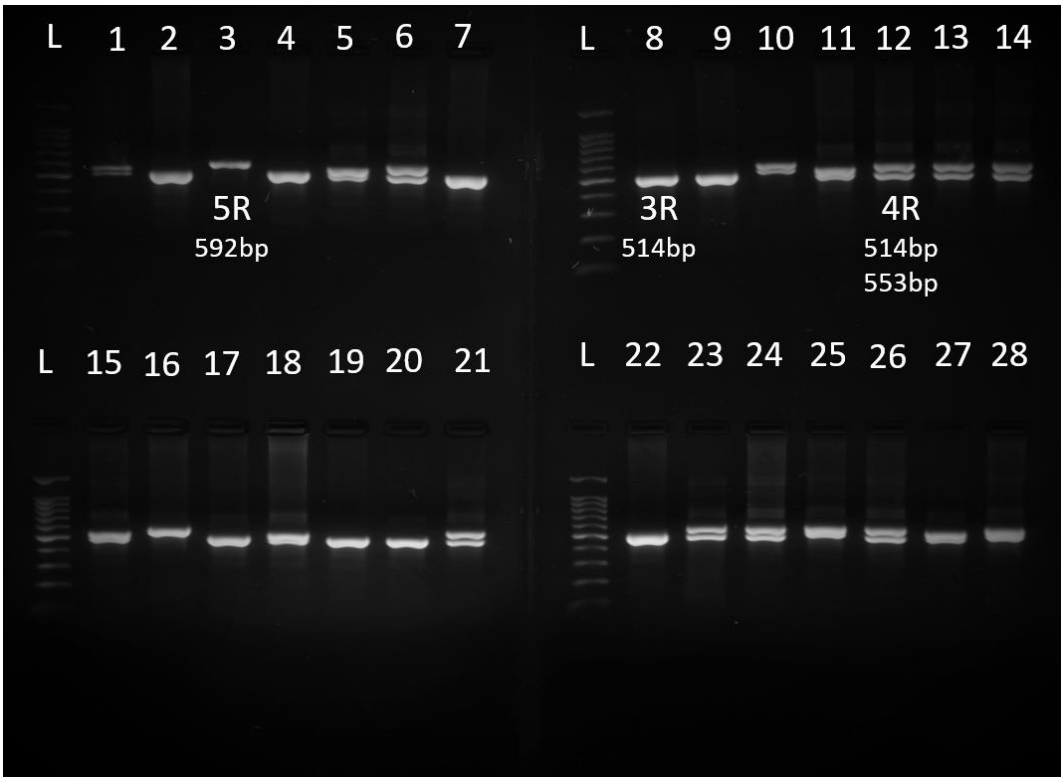

Fig 1: Gel picture with 3rd, 8th and 12th well showing 5R, 3R and 4R respectively

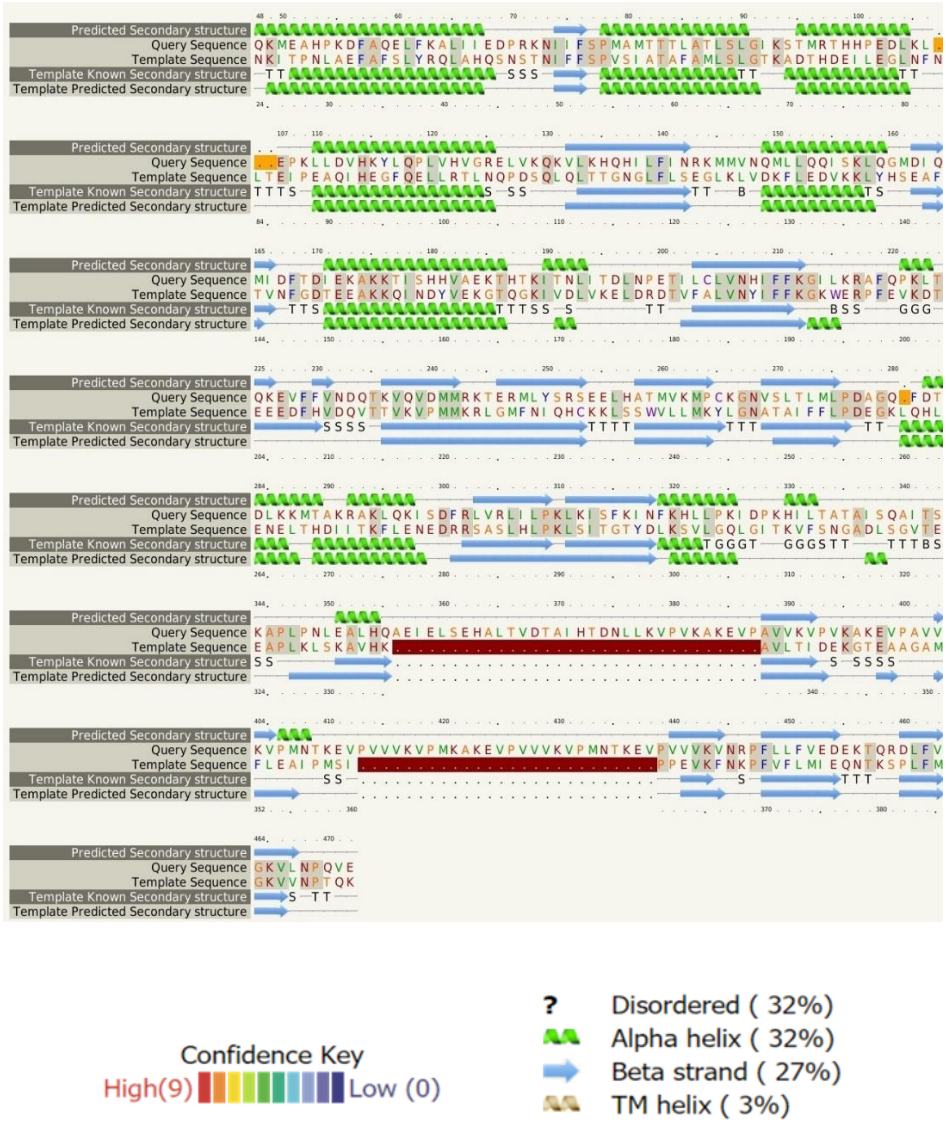

Fig 2: Secondary structure of the protein also shows the insertion of 5R. The dark band shows the insertion of the 39bp repeats.

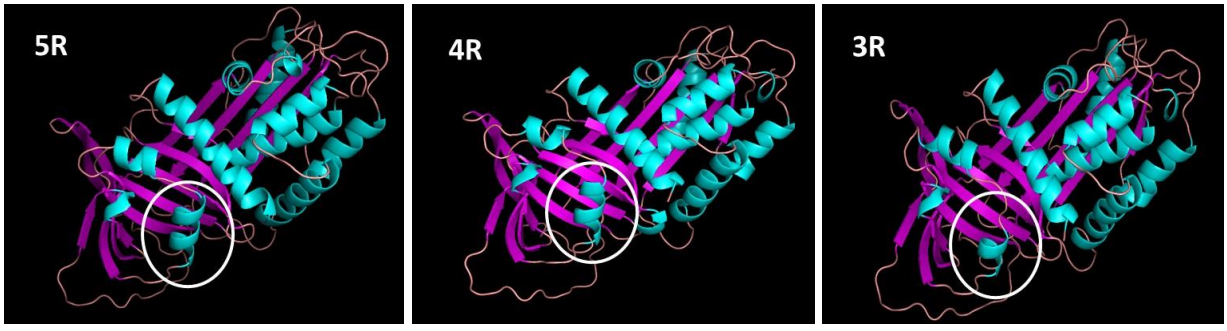

Fig 3: The final structure of the 5, 4 and 3 repeats using PHYRE2

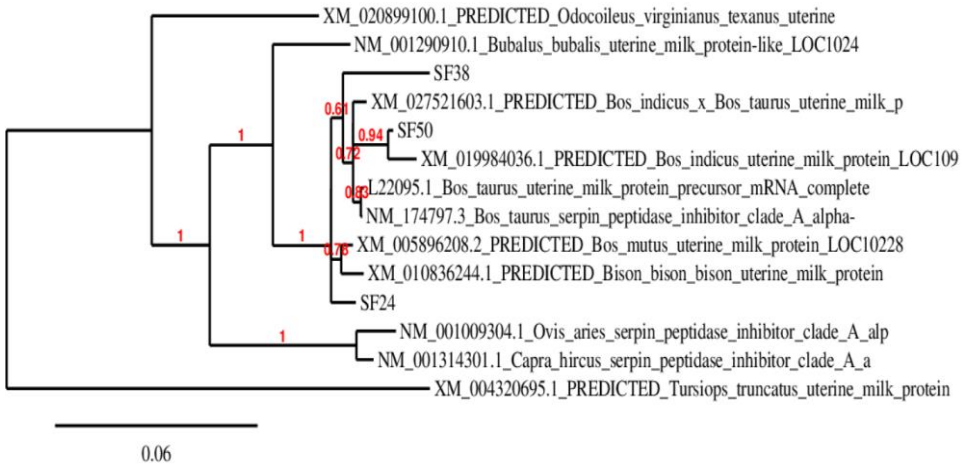

Fig 4: The phylogenetic tree of Exon 4 of SERPINA14 gene