Comparative study of SNP density in genes associated with milk traits in mammals including livestock species

Avnish K. Bhatia*, Aruna Pandey, Birham Prakash

ICAR-National Bureau of Animal Genetic Resources, Karnal-132001 (Haryana) India

ABSTRACT

Single Nucleotide polymorphisms (SNPs) are a common form of genetic diversity in organisms. SNPs are discovered throughout genomic sequence that warrants prioritization of selection of genomic regions rich in SNPs for efficient discovery and validation in laboratory experiments. Species like human, mouse, and cow have been extensively studied in the last decade and a large number of SNPs are available in online databases for these species. Assuming homology of the pattern of distribution of SNPs in genomes of various mammalian species, SNP rich genomic regions of extensively investigated species can be picked and the corresponding genomic regions in less studied species can be prioritized for SNP discovery. In the present study, 12 genes for milk traits in cow have been considered for comparative plotting of SNP density on homologous genomic regions in cow and five other mammalian species namely human, mouse, dog, horse and pig through multiple sequence alignment of genomic sequences. The bioinformatics analysis revealed conservation in SNP density in corresponding genomic regions of the investigated species providing targeted regions for increasing SNP discovery.

Keywords: Single nucleotide polymorphism, SNP density, SNP discovery, livestock, mammals, milk genes.

***Corresponding author:** avnishbhatia@gmail.com

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common and abundant type of genetic variation. SNPs are distributed throughout the genomes (Morin et al, 2004) of all species and are characterized on the basis of their location and functions. SNPs located within introns are called intronic, SNPs located in coding region that do not result in a change of the amino acid are called synonymous while those causing a change in amino acids are called non-synonymous and are the preferred SNPs for investigations attempting to link genotype to phenotype (Kharabian, 2010). SNPs have been utilised as valuable tools for genetic linkage mapping, fine mapping of candidate regions, haplotype reconstruction (Kerstens et al, 2009) and QTL identification (Panitz et al, 2007). Understanding genetic variation is the basis for diagnosis of inherited diseases in various species. For example, Shastry, 2007 examined the role of SNPs in understanding molecular mechanisms of sequence evolution, finding disease genes in human and in developing individualized medicine. Eveline *et al*, 2008 reported disease associated DNA polymorphism in genes of cattle, sheep, goat, sheep and pig.

Various online repositories of SNPs are available for a number of species. NCBI dbSNP database (Sherry et al, 2001) is the primary database of variation containing validated information on SNPs. The dbSNP database can be readily investigated using genome browsers such as Ensembl (Birney et al, 2004) and UCSC Genome Browser (Karolchik et al 2003), which provide information about SNPs characterisation such as its location on a gene, function, etc.

Species such as human, mouse, and cow have been studied extensively for SNP discovery during last

decade; consequently millions of SNPs are available for these species. Conversely, many economically important species, such as farm animal species are comparatively poorly studied and few SNPs are known. Assuming homology between these species, it can be postulated that a conserved pattern of distribution of SNPs on genomes may exist and that this may focus SNP discovery efforts into SNP rich regions of a genome. In an attempt to test this we identify SNP rich regions in genes associated with milk production conserved between model mammalian species such as human, mouse and cow. This information then provides set of target genomic regions to investigate in less well studied species such as pig, dog and horse.

METHODS

We have considered 12 genes associated with milk traits in cow [Ogorevc et al, 2009] *viz*. ABCG2, APOD, BTG3, DGAT1, CSN3, CSN1S1, LEP, PRL, LTF, CLDN8, CXCL14 and GHR to study SNPs density in homologous genomic regions of six mammalian species - human, mouse, cow, pig, dog and horse. Homologous chromosomal regions / genes and annotation were retrieved using Ensembl Biomart (http://www.ensembl.org/). The retrieved annotation included chromosome name, gene start, gene end and gene strand for each of the six species.

Multiple sequence alignments of transcripts of the twelve genes for 45 vertebrate species were obtained from UCSC genome browser. Multiple sequence alignment (MSA) of genomic DNA from the six reference species and transcripts of 45 species was performed using EMBL-EBI Kalign tool (http://www.ebi.ac.uk/Tools/msa/kalign/).

SNP data for the 12 genes in six species was retrieved using Ensembl- Biomart Variation databases. The retrieved information included Ensembl Gene-id, chromosome name, variant name, variation source, position on chromosome, strand and variant alleles.

For each gene, the corresponding position of each SNP within a transcript was determined according to the following rules. If both the SNP and the gene are on positive strand, the SNP position on gene was determined as SNP position on chromosome minus gene start position. If the gene was on the negative strand and the SNP on the positive stand, the gene SNP position is equal to the gene end minus SNP chromosome position. Where the gene is located on the positive strand and the SNP on the negative strand, the gene SNP position was calculated as SNP chromosome position minus gene start.

A program in C++ was written to insert SNPs for each species at the genomic position on multiple sequence alignment (MSA) of genomic sequences of gene and

Table 1: Gene sizes (bp) of 12 milk genes in six mammalian species.

S. N.	Gene Name	Gene Size (bp)					
		Human	Mouse	Cow	Dog	Horse	Pig
1.	ABCG2	141059	108346	117473	77398	36905	126310
2.	CSN1S1	15490	16349	17540	7304	13930	11703
3.	DGAT1	10619	9936	9211	10495	9322	9575
4.	CSN3	8840	7086	14352	9811	10292	4049
5.	GHR	298100	265732	309260	142584	130615	161415
6.	LEP	16344	13657	16750	16688	2200	2885
7.	LTF	49588	23474	34183	31415	28613	31703
8.	PRL	10250	7634	8624	30479	9534	12349
9.	APOD	15503	18616	13306	16051	16561	16320
10.	BTG3	19294	17332	18957	38198	14835	13842
11.	CLDN8	2067	2355	1987	14042	none	677
12.	CXCL14	8596	7904	8364	4223	6926	9298

Table 2: Number of SNPs for the 12 genes in six mammalian species as retrieved from Ensembl Variation Databases (Accessed in the year 2013).

S. N.	Gene Name	Number of SNPs in species					
		Human	Mouse	Cow	Dog	Horse	Pig
1.	ABCG2	2734	1900	883	136	3	0
2.	CSN1S1	43	519	189	25	2	0
3.	CSN3	41	127	132	12	2	2
4.	DGAT1	97	135	27	5	0	8
5.	GHR	216	967	1372	139	66	32
6.	LEP	264	343	178	11	0	8
7.	LTF	255	855	259	39	13	12
8.	PRL	56	220	58	0	2	10
9.	APOD	13	121	45	16	37	7
10.	BTG3	10	2144	73	45	0	85
11.	Cxcl14	3	256	42	1	3	140
12.	Cldn8	19	240	12	11	0	5

transcripts of 45 vertebrate species. Another program in C++ was written to find start and end positions of transcripts / exons on MSA of genomic sequences. A third program was written to find start and end positions of genomic segments of genes on the MSA. Finally, a script in R-programming language was used to calculate and depict the density of SNPs along each MSA. Source codes for all the programs in C++ and R-script are made available on http://figshare.com.

RESULTS

Table-1 shows gene sizes in the six mammalian species. GHR is the largest gene with size equal to 309260 base pairs in cow, while CLDN8 is the smallest gene studied with size of 1987 base pairs. A large number of SNPs are available in most of the genes for human, mouse and cow in dbSNP database. On the contrary very small numbers of SNPs are available in the dbSNP database for the other three species included in the study (see Table 2). There were 2734 SNPs for ABCG2 gene in human while there was no SNP reported in three genes in horse.

SNP density in 100bp windows along MSA of genomic sequences of BTG3 gene in the six species is shown in Figure-2. Rectangles on the alignment represent exons in the gene. The highest SNP density of 8 in human is visible in the window 47301-47400.

Further analysis was performed to determine if a relationship of SNP density exists between genomic regions of two species. SNP densities on ABCG2 gene were sorted in decreasing order on a species to find corresponding SNP density in other species so as to demonstrate similarity in distribution. ABCG2 gene was selected for the

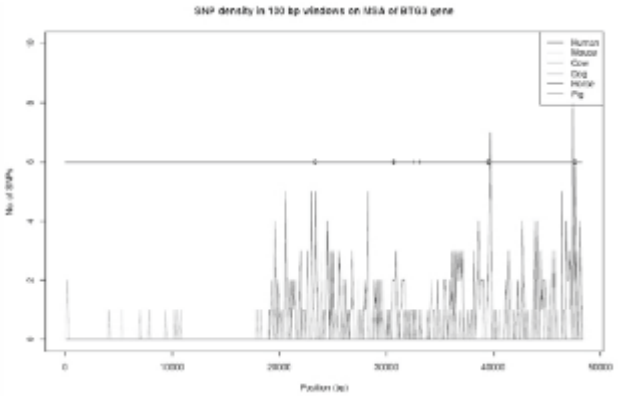

Figure 2: *SNPs density along multiple sequence alignment of genomic sequences for BTG3 gene of the six species. Rectangles on the alignment plot depict transcripts (exons).*

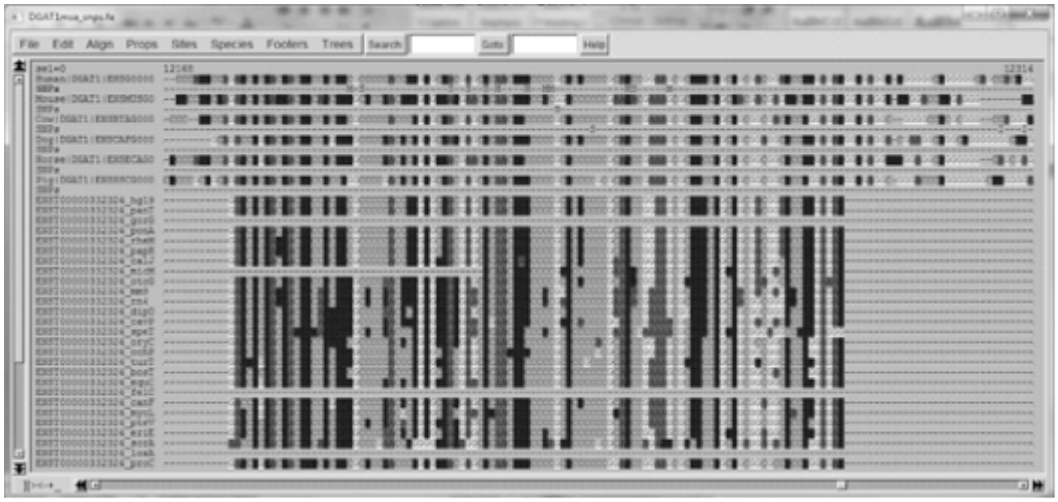

Figure 1: *Multiple sequence alignment of genomic sequences and transcripts of DGAT1 gene as viewed using SeaView 4.0 software. SNPs in the gene for each of the six species in this study are also placed in the alignment.*

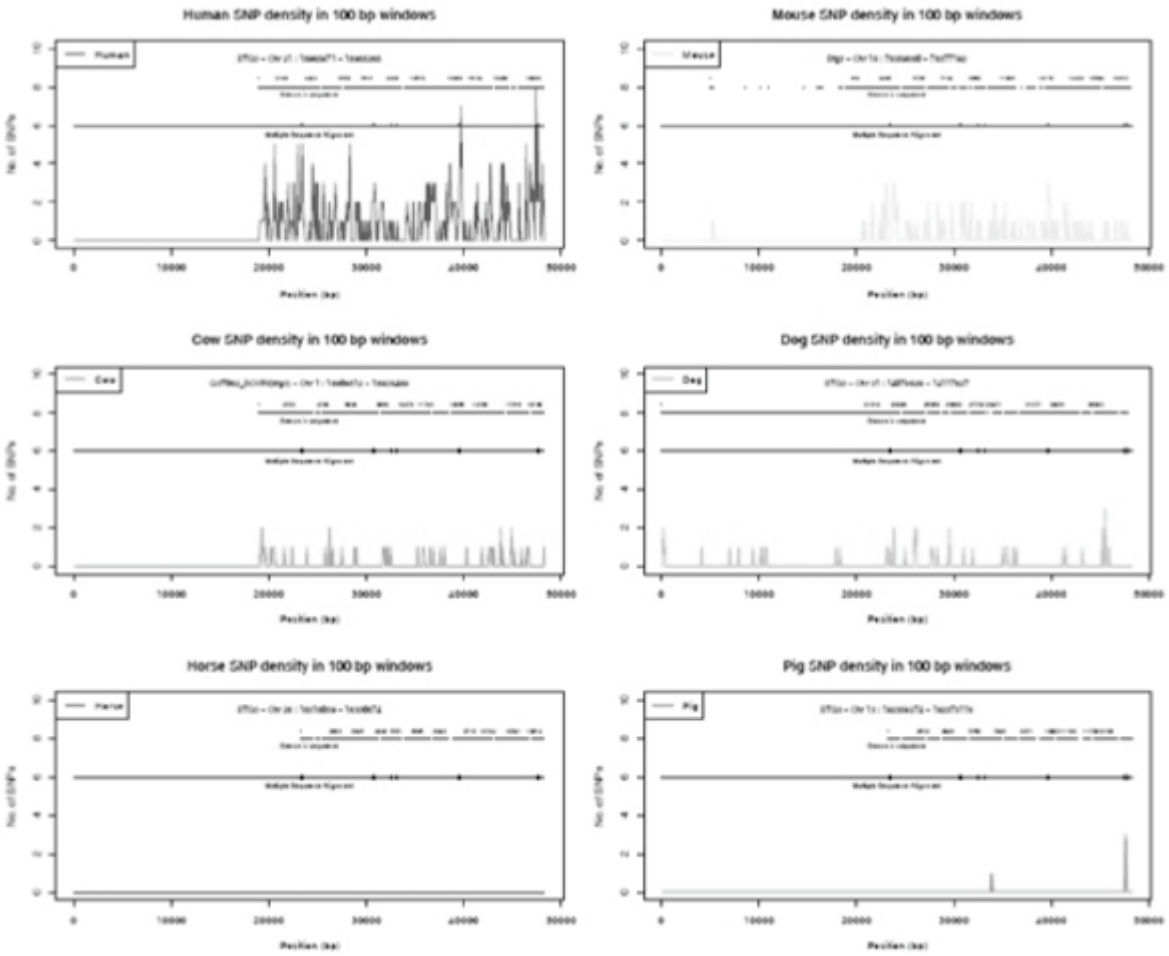

Figure 3: *SNPs density along multiple sequence alignment of genomic DNA of BTG3 gene in each of the six species. Genomic sequence of the gene for a species is also shown along with the alignment.*

Table 3: Corresponding SNP density in the other two species corresponding to SNP-rich and SNP- poor regions in multiple sequence alignment (MSA) of genomic sequences of ABCG2 gene in windows sorted in decreasing order on the third species, considering human, mouse and cow

MSA window				
From	To	Human	Mouse	Cow
Sorted in decreasing order on human				
75757	76110	49	4	2
82483	82836	33	16	2
82837	83190	30	11	7
170275	170628	0	0	0
170629	170982	0	0	0
174523	174876	0	0	0
Sorted in decreasing order on mouse				
104785	105138	6	21	1
154345	154698	3	19	1
82129	82482	4	17	2
170275	170628	0	0	0
170629	170982	0	0	0
174523	174876	0	0	0
Sorted in decreasing order on cow				
50269	50622	2	0	33
146557	146910	2	1	18
60181	60534	11	6	12
170275	170628	0	0	0
170629	170982	0	0	0
174523	174876	0	0	0

task, as it had a good number of SNPs in human, mouse and cow. Table-3 shows SNP density in the other two species corresponding to first three (SNP-rich) and last three (SNP-poor) regions along MSA of genomic sequences of ABCG2 gene in windows sorted in decreasing order on the third species. Number of SNPs corresponding to SNP-rich regions are more than zero in the other two species while number of SNPs in SNP-poor windows are equal to zero in all the three species. When the windows were sorted on mouse, there were 6, 21, and 1 SNPs in the MSA window 104785-105138 in human, mouse and cow, respectively. However, there were no SNPs in the window 170275-170628 in all the three species.

Correlation coefficients were calculated between number of SNPs in corresponding MSA windows using combinations of two species out of human, mouse and cow taking four genes - ABCG2, GHR, LEP, LTF. These four genes were selected due to availability of large number of SNPs (more than 100) in the gene in each of the three selected species. Further, MSA windows were sorted in decreasing order on number of SNPs in a gene in the species with highest number of SNPs. Correlation coefficient were calculated between combinations of two species for the first fifty windows of sorted data. Table-4 shows the correlation coefficients when all the windows in the MSA were considered and also when only the first fifty sorted windows on a species were considered. The value of correlation coefficient for first fifty sorted windows changed to positive or higher value in eight of the twelve combinations of species when compared to the correlation coefficient for all the data values. In LTF gene, the correlation coefficient between mouse-cow was equal to -0.096 when all the windows were taken into consideration but the value increased to 0.256 when the first fifty windows sorted on mouse were considered. Whilst increased correlation coefficient is not a measure of homology, it suggests a positive relationship between SNPs in corresponding genomic regions of two species.

Table 4: Correlation coefficients between number of SNPs in genomic windows (bp) on multiple sequence alignment (MSA) of genes, taking combinations of two species from human, mouse and cow. Column 'All the windows' shows the correlation coefficient for all the data values. Column 'Sorted in Decreasing Order' shows correlation coefficient for the first 50 windows in the data sorted in decreasing order on a species with largest number of SNPs in a gene.

Gene	Window size (bp)	Number of windows	All the windows			Sorted in Decreasing Order - First 50 Windows			
			Human-mouse	Human-cow	Mouse-cow	Sorted on	Human-mouse	Human-cow	Mouse-cow
ABCG2	354	500	0.101	-0.018	0.117	Human	0.318	0.103	0.153
GHR	900	500	0.006	0.138	-0.063	Cow	0.061	-0.030	0.075
LEP	100	241	-0.025	0.205	0.044	Mouse	0.0163	0.075	-0.050
LTF	132	498	0.045	-0.006	-0.096	Mouse	-0.041	0.079	0.256

DISCUSSION

SNP discovery is a major area of research in species of economic importance, such as livestock and poultry species as their association with production traits leads to improvement of animals. SNPs are distributed non-randomly in the human genome [Amos, 2010] indicating some genomic regions are SNP rich compared to other regions. Also, positions of disease associated mutations are largely conserved [Mooney, 2005] indicating conservation of SNP location in functional regions of genomes. A few model species like human, mouse and cow have been sequenced on priority and have been evaluated extensively for SNPs during the last decade and consequently a large number of SNPs for these species is available in publically accessible databases. This study has highlighted SNP-rich areas in the model species. Assuming conservation in SNP density, it has also revealed corresponding genomic regions in other species. Genomic regions in lesser explored species such as horse, sheep, goat, buffalo, etc. corresponding to SNP-rich regions of the model species can be investigated on priority for higher possibility of SNP discovery in these species.

Although all the SNPs located in genomic regions of genes are important for diversity and association studies, researchers target small portion of genome such as a few exons on a small number of candidate genes or part thereof for finding SNPs and their effect on phenotype due to cost constraints [Gupta et al, 2009; Sodhi et al, 2013]. SNPs in 24895 bp coding region of 18 canine dopamine and serotonin-related genes have been examined [Jorn and Frode, 2008]. Another consideration for exploring SNPs is to target relatively few genes in a specific pathway [Raven et al, 2013] that it can be an alternative approach to genome-wide association.

The procedure for finding SNP density in genomic sequences of genes has resulted in locating SNP-rich and SNP-poor regions in model species as well as corresponding genomic regions in other species. The analysis also suggests conservation in SNP density in corresponding genomic regions of various mammalian species. Genomic regions with high SNP density in a species have been found to correlate to at least a few SNPs in corresponding genomic regions in other species (Table 3). Significant positive correlation coefficient values between SNP rich windows also indicate positive relations of SNP density in corresponding genomic regions in different species.

The SNP rich regions in a species in the plots of MSA can be matched to genomic region in other species having little information on SNPs (Figure-3). Such genomic regions can be targeted for SNP discovery on priority. This would enhance chances of success in finding SNPs, saving time and cost of SNPs discovery.

CONCLUSION

We have defined a pipeline to draw SNPs density plots on alignments of genomic DNA of twelve milk genes in six mammalian species. The results suggest conservation of SNP density in corresponding windows on genomic region of genes. This conservation based SNP density study will be helpful for speedy SNP discovery in less studied mammalian species like goat, sheep, horse, camel etc. It will be helpful in focusing efforts on the genomic regions,

which are rich in SNPs rather than taking whole genome for exploring SNPs, thus saving time and resources for targeting SNPs responsible for diseases and production parameters in farm animal species. The bioinformatics pipeline can be replicated for other genes involved in various pathways and a database on SNP density on genomic regions of genes can be developed for use by researchers. Laboratory validation of SNP density in a few windows is desired to confirm the results of the bioinformatics analysis.

Acknowledgements: The first author is thankful to Dr. Richard Emes, School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Leicestershire, LE12 5RD, United Kingdom for hosting his visit during June-August, 2012 for an international training in animal bioinformatics, and valuable suggestions on the manuscript. Funding of the visit by Indian Council of Agricultural Research under National Agricultural Innovation Project is thankfully acknowledged. Development of R-script for plotting SNPs by Harry Clifford, School of Biology, University of Nottingham, United Kingdom is thankfully acknowledged.

REFERENCES

Amos W. 2010. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proc Royal Society B*, 277:1443-1449.

Birney E, Andrews TD, Bevan P *et al.* (2004) An Overview of Ensembl. Genome Research 14(5): 925–928.

Eveline M, beagha-Awemu I, Kgwatalala P, Ibeagha AE and Zhao X. 2008. A critical analysis of disease-associated DNA polymorphisms in the genes of cattle, goat, sheep, and pig. *Mammalian Genome* 19:226–245.

Gouy M, Guindon S and Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27(2): 221-224.

Gupta SC, Kumar D, Pandey A, Malik G and Gupta N. 2009. New *k*-Casein alleles in Jakhrana goat affecting milk processing properties. *Food Bioitechnology* 23:83-96.

Jorn V and Frode L. (2008) Single nucleotide polymorphisms (SNPs) in coding regions of canine dopamine- and serotonin-related genes. *BMC Genetics* 9:10.

Karolchik D, Baertsch R, Diekhans M *et al.* 2003. The UCSC genome browser database. *Nucleic Acids Research* 31(1): 51-54.

Kharabian A. 2010. An efficient computational method for screening functional SNPs in plants. *Journal of Theoretical Biology* 265(1): 55-62.

Kerstens HH, Kollers S, Kommadath A *et al.* 2009. Mining for single nucleotide polymorphisms in pig genome sequence data. *BMC Genomics* 10:4.

Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism analysis. *Briefings in Bioinformatics* 6(1):44-56.

Morin PA, Luikart G and Wayne RK. 2004. The SNP Workshop Group. SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution* 19: 208-216.

Ogorevc J, Kunej T, Razpet A and Dovc P. 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics* 40(6):832-51

Panitz F, Stengaard H, Hornshøj H *et al.* 2007. SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. *Bioinformatics* 23(13): i387-391.

Raven L-A, Cocks BG, Pryce JE, Cottrell JJ and Hayes BJ. 2013. Genes of the RNASE5 pathway contain SNP associated with milk production traits in dairy cattle. *Genetics Selection Evolution*, 45:25.

Shastry BS. 2007. SNPs in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics* 52:871–880.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L and Smigielski EM. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29 (1): 308-311.

Sodhi M, Mukesh M, Kishore A, Mishra BP, Kataria RS and Joshi BK. 2013. Novel polymorphisms in UTR and coding region of inducible heat shock protein 70.1 gene in tropically adapted Indian zebu cattle (*Bos indicus*) and riverine buffalo (*Bubalus bubalis*). *Gene* 527(2):606-15.