Indian Journal of Animal Sciences **84** (5): 475–489, May 2014/Article

Statistical and bio-computational applications in animal sciences

A R RAO[1], MANOSWINI DASH[2], TANMAYA KUMAR SAHU[3], S D WAHI[4], B K BEHERA[5], A P SHARMA[6] and V K BHATIA[7]

Indian Agricultural Statistics Research Institute, New Delhi 110 012 India

Received: 19 August 2013; Accepted: 27 January 2014

ABSTRACT

The demand for food proteins, including plant and animal proteins is increasing at an exponential rate. The demand for animal products will nearly be doubled by 2030. Thus, to improve livestock production and meet the animal protein demand, it is essential to go for application of interventions based on genomics, statistics and informatics. Such interventions are quite often used in the animal improvement programs to develop offspring with desirable traits. More recently, with the emergence of high throughput sequencing technologies, genomes of farm animals, fishes and model organisms were sequenced and the same are available in public domain. Also, with the advent of new silicon technologies, it has become possible to manage the generated data from genome sequencing projects. Now, the challenge lies with the analysis and interpretation of sequence data in a biologically meaningful manner, for which many algorithmic based analytical techniques and high performance computing methods were developed. Here, a brief review is presented on the application of various statistical and computational approaches used in genomic data analysis. Applications of the above mentioned approaches for health management and sustainable animal and fish production from the view point of vaccine and drug designing, disease risk management, epigenomics and whole genome level SNP/CNV associations with traits at are also discussed here. Besides, this paper allows the molecular biologists and other application scientists to analyze overwhelming amount of genomic data by different methods outlined here.

Key words: Allele mining, CNV, Drug discovery, GWAS, Machine learning, SNP, Transcriptomics, Vaccine designing

Farming and domestication of livestock was integrated to human civilization in such a way that it contributes to the sustainable livelihoods and security to poor and smallholding farmers (Thornton *et al.* 2006). Livestock accounts for about 30% of the Agricultural Gross Domestic Product (AGDP) in the developing world and about 40% of the global Agricultural GDP (World Bank 2009). Feeding the world's population is one of the major challenges in the present day situation due to alarming population growth. Estimate of the world's population is approximately 7 billion with annual growth rate of 1.15%. While in India alone, it is approximately 1.2 billion with annual growth rate of 1.7% contributing a major share to the global population (http: / /censusindia.gov.in/). Livestock play an important role to meet out the demand of animal proteins. It provides around 12.9 % of global calories and contributes around 27.9 % of proteins directly through provision of meat, milk, eggs and offal. It also contributes to crop production in the form of transport and manure (FAO 2011). The rapid growth in the meat sector was also underpinned by rising demand for poultry meat, which has consistently increased at around three times the rate of population growth over each of the past 5 decades (FAO 2012a). FAO study on the distribution of livestock reveals the fact that Asia occupies top position in livestock production. According to the 18th Livestock census of India (2007), population of livestock is 3.28 crore and of poultry 4.24 crore. The number of cattle and buffaloes (domestic) stood at 105.07 lakh and 43.29 lakh respectively (http: //www.dahd.nic.in/dahd/statistics/livestock-census.aspx). As per FAOSTAT (2010), the livestock production and its contribution to agricultural GDP of India increased from 27.2% in 1995 to 31.9% in 2007. Although livestock population in the world is growing, the overall growth rate in livestock sector is steady and found around 4–5 % (Thornton 2010). Livestock products contribute remarkably in fulfilling the food demand, and helps in reducing poverty in developing countries. Besides, livestock also provides animal traction to almost a quarter of the total area under crop production (Devendra 2010). In addition to livestock production, fish production has a major contribution towards human nutrition in terms of protein supplement. According to FAO reports, aquaculture

Present address: [1]Senior Scientist (rao.cshl.work@gmail.com), Center for Agricultural bioinformatics, [2-3]Research Associate (manoswini.dash@gmail.com, tanmayabioinfo@gmail.com), [4]Principal Scientist and Head (sdwahi@iasri.res.in), Statistical Genetics, [7]Former Director (vkbhatia7251@yahoo.co.in). [5] Senior Scientist (beherabk18@yahoo.co.in), [6]Director (apsharma1 @gmail.com), Central Inland Fisheries Research Institute, Barrackpore, West Bengal.

represents the fastest-growing animal-based food producing sector. Around 56 million tonnes of fish and fishery products were produced by aquaculture in 2009, which accounts for around half of fish and fishery product world supplies for human consumption across the world. China has a share of over 60 % of world's aquaculture production, while Asia accounts for almost 90 % (FAO 2012b). Further, the ever increasing world's human population demands more production of livestock, fish and animal protein. It projects that the demand for animal products will nearly be doubled by 2030. Thus, to meet the increasing animal protein demand, it is essential to go for application of interventions based on genetic engineering, biotechnology, genomics, statistics and computer applications for improving livestock and fish production. Therefore, the integration and implementation of new techniques evolved in these emerging fields is highly needed for sustainable production of livestock and fish.

With the advent of molecular biology techniques, genome sequencing has become technically more feasible as well as less expensive, which lead to generation of massive amount of data. The International Technical Conference of FAO has documented an unprecedented surge in the development of biotechnology in animal production and health during the last 4 decades (FAO 2010). Genomes Online Database (GOLD) has documented 4,015 completed genome projects, which include 175 Archeal, 3,657 bacterial and 183 Eukaryal genomes as on December 2012. While GenBank has released the final genome sequences of 2,343 completed genome projects. Many genome sequencing projects were completed at the level of a draft genome, identified as *permanent draft*, and their final sequences were submitted in GenBank. There have been 1,672 *permanent draft* genomes available at GenBank till December 2012 (Pagani 2012). To be more precise, there are 381 chordate genome projects finished and genome data of model organisms, farm animals and fishes are available in the public domain. According to National Center for Biotechnology Information (NCBI) statistics on genome sequencing projects, so far 1,153 projects have been completed while in 1,285 projects the draft assemblies of genomes are available. Till date, genomes of fish species, viz. Zebra fish, Tetradon, Fugu, Medaka, Salmon, Stickleback, and Trout have been sequenced and genome sequencing of Tilapia, Coelacanth, Gar (Spotted), Hagfish, Lamprey, Shark, Skate, Slender lung fish, turtle is in progress (NHGRI 2011). The MitoFish database (http: // mitofish.aori.u-tokyo.ac.jp/) has mitochondrial genomes of 1,044 fish species sequenced till September 2012.

In order to deal with such massive amount of genomic data, there is a need to store, manage, analyze and interpret the data in a biologically meaningful manner, so that retrieving all kinds of underlying biological information from sequence data (DNA/RNA/Protein) will be possible through various computational and statistical approaches. Thus, the aim of this manuscript is to acquaint the application scientists with the various statistical and bio-computational techniques currently being used in animal and fish genomics, with special reference to health management and sustainable animal and fish production.

Statistical and Algorithmic Applications in Animal Sciences

Breeding and genetics: Most of the economic characters in livestock and fish species are quantitative in nature. The existence of a wide range of variability in these characters depends on the genetic makeup of the individuals and the environment in which they are grown. Breeders use this variability for getting improvement in economic characters through efficient selection strategies. The information on genetic parameters, such as heritability, repeatability and genetic correlation is a prerequisite for making efficient selection strategies by the geneticists and breeders. Prabhakaran (1998) reviewed the research work done in the field of statistical genetics in India. Rao *et al.* (2009) gave a brief account on the important research and academic accomplishment in the area of statistical application in genetics and breeding. They have discussed various statistical applications in animal sciences, viz. improved estimation of genetic parameters using resampling techniques (Bhatia *et al.* 1994, Bhatia and Jayasankar 1996a,b, Singh *et al.* 2006, Kiran *et al.* 2007, Prabhakaran and Rao 2008), Bayesian estimation (Rao and Kumar 2001, Kumar *et al.* 2004), robust estimation (Kiran *et al.* 2003, 2004, Singh *et al.* 2007), yield survival relationships and culling pattern (Bhatia *et al.* 1992a,b, Bhatia and Malhotra 1995a,b, Paul and Bhatia 2000a,b, Paul and Bhatia 2001, Paul and Bhatia 2002a,b), improved estimation and effect of outliers on estimation of genetic correlations (Sarika *et al.* 2006a,b, Wahi *et al.* 2006), modeling of biological phenomena in animal breeding using discriminant analysis, fitting of growth models of milk and body weight traits (Wahi and Bhatia 1995, Chand *et al.* 1997, Chand *et al.* 1998, Wahi *et al.* 2004). Meher *et al.* (2010) discussed various methods of detection of multivariate outliers in breeding data. Wahi and Rao (2011) made an attempt to compare the estimated, predicted, empirical and bootstrap Standard Errors (SE) of genetic correlation for different family sizes and structures under half-sib mating design. Behera *et al.* (2013) studied the genetic diversity of 5 feral populations of Asian Sea bass, *Lates calcarifer* in India using randomly amplified polymorphic DNA.

In the recent past many statistical and Machine Learning Approaches (MLA) have been used to solve the problems in the fields of animal breeding and genetics. The MLAs are computationally intensive and benefit greatly from progress in computer speed. Application of machine learning methods offers the cost-effective tools for building predictive models from biological data, e.g., for annotating new genomic sequences, for predicting macromolecular functions, for identifying functionally important sites in proteins, for identifying genetic markers of diseases, and for discovering the networks of genetic interactions that orchestrate important biological processes. The application of artificial neural network (ANN) in goats to predict the

weekly milk production and clustering of goat flocks was studied by Fernandez *et al.* (2006). Caraviello *et al.* (2006) used the alternating decision tree algorithm to identify factors affecting the reproductive performance of lactating Holstein cows on large dairy farms using data from farms in the Alta Genetics Advantage progeny-testing program. They identified variables affecting first-service conception rate and probability of pregnancy by 150 DIM on large dairy farms, showing interactions between the variables, and also build predictive models for the important reproduction traits. Their results indicated that pregnancy status at 150 Days-In-Milk *(*DIM) is a more robust measure of the overall reproductive performance of a commercial dairy farm than first-service conception rate. Druet and Georges (2010) applied Hidden Markov Model (HMM) with Linkage and Linkage Disequilibrium information for Haplotype reconstruction and Quantitative Trait Loci (QTL) fine mapping applicable in animal genetics. They proposed that this method is computationally effective at handling large data sets based on high-density SNP panels and can directly be used to fine map QTL. Dale *et al.* (2010) tried several machine learning methods in metabolic pathway predictions. Shahinfar *et al.* (2012) investigated the prospective of two types of intelligent learning methods *i.e.* Artificial Neural Networks (ANN) and Neuro-Fuzzy Systems (NFS), in order to estimate breeding values (EBV) of Iranian dairy cattle. Their study revealed that NFS tended to provide greater predictive ability, consistent results than ANN. Besides, NFS may be more robust to "noise" in specific data sets. Oudendag *et al.* (2012) applied Multiple Linear Regression (MLR) and Neural Networks (NN), to predict dairy farm size change in the Netherlands. Their results revealed that the chosen methods are able to predict the change in farm size effectively, and that prediction quality is best when the aim is to predict farm size for 4 years. A comparative analysis on the efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation (305-day milk yield) in Sahiwal cattle carried out by Dongre *et al.* (2012) have recommended that ANN was better than the multiple linear regression analysis for the prediction of the first lactation in Sahiwal cows. However, instead of traditional techniques, the advanced MLAs, decision trees and statistical methods could be implemented to manage the problems associated with animal breeding, reproduction and genetics thereby contributing to livestock improvement.

Genome selection: The main aim of any animal improvement programme is to change the genetic endowment of animals to improve the performance of economically important traits. Traditionally, selection was done on the basis of phenotypic performance. Smith (1967) and Soller and Beckman (1983) have initially given the idea to improve rate of genetic gain in cattle by using DNA markers. Later on adaption of marker assisted selection (MAS) was accepted in the dairy industry. However adaption of MAS is slow due to several reasons like large number of loci affecting the traits with each locus capturing a limited proportion of total genetic variance. Hence, small gains were possible with the available limited number of markers. However, with the advent of recent genome sequencing techniques it has become possible to discover thousands of single nucleotide polymorphism (SNPs) from dense SNP arrays that cover the whole genome and that explain majority of genetic variation in important traits through an approach called genomic selection or whole-genome selection proposed by Meuwissen *et al.* (2001). Genomic selection (GS) has recently revolutionized dairy cattle breeding in animal improvement programs. It offers the animal breeders to reduce cost, decrease generation interval, increase the accuracy and intensity of selection, which in turn increases the rate of genetic progress for economically important dairy traits. GS refers to selection decision based on genomic estimated breeding values (GEBV). The genomic breeding values are calculated using a prediction equation based on SNP effects derived from a subset of animals in the population (a reference population) that have SNP genotypes and phenotypes for traits of interest. Schaeffer (2006) divulged that genomic selection could lead to a doubling of the rate genetic gain obtained through selection and breeding from bulls at 2 years of age rather than 5 years of age or later. Goddard and Hayes (2007) explained that selection in which genetic markers covering the whole genome are used so that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker. Hayes *et al.* (2009) reviewed the progress of genomic selection and discussed the issues related to improvement of accuracy of genomic breeding values as well as the effect of genomic selection of long term genetic gain.

GEBV using genome wide genetic markers had gained considerable appreciation in the recent past. A variety of methods, viz. least squares (LS) approach, best linear unbiased prediction (BLUP), and Bayesian methods were proposed to estimate the total breeding value (Meuwissen *et al.* 2009). Lillehammer *et al.* (2011) compared a conventional breeding scheme in which no genomic selection was used with a breeding scheme in which genomic selection was used to preselect young bulls for progeny testing and also to select elite sires directly at a younger age. Pungpapong *et al.* (2012) applied a penalized-orthogonal-components-regression (POCRE) method to estimate / predict breeding values and enforced genomic selection. Their study revealed that POCRE provides similar or even better accuracy of predicting breeding values than methods based on Bayesian approaches in terms of both simulation studies and real data analyses.

Genomic selection has gained much attention in livestock to increase the predictive accuracy and the genetic gain by using dense marker information. Recio and Forni (2011) analyzed discrete traits like pregnancy outcome and disease risk in livestock, using 2 threshold versions of Bayesian regressions (Bayes A and Bayesian LASSO) and 2 machine learning algorithms (boosting and random forest), on a genome-wide prediction context. The results of this study suggested that the best method for genome-wide prediction

may depend on the genetic basis of the analyzed population. However, machine-learning showed some advantages over Bayesian regressions.

Genome wide association studies: Whole genome association (WGA) studies also referred as genome wide association studies (GWAS) is an approach that involves scanning of markers (SNPs) over genomes of 2 types of individuals (Cases and Controls) to find the genetic variations associated with a particular trait (disease, height, color etc.). This approach is mostly useful to identify the genetic variants associated with complex diseases and the identified variants can thus be used for disease diagnosis, treatment and prevention.

With the completion of many genome projects, it is now feasible to analyze whole genome level markers and select only the specific markers related to a trait. Compared with traditional QTL mapping strategies, GWAS confers major advantages both in the power to detect causal variants with modest effects and in defining narrower genomic regions that harbor causal variants (Hirschhorn and Daly 2005). GWAS was extensively applied to study human diseases and many success stories identified genetic variation in type-2 diabetes, Parkinson's disease, heart disorders, obesity, Crohn's disease and prostate cancer (http: // www.genome.gov/20019523). At present, GWAS is applied on farm animals and many genes or regulatory elements which are associated with important economic traits were identified. Zhang *et al.* (2012) reviewed that 734 SNPs, located close to the DGAT1 gene, were associated with milk yielding trait of cattle and also reported that 547 SNPs on chromosomes 5, 6, 11, 14, 19 and 26 and related genes were identified to be associated significantly with milk quality in cattle. Besides, they have discussed that GWAS also was applied to other livestock species like pig, horse, sheep and chicken to identify the genes associated with economically important traits like fertility, growth, meat quality, fatness, body weight, egg production, diseases etc. These findings certainly help the animal breeders and animal biotechnologists to understand the genetic architecture of complex traits and thereby improve the health as well as production of livestock and animal proteins.

Traditional computational methods used in GWAS have shown their inability to analyse high dimensional data, i.e., large number of SNPs distributed over genome. Some of the most recently used statistical approaches in genome wide association studies are least absolute shrinkage and selection Operator (LASSO, Tibshirani 1996), support vector machine (SVM, Witten and Frank 2005, Ban *et al.* 2010), and random forest (Breiman 2001) etc. Goddard *et al.* (2009) suggested an integrated approach to estimate SNP effects and predict the trait value by treating SNP effects as random instead of fixed. They suggested that statistical methods traditionally used in the prediction of trait values in the genetics of livestock, can be applied to analysis of GWAS, giving better estimates of the SNP effects and predicted phenotypic and genetic performance in individuals.

Disease risk prediction of complex diseases is another important field in epidemiology where GWAS play an important role. To predict the onset and progression of disease is one of the cornerstones of epidemiology (Jostins and Barrett 2011). Traditionally, disease risk prediction was based on certain environmental factors and clinical factors associated with a disease, which have shown very less success ratio. Even though, significance of genetic factors in disease risk prediction had gained appreciation, with the intervention of GWAS it seems more practical to find significant number of unique variants genetically associated with the disease.

Wray *et al.* (2007) investigated the number and effect size of risk loci that underlie complex disease constrained by the disease parameters of prevalence and heritability. Their approach can be used to assess the genetic risk of a disease in healthy individuals, based on dense genome wide SNP panels. Daetwyler *et al.* (2008) presented simple expressions for the genome wide accuracy of prediction of genetic disease risk and derived expressions for continuous traits as well as dichotomous disease traits data obtained either from population or case-control studies. Recently Vazquez *et al.* (2012) developed a whole genome prediction (WGP) model to predict skin cancer risk in human. They incorporated a large no. of markers from GWAS, family history or ancestry and extended the Bayesian LASSO regression with a probit link to model skin cancer. Prediction accuracy was evaluated using the Area Under the receiver operating characteristic Curve (AUC) and a significant increase in prediction accuracy was reported. Colombani *et al.* (2012) applied Bayes A and Bayesian LASSO to 2 reference populations consisting of 3,940 Holstein bulls and 1,172 Montbeliarde bulls with approximately 40,000 polymorphic SNPs. They compared the accuracy of the Bayesian methods with pedigree-based BLUP, genomic BLUP, partial least squares (PLS) regression, and sparse PLS regression, for the prediction of traits like milk yield, fat content and conception rate. Based on the study, they suggested that Bayesian methods are more efficient for genomic evaluations of French dairy cattle. Similarly, Thomas *et al.* (2012) used regularized S-LASSO to examine how long different T cells may spend in an individual lymph node by observing data from long term cannulation of blood and efferent lymphatics of a single lymph node in the sheep. Their results demonstrated that the rapid recirculation of lymphocytes, observed at a macro level, is compatible with their predominant randomized movement within lymph nodes.

Martiskainen *et al.* (2009) used a 3-dimensional accelerometer and a multiclass support vector machine (SVM) to develop and pilot a method which will automatically measure and recognize several behavioral patterns of dairy cows. Based on the results, they revealed that SVMs are very useful in classification of measured behavior patterns. Support vector machine also being used for intelligent diagnosis of animal diseases and the results showed that the model can carry out animal diseases

diagnosis more accurately, rapidly with good predictability even in small sample situation and provide a new approach for animal disease diagnosis (Wan and Bao 2009). Application of SVM was also reported in fish species by Masoum *et al.* (2007) where they successfully distinguished wild and farm salmons using fish oils extracted from their white muscles. Though random forest is not fully explored in field of live stock genomics, it can be applied to solve protein folding problem, marker association with the traits, prediction of splice sites etc.

Copy number variation (CNV) analysis: Similar to single nucleotide polymorphism (SNP), copy number variation (CNV) is another important constituent of genetic diversity found in different populations. The CNV region is a DNA fragment of length 1kb to 5mb, which is at the cytogenetic level of resolution, present in individual's genome at a variable copy number while compared to reference genome (Eichler 2008). It includes both insertions and deletions. Recent evidences suggested that CNVs play a very important role in evolution, gene expression and diseases. CNVs were extensively studied in human after sequencing the whole genome of human. According to the Database of Genomic Variants (DGV; http: //dgvbeta.tcag,ca/dgv/app/home?ref= NCBI36/hg18), there are 179,450 CNVs identified so far in human genome which covers more than 53% of the human genome.

Recently, CNV studies were conducted in farm animals especially in cattle, swine, goat, horse and chicken. In goat, 127 CNV regions covering about 11.47 Mb of the virtual goat genome referred to the bovine genome were identified by Fontanesi *et al.* (2010) and their significance also proved with reference to cattle CNV regions. The total length of CNV regions reported in cattle only covers 0.13% (3.3 Mb) to 5.57% (141.8 Mb) of the cattle genome, which is very low as compared to human genome (Jiang *et al.* 2012). Wang *et al.* (2012a) identified 382 CNV regions, which cover 95.76Mb of the swine genome and corresponds to 4.23% of the autosomal genome sequence. The length of these CNV regions ranged from 5.03 to 2,702.7kb with an average of 250.7kb, and the frequencies of them varied from 0.42 to 20.87%. Likewise, 775 CNV regions were identified in horse which includes genes regulating blood group antigens, coat color, fecundity, lactation, keratin formation, neuronal homeostasis, and height (Doan *et al.* 2012). In chicken, 130 CNV regions were identified and out of which 104 were reported for the first time. Comparative and gene ontology analysis of the novel CNV regions revealed that the gene for positive regulation of phospholipase A2 is significantly over represented (Wang *et al.* 2012b).

From different studies on CNVs of human and other species, it is well understood that CNVs are not just innocent bystanders that deny causal relationship to variable traits, but are likely to be involved actively in phenotypic diversity, challenging current concepts, that genetic variation is more than a heritage of the past. Dube *et al.* (2008) presented an algorithm which is mathematically sound and computationally efficient to accurately analyze CNV in a DNA sample utilizing a nanofluidic device, known as the digital array. This numerical algorithm is utilized to compute copy number variation and the associated statistical confidence interval and is based on results from probability theory and statistics. Noushin (2012) described various computational tools for CNV detection using probe-level analysis of affymetrix SNP arrays. However, application of CNV studies in aquatic species and other livestock species will provide a more comprehensive map of copy number variation and also will be an important resource for investigation of genome structure and disease in future.

BIO-COMPUTATIONAL APPROACHES IN ANIMAL GENOMICS

With the advent of sophisticated sequencing technologies, a number of sequencing projects were completed by leaving behind a voluminous data. Bioinformatics consisting of various bio-computational approaches play an important role in deriving information and knowledge from such data. Bioinformatics being considered as a mixture of computer science, data management and genome sciences, bioinformatics now encompasses both conceptual and practical tools for the propagation, generation, processing and understanding of scientific ideas and biological information. In the following subsections major biological problems and their solutions using bio-computational approaches are discussed.

Transcriptome analysis: Transcriptome is the total RNA content present in a single cell or population of cells. It is again broadly classified into coding (mRNA) and non-coding RNAs, viz. rRNA, tRNA, miRNA, siRNA, shRNA, snRNA, snoRNA. In comparison to genome, transcriptome contribute a very less percent age to the total cell content. The total quantity of transcriptome varies as it represents the number of genes expressed in a particular cell at a given instance of time. These variations underlie the wide range of physical, biochemical, and developmental differences seen among various cells and tissues and may play a role to differentiate between healthy and diseased individuals (Adams 2008). Transcriptome attains more complexity as compared to genome with the increase in the complexity of organism. Therefore, collection and comparison of the transcriptome will help researchers to get an in-depth knowledge about gene expression like, what a specific cell type comprises of and what is the role of transcriptional activity towards a disease condition. With greater depth of sequencing and application of unbiased sampling approaches, it was revealed that majority of the genomes are transcribed and that much of the transcription is overlapping, such that multiple transcripts on either or both strands can use the same stretch of genomic DNA (Katayama *et al.* 2005). Earlier the gene expression study was based on microarray techniques or RT-PCR techniques. Limitation of these 2 techniques is lack of exquisite sensitivity and expensiveness for genome wide study respectively (Mardis 2008). Recently, the next generation sequencing (NGS), a much faster and cheaper method than

the traditional sanger sequencing method, provided a better opportunity for transcriptome analysis.

Next generation sequencing (NGS): NGS is based on sequencing-by-synthesis methodology called as pyrosequencing. Pyrosequencing technique used for transcriptome analysis is called as short read massively parallel sequencing or RNA-Seq or whole transcriptome shotgun sequencing (WTSS). Different sequencing platforms are available with little technical differences, read length, speed and cost. These platforms are used as per the requirement.

Most commonly used platforms are 454 sequencing, genome analyzer, SOLiD, Pyromark ID. 454 sequencing generates reads of length 250nt is a faster technique as up to 400000 reactions can be run in parallel (Mardis 2008). On the other side, non-synchronised extension due to less efficiency of apyrase to degrade excess nucleotide is one of the limiting factors. Genome analyzer is another platform which uses traditional sequencing method with single molecule amplification technology. However, reads generated are of 25 to 35nt length which reduces the enzyme activity but generate stronger signals allowing a much higher density of reaction. SOLiD is another platform, which is based on ligation based sequencing with read length 35 to 100nt. Like genome analyzer, it also generate strong signal but degradation of templates over time and

Table 1. Different NGS analysis software

Name	URL	Purpose
CLC Bio Genomics WorkBench	http://www.clcbio.com/	De novo and reference assembly of Sanger, Roche FLX, Illumina, Helicos, and SOLiD data
Galaxy	https://main.g2.bx.psu.edu/	Interactive and reproducible genomics. A job webportal
Genomatix	http://www.genomatix.de/index.html	Integrated Solutions for Next Generation Sequencing data analysis
JMP genomics	http://www.jmp.com/software/genomics/	Next generation visualization and statistical tool from SAS.
NextGENe	http://softgenetics.com/NextGENe.html	Analysis of sequencing data produced by the ION PGM™, Roche Junior, Illumina MiSeq as well as high throughput systems as the Ion Torrent Proton, Roche FLX, Applied BioSystems SOLiD™ and Illumina® GA, and HiSeq systems
SeqMan genome analyser	http://www.dnastar.com/t-sub-products-genomics-seqman-ngen.aspx	Next Generation sequence assembly of Illumina, Roche FLX and Sanger data integrating with Lasergene Sequence Analysis software for additional analysis and visualization capabilities. Can use a hybrid templated/de novo approach
SHORE		Used as a mapping and analysis pipeline for short DNA sequences produced on an Illumina Genome Analyzer. A suite created by the 1001 Genomes project
Real time genomics	http://www.realtimegenomics.com/	High throughput NGS data analysis
Avadis NGS	http://www.avadis-ngs.com/	State-of-the-art algorithms for analysis, visualization and management of next-generation sequencing data on a wide range of computing infrastructures
BC platforms	http://www.bcplatforms.com/Solutions/Sequence-Data-Management.html	Statistical genetic epidemiology analysis of variation data produced using Next Generation Sequencing devices
BIOMOL-NGS	http://www.biomol-ngs.com/	Fast and accurate analysis of data from large next generation sequencing projects, including whole genome resequencing as well as targeted resequencing of specific regions of known genomes
DNA nexus	https://dnanexus.com/	NGS Sequence analysis and visualization

comparatively-slow-process are the two loopholes (Hossain *et al.* 2009). Pyromark ID is a medium throughput platform, which can run 96 samples in parallel and read-lengths are around 40nt. Basically it is used for SNP analysis, mutation study etc. Among all, Roche 454 sequencing is the most commonly used NGS platform and also supported by wide publications. In literature, mammalian transcriptomes were sequenced using genome analyzer (Mortazavi *et al.* 2008) and transcriptome profiling of stem cell was done using ABI SOLiD sequencing (Cloonan *et al.* 2008). Glazov *et al.* (2008) used Illumina sequencing platform and detected 449 novel and all known chicken miRNAs in the chicken embryo.

Transcriptome data generated from different NGS platforms are subjected to different computational and statistical analysis procedures. There are many softwares available and a partial list of the same is shown in Table 1. Although the algorithms and approaches are different from each other, the basic procedure followed by them is almost same. Initially, the data obtained from NGS platforms are in the form of fluorescence signals, which are converted into short reads by base calling algorithms that are specific to the platforms (Nielsen *et al.* 2011). Additionally, the quality score for each base call is calculated to get the consistency of each base. Various methods used for quality scores calculation are again platform dependent. Finally, the output will be saved in standard FASTQ format. Many a time, a reference sequence (genome or transcriptome) is needed against which the reads will be aligned, which can be obtained from public domains like GOLD. The level of expression of genes can then be calculated by counting all the mapped reads against reference genome which is done by indexing. Currently, most algorithms either use hash look up tables, or Burrow Wheeler transformations for indexing. While hash table based tools show a high sensitivity, Burrow Wheeler methods are much faster (Mutz *et al.* 2012). Also, few statistical approaches are followed to determine differential gene expression. Quantification of reads mapping to a particular gene can be done by using a window of a definite size. By moving the window across the genome will generate the gene expression profile and expression score, which need to be normalized to avoid inherent bias during quantification. The genes, which are differentially expressed, are detected by various computational tools using normalized gene expression scores and statistical tests, *i.e.*, either parametric or non-parametric tests. The parametric methods use probability distributions such as Binomial, Poisson, Normal whereas the Non-parametric ones model the noise distribution based on actual data (Li and Tibshirani 2011). It was demonstrated that non-parametric algorithms show a lower dependency on sequencing depth and consequently achieve more robust results (Tarazona *et al.* 2011). CLC Genomics workbench Version 5.0, efficient software meant for NGS data analysis, uses two types of tests to study differential expression of gene, i.e., Gaussian test and test based on proportion. Besides transcriptome analysis, NGS is also used to discover novel nuclear RNAs (ncRNAs) and their expression (MacLean *et al.* 2009). piRNA is one of the small ncRNAs, which is bound to mammalian piwi protein and have role in germ cell development. This was discovered using Roche 454 sequencing. Morin *et al.* (2008) used this approach to study miRNA expression in human embryonic stem cell.

Despite the recent and rapid acceptance of second generation sequencing technologies, a new generation of single-molecule sequencing (SMS) technology has emerged, which is also referred as third generation sequencing. Recently, Pacific BioScience has released a third generation sequencing machine named PACBIO RS that incorporates novel, single molecule sequencing techniques and advanced analytics to reveal true biology in real time. Although these machines currently read 3 bases per second, it aims to produce entire human genomes in less than 3 min by 2013. It has also promised to deliver longer read lengths (Nature NEWS 2009). A comprehensive list of NGS tools are listed in Table 1.

EST analysis: The assembled contigs or transcripts obtained from NGS are generally annotated using EST databases. The ESTs are vital from gene identification view point. Thus various EST databases and libraries are developed for the purpose. These ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene and used as a marker for gene identification and to locate the gene in the whole genome. dbEST is the database developed and managed by NCBI which contain EST data of all most all organisms. Besides, there are many specialized databases for farm animal and fishes which contain EST information. Table 2 contains a comprehensive list of EST databases of farm animals and fish species.

Vaccine designing: Vaccination also referred as immunization is a tried, tested and most effective method among all the medical interventions and it assists in continual fight against disease in humans and animals. The process of vaccination although started 200 years back with the development of vaccine against small pox by Edward Jenner, there is still a long way to go for vaccine development to combat and eradicate many fatal microbial diseases in livestock and fish species. Although the conventional method of vaccine development is successful in many cases but it takes substantial time and failed to provide a solution for those bacteria and parasites that do not have obvious immune-dominant protective antigens (Raju and Rao, 2010). To overcome some of the difficulties, an *in silico* approach was proposed, called as reverse vaccinology or *in silico* vaccine designing.

Reverse vaccinology is one of the recent approaches in which vaccine targets were predicted by analyzing the microbial genome through computational tools (Sette and Rappuoli 2010). Wet lab experiments are then performed in a later stage on few selected targets thereby helping in reduction of cost and time consumption. Pizza *et al.* (2000) applied the reverse vaccinology approach to develop vaccine against *Neisseria meningitides*, during which 350 novel

Table 2. List of EST databases of farm animals, fishes and other related species

Name	Hosted by	URL	Contain
dbEST	NCBI	http://www.ncbi.nlm.nih.gov/dbEST/	Collection of short single-read transcript sequences from GenBank
CleanEST	KOBIC	http://cleanest.kobic.re.kr/	A novel database server that classified dbEST libraries and removes contaminants
ChickEST Database	BBSRC	http://www.chick.manchester.ac.uk/	Provides access to 339,314 *Gallus gallus* ESTs which were sequenced from 64 cDNA libraries generated from 21 different embryonic and adult tissues (Boardman *et al.* 2002)
UniGene	NCBI	http://www.ncbi.nlm.nih.gov/unigene	Cluster EST sequences with traditional gene sequences
TiGER	Wilmer Eye Institute, Johns Hopkins University	http://bioinfo.wilmer.jhu.edu/tiger/	tissue-specific gene expression profiles or expressed sequence tag (EST) data
GoSh		http://www.itb.cnr.it/gosh/	The GoSh DB is a collection of 58,990 *Capra hircus* and *Ovis aries* EST sequences
PEDE	Swine Genome Sequencing Consortium	http://www.piggenome.org/	Contain Pig EST data and also having the explore tool
Arabian Camel Genome	KACST	http://camel.kacst.edu.sa/	70,000 EST sequences from the camel cDNA libraries
Songbird EST	Funded by NIH	http://titan.biotec.uiuc.edu/cgi-bin/ESTWebsite/estima_start?seq Set=songbird	58, 211 ESTs representing RNAs expressed in the zebra finch brain

vaccine candidates were predicted and expressed in *E.coli*, out of which 28 were found to elicit protective immunity. The whole process of identifying a vaccine candidate through reverse vaccinology can take approximately 18 months whereas it takes 40 years by conventional methods (Rinaudo *et al.* 2009). Countable number of databases, web servers and software are available for vaccine target prediction in human, rat, mouse etc. (Tomar and De 2010), but in case of livestock species such availability is almost negligible due to their genetic diversity, minimum knowledge in cell surface markers and immune modulators, route of delivery of vaccine and environmental factors (Nandedkar 2009). However, now-a-days epitope based vaccine designing has become a prime interest in the field of modern computational biology, in which candidate epitopes were identified and are expected to bind with the alleles of major histocompatibility complex proteins to trigger the immunity in the organism against specific pathogens. Gamberini *et al.* (2005) used an *in silico* approach to identify 4 unique proteins (potential vaccine targets) in *Leptospira interrogans*, which cause leptospirosis disease in cattle and humans. Further, Gan *et al.* (2010) used a similar approach and identified a tegumental membrane protein enolase of *Echinococcus granulosus* as a vaccine candidate against cystic echinococciasis (CE). Guerrero *et al.* (2012) reviewed various approaches followed in designing vaccines against cattle tick.

Nanovaccine is another fresh target specific approach of vaccine designing in which nanomaterials are delivered in the form of microspheres, nano beads or micro-nanoprojections, which directly target the site in the body where the disease or infection is originated, whereas traditional vaccines affect all parts of the body (Nandedkar 2009, Sekhon and Saluja 2011). Another benefit of nanovaccine over conventional vaccines is that it induces both humoral and cell mediated immune response. Significant advantages of these biodegradable polymers are— longer safety, proven biocompatibility and control time and rate of polymer degradation and antigen release (Kersten and Hirshberg 2004). This approach was successfully applied in smallpox, influenza, tetanus and HIV and was safer than the present live-vaccinia virus vaccine because it uses nano-emulsion-inactivated vaccinia virus (Bielinska *et al.* 2008). Cervarix, an HPV vaccine approved

in United States during October 2009, is the first vaccine containing the adjuvant monophosphoryl lipid A (MPL) to be licensed by the FDA (FDA 2009). This method was also followed in livestock and fish species and many are in developmental and experimental stage. An extensive implementation of this approach in livestock and fish can improve health management and productivity with less effort, time and cost.

Drug discovery: Diseases are greatest threat to livestock and fish health and thus most of the times higher priority is given. The livestock diseases are mainly due to microbes, parasites, climate change, nutritional deficiency etc. There are various measures to control or eradicate a disease. One such measure is the use of drug or medicine. Drug may be a natural or synthetic substance which when taken have medicinal or physiological effects on body. Traditional approach of drug development was based on a screening approach and trial-and-error method, as nobody knew which compound or approach could serve as a drug or therapy (Khan *et al.* 2011). The major shortcomings of this whole approach were (i) time consuming, (ii) laborious, (iii) high attrition rates with failures attributed to poor pharmacokinetics (39%), (iv) lack of efficacy (30%), (v) animal toxicity (11%), (vi) adverse effects in humans (10%), and (vii) various commercial and miscellaneous factors (van de Waterbeemd and Gifford 2003). In 2001 Pharmaceutical Research and Manufacturers of America (PhRMA) estimated the cost at US$802 million over a period of 11 years, i.e., from the initial research stage to the successful marketing of a new drug (Yuliana *et al.* 2013). Only 1 out of 12 drugs entering clinical trials becomes a new drug. Hence, intervention of the computational approaches is essential for screening the drug computationally thereby minimizing the cost and time required in the process of drug discovery.

In silico drug designing also referred as computer-aided drug designing (CADD) or rational drug designing is the process of finding new drugs based on the knowledge of biological target. The important steps followed during this process are target identification (DNA or protein in body), ligand selection and validation, *in silico* ADMET (absorption, distribution, metabolism, excretion and toxicity) prediction, docking of target and ligand and *in silico* toxicity prediction. Collar (2010) developed and validated a novel computational approach to identify the exact pharmacophore for the external binding site of the aminopurine P2 transporter of *Trypanosoma brucei,* thereby inhibiting its effect, which cause African sleeping sickness in humans and nagana in livestock. However, very little effort has been made in the field of *in silico* drug discovery in livestock and fish diseases. Thus, there is a need to implement this technique for better management of livestock and fish species.

Motif discovery: Motifs are biologically significant fragments of nucleotide or peptide sequences in a specific pattern. Indeed, these motifs control the expression of genes. But direct experimental determination of these motifs is not practical or efficient in any biological systems (Conlon *et al.* 2003). Broadly, motifs are classified into sequence, structure and network motifs. Sequence motifs are nucleotide or peptide sequence patterns, believed to have a biological significance. Sequence motifs explain and recognize the specific characteristics of DNA, RNA and protein sequences such as TF binding sites, splice junctions and protein-protein interaction sites. Structural motifs are the super secondary structures in DNA or protein. DNA structural motifs occur in unimolecular folded DNAs as well as bimolecular hybridization. In proteins; the structural motifs are formed by 3-dimensional arrangements of continuous or discontinuous amino acids. The structural motifs having a contribution to an independently folded functional domain are called as functional motifs. The network motifs are the connectivity-patterns (sub-graphs), which occur very commonly in specific types of gene and protein interaction networks. Each type of network seems to display its own set of characteristic motifs. Alon (2006) presented the concept of network motifs for the first time when such motifs were discovered in the gene regulation (transcription) network of *E. coli* (Shen-Orr *et al.* 2002). The purpose of motif discovery is to discover patterns in bio-polymeric (nucleotide or peptide) sequences (Bailey 2007) and connectivity-patterns in gene or protein interaction networks to understand and unravel the structure and function of the bio-molecules represented by these motifs. There is countable number of databases and tools available in public domain for motifs. However, for a brief overview about the algorithms and approaches followed for motif discovery along with the databases and tools for motif search, one can refer to Sahu *et al.* (2012). Discovery of regulatory motifs in different biological activities of livestock and fisheries is a subject of interest in present day computational biology. Meher *et al.* (2014) have successfully applied Gibbs sampling in the identification of transcription factor binding site (TFBS) and discussed the parameters, model and procedures using the sequence data of *Arabidopsis thaliana*. However, this technique can also be applied for the identification of TFBS using the nucleotide sequences of livestock species.

Abiotic stress tolerance: Understanding the molecular mechanism of life cycles of livestock and fish species and identification of the corresponding genes or proteins and their successful intervention helps in the development of stress tolerant breeds. For example, anoxia tolerance is found across species and studies on genetic relationships by comparing amino acid sequences or nucleotide sequences can help find the key residues responsible for it. An *in silico* approach was proposed by Sahu *et al.* (2011) to identify probable residues responsible for anoxia tolerance. The results from their study revealed that arginine is one of the key residues involved in anoxia tolerance across species. Rao *et al.* (2014) followed a similar approach to identify the key residues of super oxide dismutase (SOD), a metallo enzyme, which neutralize ROS (reactive oxygen species) by converting it to hydrogen peroxide. It is well documented

that excess ROS generated during the photosynthesis, respiration etc., cause salt and oxidative stress in organisms.

Besides different stresses that the livestock receives from its surrounding, a number of other factors like, heavy metal, mycotoxin, plant toxin, antibiotics, radionuclides, reduce the reproductive capacity in animal and fish species and adversely affect the livestock health. One of them is the consumption of feed and fodder on which high toxic pesticides are directly sprayed. A survey indicated that 21% of feeds in the United Kingdom contain pesticide residues (D'Mello 2004). Likewise, industrial waste disposal in water bodies make it toxic and subsequently hazardous to livestock health. These toxic compounds can behave like mutagens, which may cause a change in the genetic material of a cell, or by teratogens that can cause birth defects in an embryo. So, better understanding of the toxic compounds and their mode of action at molecular level can help find the genes or proteins that are getting changed. Further, new peptide drug molecules can be designed with the same architecture but without any side effect to inhibit the binding of toxic compound thereby preventing the subsequent loss.

Computational epigenetics: Epigenetics is the study of heritable changes in gene expression and other genomic functions without altering the underlying DNA sequence (Richards 2006). It refers to a gene activity-state that may be stable over long period of time, persist through many cell divisions, or even be inherited through several generations, all without any change to the primary DNA sequence (Probst *et al.* 2009). The major epigenetic processes, viz. DNA methylation, histone modification and RNA interference are generally responsible for the gene expression and regulation. DNA methylation occurs mostly at CpG sites, at which a methyl group is added to 5' position of the cytosine to convert itto 5-methylcytosine by DNA methyltransferase (DNMT) enzymes. These modified cytosines although behave like the regular one, but methylated sites are found to be transcriptionally less active and many time lead to gene silencing. The post-translational modifications of proteins are usually governed by the histones. The histone proteins are the primary component of chromatin, which acts as a spool around which DNA can wind. Thus, histone modification occurs either by acetylation or methylation during which acetyl or methyl group is added to lysine residues of histone protein respectively. Similarly, the RNA interference (RNAi) is another phenomena involved in the regulation of gene expression. Here, the double-stranded RNA (dsRNA) is processed into short interfering RNAs (siRNAs) that regulate the gene expression. All interfering RNAs that act through RNA interference (RNAi) pathways silence the gene expression either at the transcriptional or post-transcriptional level. A number of non coding RNAs (NcRNAs) also regulate the gene expression. They are classified into different group like microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), small interfering RNAs (siRNAs), and long non-coding RNAs (lncRNAs), promoter-associated RNAs (PARs) and enhancer RNAs (eRNAs). However, the most common RNAi agents are siRNA, miRNA and shRNA that have both natural and artificial importance in controlling the gene expression. It was revealed that miRNAs regulate the expression of other ncRNAs, suggesting that small RNAs exhibit a significant impact upon transcriptomic networks (Rossi *et al.* 2008). These small RNAs played a role in DNA methylation and chromatin modification. They are also associated with regulation of protein expression and many diseases. Hence, it can be used as a marker for early diagnosis and drug development.

The study of epigenetics had gained considerable importance in genetic studies of human diseases. Ortega *et al.* (2010) designed short hairpin RNAs (shRNAs) against HIV-1 by considering the variability observed in naïve and drug-resistant isolates available at public databases using a bioperl-based algorithm. Recently, Singh *et al.* (2012) developed a user-friendly command-line user interface (CUI) based software, shRNAPred(1.0), to predict shRNA-like regions from a large set of nucleotide sequences. The software was developed using PERL Version 5.12.5 taking into account the parameters such as stem and loop length combinations, specific loop sequence, GC content, melting temperature, position specific nucleotides and low complexity filter.

Application of epigenetics in livestock has helped to find missing heritability and causality of complex traits. A recent study on epigenetics in cattle revealed that a number of dietary components can cause epigenetics as butyrate induces histone modification followed by changes in biological processes in bovine cell (Li *et al.* 2009). The above fact indicates the importance of epigenetics to better understand the role of dietary components in epigenetic change. This information will definitely assist in rising farm animals under favorable circumstances and minimizes the epigenetic modifications.

Nijland *et al.* (2008) revealed that the environment can force the pattern of methylation in human as well as in sheep. Few methods are though available to identify or locate DNA methylation at CpG site, second generation sequencing has made it more feasible. This knowledge can help animal breeders to select the genotypes which are less susceptible to unfavorable methylation. Besides, drug molecules can be designed for the diseases occurring due to epigenetic changes. Thus, its potential application could lead to perform breeding and management of livestock in a more efficient and sustainable manner.

Even though the application of epigenetics in livestock has created more opportunities to deal with complex traits, it has its own difficulties that need to be addressed through statistical approaches. Few statistical approaches were proposed but lot more has to be done on collecting and analyzing genome wide epigenetic information (Slatkin 2009, Tal *et al.* 2010, Godfrey *et al.* 2011, Takeshima *et al.* 2011). Epigenomics in livestock is at its infancy and there is still a long way to realize the relevance of epigenetics with respect to complex traits. Efficient implementation of

Table 3. A comprehensive list of epigenetic databases / tools

Name of the database	URL	Contains	Reference
MethDB	http://www.methdb.de/	Methylation pattern, profile and methylation content data	Amoreira *et al.* (2003)
PubMeth	http://www.pubmeth.org/	Cancer methylation data from literature	Ongenaert *et al.* (2008)
ChromDB	http://www.chromdb.org/		Gendler *et al.* (2008)
HIstome	http://www.actrec.gov.in/histome/index.php	Histone proteins and histone modifying enzymes	Khare *et al.* (2011)
CREMOFAC	http://www.jncasr.ac.in/cremofac/	Chromatin remodeling factors from different organism	Agrawal *et al.* (2006)
MethPrimerDB	http://medgen.ugent.be/methprimerdb	PCR primers for DNA methylation analysis	Pattyn *et al.* (2006)

the knowledge gained from epigenetics in genome selection and breeding can assist in sustainable production of livestock. Many databases and tools were developed to store and analyze the epigenetic information and some of them are listed in Table 3.

Demand for livestock is likely to increase in coming years because of ever increasing global population. Simultaneously, cost for production is also likely to increase exponentially due to competition for land, water, nutrition etc. In order to meet the challenges, livestock and fish production, in turn, animal protein production needs to be increased. Thus, an integrated approach based on various disciplines like, genomics, computer science and statistics is necessary to deal with health management and productivity of livestock. Work needs to be initiated to demonstrate how the gene regulation is far more complex than previously believed in animal and fish species. Besides, certain facts like, how much percent age of non-coding DNA in animal and fish genomes is functional and transcribed with no known functions, how much per cent of non-coding DNA is involved in the regulation of the expression of coding genes, whether the regulatory sites located near and distant from the genes controlling the expression of each of such genes are to be understood using statistical and computational techniques.

As most of the economically important traits are complex and quantitative in nature, i.e., regulated by more than one gene and environmental factors, accurate association of all such important traits using high-throughput marker data has to be studied in animal and fish species. Advantage of whole genome selection (WGS) need to be explored as it not only puts greater emphasis on the regions with larger effects but also considers remainder of the genome, which helps to get the SNPs that have lesser and cumulative effect on the trait of interest. Such selection was successfully applied for the selection of young bulls and need to be applied in other farm animals too. New forms of computational models such as process calculi and constraint-based modeling may be tried to model biological processes. Besides the mathematical modeling involving the machine learning approaches (MLAs) may be applied on livestock and fish genomic resources to mine meaningful biological information.

Now, it is the need of the hour to develop genome explorers / browsers, genomic databases and information systems on indigenous species with user interactive facility for deriving the underlying biological information and knowledge in animal and fish species. More sequencing projects need to be undertaken to get dense marker data at whole genome level so that the association of SNPs / CNVs with traits like diseases or stress tolerance can be studied. Also, integration of information from literature, using techniques of information extraction and text mining is also necessary in the present context. In addition, online interfaces may have to be developed for sharing data and models. Also, various approaches to database integration and software interoperability via loose coupling of software, commercial suites are to be developed to represent the biological models and processes in animal species. Emerging approaches like systems biology, a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems, using a more holistic perspective approach, may be applied to understand the complex biological phenomena involved in performance of economically important traits.

ACKNOWLEDGEMENTS

Authors acknowledge World Bank Funded National Agricultural Innovation Project (NAIP), NAIP/Comp-4/C4/C-30033/2008-09 and ICAR Grants 30(68)/2009/Bio Informatics/NAIP/ O&M.

REFERENCES

Adams J. 2008. Transcriptome: connecting the genome to gene function. *Nature Education* **1**(1).

Agrawal S, Chetan K and Rao M R S. 2006. CREMOFAC-a database of chromatin remodeling factors. *Bioinformatics* **22** (23): 2940–44.

Alon U. 2006. *An Introduction to Systems Biology: Design Principles of Biological Circuits.* Boca Raton: CRC, FL, USA.

Amoreira C, Hindermann W and Grunau C. 2003. An improved version of the DNA methylation database (MethDB). *Nucleic Acids Research* **31**: 75–77.

Bailey T L. 2007. Discovering sequence motifs. *Methods in Molecular Biology* **395**: 271–92.

Ban H J, Heo J Y, Oh K S and Park K J .2010. Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC Genetics* **11**: 26.

Bhatia V K and Jayasankar J. 1996a. On use of bootstrapping for estimation of confidence interval of heritability. *Journal of the Indian Society of Agricultural Statistics* **48**: 217–23.

Bhatia V K and Jayasankar J. 1996b. A bootstrap look on the precision of Heritability estimators in unbalanced data. *Journal of the Indian Society of Agricultural Statistics* **49**: 289–306.

Bhatia V K and Malhotra P K. 1995a. On some aspects of estimation and comparison of retention times of dairy cattle. *Indian Journal of Dairy Science* **48** (2): 103–09.

Bhatia V K and Malhotra P K. 1995b. Relationship between retention times and explanatory variables in dairy cattle. *Indian Journal of Dairy Science* **48**(6): 438–41.

Bhatia V K, Jayasankar J and Wahi S D. 1994. Use of bootstrap technique for variance estimation of heritability estimators. *Annals of Agricultural Research* **15**(4): 476–80.

Bhatia V K, Narain P and Malhotra P K. 1992a. Comparison of culling patterns of different categories of dairy animals - Non-parametric approach. *Indian Journal of Dairy Science* **45**(3): 114–18.

Bhatia V K, Narain P and Malhotra P K. 1992b. Genetic parameters of stayability of different categories of dairy animals. *Journal of the Indian Society of Agricultural Statistics* **44 (2)**: 122–137.

Bielinska A U, Chepurnov A A , Landers J J, Janczak K W and Chepurnov C S .2008. A novel, killed-virus nasal vaccinia virus vaccine. *Clinical Vaccine Immunology* **15**: 348–58.

Breiman L.2001. Random Forests. *Machine learning* **45**: 5–32.

Caraviello D Z, Weigel K A and Craven M .2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. *Journal of Dairy Science* **89**: 4703–22.

Chand L, Wahi S D and Bhatia V K. 1997. Relationship among different genetic groups for growth traits in goats. *Indian Journal of Dairy Science* **49**(8): 511–16.

Chand L, Wahi S D and Parkash Lal. 1998a. Estimation of heritability and genetic correlation in the parameter space. *Indian Journal of Animal Genetic and Breeding* **20**(1): 37–40.

Cloonan N, Forrest A R, Kolle G, Gardiner B B A and Faulkner G J. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**: 613–19.

Collar C J. 2010. Rational Drug Design for Neglected Diseases: Implementation of Computational Methods to Construct Predictive Devices and Examine Mechanisms. Chemistry Dissertations Paper 48. http://digitalarchive.gsu.edu/chemistry_diss/48.

Colombani C, Legarra A, Fritz S, Guillaume F and Croiseau P . 2012. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and Bayes C P methods for genomic selection in French Holstein and Montbéliarde breeds. *Journal of Dairy Science* **96**: 575–91.

Conlon E M, Liu X S, Lieb J D and Liu J S. 2003. Integrating regulatory motif discovery and genome-wide expression analysis. *The Proceedings of the National Academy of Sciences* **100**: 3339–44.

Daetwyler H D, Villanueva B and Woolliams J A. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE **3**(10):e3395, doi:10.1371/journal.pone.0003395.

Dale J M, Popescu L, and Karp P D. 2010. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* **11**: 15. doi:10.1186/1471-2105-11-15.

Devendra C. 2010. Small farms in Asia: Revitalising agricultural production, food security and rural prosperity. *Academy of Sciences Malaysia*, Kuala Lumpur, Malaysia.

D'Mello J P F. 2004. Contaminants and toxins in animal feeds. *Assessing Quality and Safety of Animal Feeds* (Eds.) J Samuel, pp 107-128, FAO, Rome.

Doan R, Cohen N, Harrington J, Veazy K, Juras R, Cothran G, McCue M E, Skow L and Dindot S V. 2012. Identification of copy number variants in horses. *Genome Research* **22**: 899–907.

Dongre V B, Gandhi R S, Singh A and Ruhil A P. 2012. Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in Sahiwal cattle. *Livestock Science* **147**: 192–97.

Druet T and Georges M. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**: 789–98.

Dube S, Qin J, Ramakrishnan R.2008. Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device. *PLoS ONE* **3**(8): e2876. doi:10.1371/journal.pone.0002876.

Eichler E E. 2008. Copy Number Variation and Human Disease. *Nature Education* **1**(3).

FAO 2010. The International Technical Conference, Current Status and Options for Livestock Biotechnologies in Developing Countries, Guadalajara, Mexico, FAO.

FAO 2011. World Livestock 2011 – Livestock in food security. Rome, FAO.

FAO 2012a. Trends in the livestock sector, Part-3-Feeding the World, FAO Statistical Year Book 2012, Rome, FAO.

FAO 2012b. The State of World Fisheries and Aquaculture, FAO Fisheries and Aquaculture Department, Rome, FAO.

FAOSTAT 2010. FAO Statistical Databases. www.faostat.fao.org/site/452/default.aspx

FDA 2009. FDA approves new vaccine for prevention of cervical cancer. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm187048.htm.

Fernandez C, Soria E, Martin J D and Serrano A J. 2006. Neural networks for animal science applications: two case studies. *Expert Systems with Applications* **31**: 444–50.

Fontanesi L, Martelli P L, Beretti F, Riggio V and Colombo M. 2010. An initial comparative map of copy number variations in the goat *(Capra hircus)* genome. *BMC Genomics* **11**: 639.

Gamberini M, Gómez R M, Atzingen M V, Martins E A and Vasconcellos S A. 2005. Whole-genome analysis of Leptospira interrogans to identify potential vaccine candidates against leptospirosis. *FEMS Microbiology Letters* **244**: 305–13.

Gan W, Zhao G, Xu H, Wu W, Du W. 2010. Reverse vaccinology approach identify an Echinococcus granulosus tegumental membrane protein enolase as vaccine candidate. *Parasitology Research* **106**: 873–82.

Gendler K, Paulsen T and Napoli C. 2008. ChromDB: the

chromatin database. *Nucleic Acids Research* **36**: D298–D302.

Glazov E A, Cottee P A, Barris W C, Moore R J, Dalrymple B P 2008. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome Research* **18**: 957–64.

Goddard M E and Hayes B J. 2007. Genomic selection, *Journal of Animal Breeding and Genetics*. **124**: 323–30.

Goddard ME, Wray NR, Verbyla K and Visscher PM. 2009. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Statistical Science* **24**(4): 517–29.

Godfrey K M, Sheppard A, Gluckman P D, Lillycrop K A , Burdge G C. 2011. Epigenetic gene promoter methylation at birth is associated with child's later adiposity. *Diabetes* **60**: 1528–34.

Guerrero F D, Miller R J and Pérez de León A A . 2012. Cattle tick vaccines: Many candidate antigens, but will a commercially viable product emerge. *International Journal for Parasitology* **42**: 421–27.

Hayes B J, Bowman P J, Chamberlain A J and Goddard ME. 2009. Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92**(2): 433–43. doi: 10.3168/jds. 2008–1646.

Hirschhorn J N and Daly M J. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Review of Genetics* **6**: 95–108.

Hossain M S, Azimi N and Skiena S. 2009. Crystallizing short-read assemblies around seeds. *BMC Bioinformatics* **10**: (Suppl. 1), S16.

Jiang L, Jiang J, Wang J, Ding X and Liu J. 2012. Genome-Wide Identification of Copy Number Variations in Chinese Holstein. *PLoS ONE*, **7**(11), e48732.

Jostins L and Barret C J. 2011. Genetic risk prediction in complex disease. *Human Molecular Genetics* **20**(2): R182–R188.

Katayama S, Tomaru Y, Kasukawa T, Waki K and Nakanishi M. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309**: 1564–66.

Kersten G and Hirschberg H. 2004. Antigen delivery systems. *Expert Review of Vaccines* **3**(4): 453–62.

Khan F , Yadav E K , Maurya A and Shrivastav S K. 2011. Modern methods and web resources in drug design and discovery. *Letters in Drug Design and Discovery* **8**(5): 469–90.

Khare S P, Habib F, Sharma R, Gadewal N, Gupta S and Galande S. 2012. HIstome-a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Research* **40**: D337–D342.

Kiran P S, Bhatia V K and Rao A R. 2003. Sampling distributions of heritability. *Journal of the Indian Society of Agricultural Statistics* **56**: 294–301.

Kiran P S, Bhatia V K and Rao A R. 2004. Robust method of estimation of heritability. *Journal of the Indian Society of Agricultural Statistics* **57**: 116–28.

Kiran PS, Bhatia V K and Paul A K. 2007. Estimation of heritability by modified ANOVA method in different situations. Journal of Animal Science **77**(11): 1184–87.

Kumar S, Rao A R and Bhatia V K. 2004. Bayesian estimation of heritability in animal breeding experiments under 2-way nested classification. *Journal of the Indian Society of Agricultural Statistics* **58**(3): 352–62.

Li C, Elsasser T H and Li R W. 2009. Epigenetic regulation in bovine cells: nutrient-induced modulation of gene expression and cellular functions. *Animal Genetics* (Eds). L J Rechi, Chapter 4, Nova Science Publishers, New York.

Li J and Tibshirani R. 2011. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*. doi:10.1177/0962280211428386.

Lillehammer M , Meuwissen T H E and Sonesson A K .2011. A comparison of dairy cattle breeding designs that use genomic selection. *Journal of Dairy Science* **94**: 493–500.

MacLean D, Jones J D and Studholme D J. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology* **7**: 287–96.

Mardis E R. 2008 . Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Martiskainen P , Jarvinen M, Skon J, Tiirikainen J , Kolehmainen M. 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science* **119**: 32–38.

Masoum S, Malabat C, Jalali-Heravi M, Guillou C and Rezzi S. 2007. Application of support vector machines to 1H NMR data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins. *Analytical and Bioanalytical Chemistry* **387** (4):1499–1510.

Meher P K, Sahu T K, Rao A R and Wahi S D. 2014. Application of Gibbs sampling methodology for identification of transcription factor binding sites in MADS box family genes in *Arabidopsis thaliana*, *Indian Journal of Genetics and Plant Breeding* **74** (1): (In press).

Meher P K, Rao A R, Wahi S D and Jaggi N. 2010. Detection of multivariate outliers in breeding data. *International Journal of Statistics & Systems* **5**(4): 527–35.

Meuwissen T H , Solberg T R , Shepherd R, Woolliams J A.2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution* **41**: 2 doi:10.1186/1297-9686-41-2.

Meuwissen T H E, Hayes B J and Goddard M E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–29.

Morin R D, O'Connor M D, Griffith M, Kuchenbaur F and Delaney A. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research* **18**(4): 610–21.

Mortazavi A, Williams B A, McCue K, Schaeffer L , Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–28.

Mutz K O, Heilkenbrinker A and Lönne M. 2012. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* **24**: 1–9.

Nandedkar T D. 2009. Nano vaccines: recent developments in vaccination. *Journal of Biosciences*. **34**: 995–1003.

Nature NEWS 2009. Genome sequencing: the third generation. http://www.nature.com/news/2009/090206/pdf/news.2009.86.pdf.

NHGRI 2011. Approved Sequencing Targets. http://www.genome.gov/10002154

Nielsen R, Paul J S, Albrechtsen A and Song Y S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**: 443–51.

Nijland M J, Ford S P and Nathanielsz P W. 2008. Prenatal origins of adult disease. *Current Opinion in Obsterics and Gynecology* **20**: 132–38.

Noushin R F . 2012. Computational tools for CNV detection using probe-level analysis of Affymetrix SNP arrays : application to the study of CNVs in follicular lymphoma. https://circle.ubc.ca/handle/2429/43034

Noyes H, Brass A, Obara I, Anderson S and Archibald A L. 2011. Genetic and expression analysis of cattle identifies candidate

genes in pathways responding to Trypanosoma congolense infection. *PNAS* **108**(22): 9304–09.

Ongenaert M , Neste LV, De Meyer T ,Menschaert G, Bekeart S. 2008. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research* **36**: D842–D846.

Ortega M, Restrepo S and Rodríguez L M. 2010. An RNAi in silico approach to find an optimal shRNA cocktail against HIV-1. *Virology Journal* **7**: 369. doi:10.1186/1743-22X-7-369.

Oudendag D, Szlávik Z and Veen H. 2012. The use of machine learning techniques to predict farm size change – an implementation in the Dutch Dairy sector. *American Academic and Scholarly Research Journal b* **4**(5).

Pagani I, Liolios K, Jansson J, Chen I M, Smirnova T, Nosrat B, Markowitz V M and Kyrpides N C. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40**: D571- D579.

Pattyn F, Hoebeeck J, Robbrecht P. 2006. methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. *BMC Bioinformatics* **7**: 496.

Paul A K and Bhatia V K. 2000a. Heritability of stayability by path coefficient approach when herd-life is influenced by unrelated auxiliary characters. *Sankhya: The Indian Journal of Statistics, Series* B **62**(3): 463–71.

Paul A K and Bhatia V K. 2000b. Empirical comparison of different methods of estimation of heritability of stayability in dairy cattle. *Journal of the Indian Society of Agricultural Statistics* **53**(2): 196–206.

Paul A K and Bhatia V K. 2001. Modification of beta-binomial method of estimation of heritability of stayability. *Journal of the Indian Society of Agricultural Statistics* **54**(3): 385–95.

Paul A K and Bhatia V K. 2002a. Comparison of beta binomial with other methods for estimation of heritability of stayability. for different degree of unbalanced data in dairy cattle. *Indian Journal of Animal Sciences* **72**(4): 402–05.

Paul A K and Bhatia V K. 2002b. Effect of auxiliary traits on estimation of heritability of stayability. *Journal of the Indian Society of Agricultural Statistics* **55**(1): 63–69.

Pizza M, ScarlatoV, Masignani V. 2000. Identification of Vaccine Candidates against Serogroup B Meningococcus by Whole Genome Sequencing. *Science* **287**: 1816–20.

Prabhakaran V T and Rao A R. 2008. A new approach to the estimation of variance of sample heritability from full-sib analysis. *Journal of the Indian Society of Agricultural Statistics* **62** (3): 208–13.

Prabhakaran V T. 1998. Research in bio-statistics and statistical genetics in the country: A review. *Indian Journal of Animal Sciences* **68**(8): 743–54.

Probst A V, Dunleavy E and Almouzni G. 2009. Epigenetic inheritance during the cell cycle. *Nature Review on Molecular Cell Biology* **10**:192–206.

Pungpapong V, Muir W M and Li X. 2012. A Fast and efficient approach for genomic selection with high-density markers. *G3 (Bethesda)* **2**(10): 1179–84.

Raju S and Rao UM. 2010. Current development strategies for vaccines and the role of reverse vaccinology. *Journal of Pharmaceutical Research and Health Care* **2**(4): 339–46.

Rao A R and Kumar S. 2001. Bayesian estimation of heritability using Gibbs sampling for half-sib mating design. *Indian Journal of Applied Statistics* **6**: 12–26.

Rao A R, Dash M, Sahu T K, Behera B K and Mohapatra T. 2014. Detection of novel key residues of Mn SOD enzyme and its role in salinity management across species, *Journal of Genetics* (In press).

Rao A R, Wahi S D and Bhatia V K. 2009. *Statistical Applications in Breeding and Genetics. IASRI- An era of excellence 1959-2009*. IASRI publications. pp 139-59.

Recio O G and Forni S. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution* **43**: 7.

Richards E J. 2006. Inherited epigenetic variation-revisiting soft inheritance. *Nature Review of Genetics* **7**(5): 395–401.

Rinaudo C D, Telford J L, Rappuoli R and Seib K L. 2009. Vaccinology in the genome era. *J. lin. Invest.* **119**: 2515–25. doi: 10.1172/JCI38330.

Rossi S, Sevignani C, Nnadi S C. 2008. Cancer-associated genomic regions (CAGRs) and noncoding RNAs: bioinformatics and therapeutic implications. *Mammlian Genome* **19**: 526–40.

Sahu T K, Rao A R, Singh A, Behera B K and Mohapatra T. 2011. *In silico* identification of residues for anoxia tolerance across species. *Online Journal of Bioinformatics* **12**(1): 175–97.

Sahu T K, Rao A R, Vasisht S, Singh N and Singh U P. 2012. Computational Approaches, Databases and Tools for *in silico* Motif Discovery. *Interdisciplinary Sciences: Computational Life Sciences* **4**(4): 239–55.

Sariaka, Wahi S D and Rao A R. 2006a. Effect of outliers on the estimates of genetic correlation. *Indian Journal of Animal Genetics and Breeding* **27** (1, 2): 17–21.

Sariaka, Wahi S D and Rao A R. 2006b. On the robustness of estimates of genetic correlation. *Indian Journal of Animal Sciences* **76**(9): 759–63.

Schaeffer L R. 2006. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding Genetics* **123**(4): 218–23.

Sekhon B S and Saluja V. 2011. Nanovaccines-an overview. *International Journal of Pharmaceutical Frontier Research* **1**(1): 101–09.

Sette A and Rappuoli R. 2010. Reverse vaccinology: Developing vaccines in the era of genomics. *Immunity* **33**(4): 530–41.

Shahinfar S, Yeganeh H M, Lucas C *et al.* 2012. Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. *Computational and Mathematical Methods in Medicine* **2012**: doi:10.1155/2012/127130.

Shen-Orr S S, Milo R and Mangan S. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31**: 64–68.

Singh N, Sahu T K, Rao A R and Mohapatra T. 2012. shRNAPred (version 1.0): An open source and standalone software for short hairpin RNA (shRNA) prediction. *Bioinformation* **8**(13): 629–33.

Singh O N, Rao A R, Wahi S D and Singh V P. 2006. Robustness of bootstrap estimates of variance of heritability of master samples in half-sib analysis". *Indian Journal of Animal Genetics and Breeding* **27**(1, 2): 6–11.

Slatkin M. 2009. Epigenetic inheritance and the missing heritability problem. *Genetics* **182**: 845–50.

Smith C. 1967. Improvement of metric traits through specific genetic loci. *Animal Production* **9**: 349–58. doi:10.1017/S0003356100038642.

Soller M and Beckmann J S. 1983. Genetic polymorphism in varietal identification and genetic improvement. *Theoretical and Applied Genetics* **67**(1): 25–33

Takeshima E, Tomimori K, Kawakami H. 2011. Retraction: NF-*K*B activation by *Helicobacter pylori* requires Akt-mediated

phosphorylation of p65. *BMC Microbiology* **11**: 128. doi:10.1186/1471-2180-11-128.

Tal O, Kisdi K and Jablonka E. 2010. Epigenetic contribution to covariance between relatives. *Genetics* **184**: 1037–50.

Tarazona S, Garcia-Alcalde F and Dopazo J. 2011. Differential expression in RNA-seq: a matter of depth. *Genome Research* **21**: 2213–23.

Thomas N, Matejovicova L and Srikusalanuku W. 2012. Directional migration of recirculating lymphocytes through lymph nodes via random walks. *PLoS ONE* **7**(9): e45262, doi:10.1371/journal.pone.0045262.

Thornton P K, Jones P G and Owiyo T. 2006. *Mapping climate vulnerability and poverty in Africa.* International Livestock Research Institute, Nairobi, Kenya.

Thornton PK. 2010. Livestock production: recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Science* **365**: 2853– 67.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)* **58**: 67–288.

Tomar N and De R K. 2010. Immunoinformatics: an integrated scenario. *Immunology* **131**:153–68.

van de Waterbeemd H and Gifford E. 2003. ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery* **2**(3): 192–204.

Vazquez A I, de los Campos G, Klimentidis Y C, Rosa G J, Gianola D, Yi N and Allison D B. 2012. A Comprehensive Genetic approach for improving prediction of skin cancer risk in humans. *Genetics* **192**: 1493–1502.

Wahi S D and Bhatia V K. 1995. Use of bootstrap methods in comparing the performance of linear discriminant function. *Journal of the Indian Society of Agricultural Statistics* **47**(1): 12–20.

Wahi S D and Rao A R. 2011. Some investigations on sampling variance of genetic correlation. *The ICFAI Journal of Genetics & Evolution* **4**(1): 27–44.

Wahi S D, Bhatia V K and Rao A R. 2006 "Study on statistical properties of genetics correlation using bootstrap technique". *Indian Journal of Animal Sciences* **76**(9): 755–58.

Wahi S D, Chand L and Bhatia V K. 2004. A growth pattern study in crosses and pure Indian breeds of goats. *Indian Journal of Animal Sciences* **74**(9): 955–58.

Wan L and Bao W. 2009. Research and Application of Animal Disease Intelligent Diagnosis based on Support Vector Machine. Proceedings of the 2009 International Conference on Computational Intelligence and Security **2**: 66–70.

Wang J, Jiang J, Fu W, Jiang L, Ding X, Liu J F and Zhang Q. 2012a. A genome-wide detection of copy number variations using SNP genotyping arrays in swine. *BMC Genomics* **13**: 273.

Wang Y, Gu X, Feng C, Song C, Hu X and Li N. 2012b. A genome-wide survey of copy number variation regions in various chicken breeds by array comparative genomic hybridization method. *Animal Genetics* **43**(3): 282–89.

Witten I H and Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, (Second Edition). San Francisco: Morgan Kaufmann.

World Bank 2009. Minding the stock: Bringing public policy to bear on livestock sector development. Report No. 44010-GLB. The World Bank, Washington D.C., USA.

Wray N R, Goddard M E and Visscher P M. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**: 1520–28.

Yuliana D, Mursalin and Najib A. 2013. *In silico* screening of chemical compounds from sweet flag (*Aracus calamus* L) as á-Glucosidase inhibitor. *International Research Journal of Pharmacy* **4**(3): 110-12.

Zhang H, Wang Z, Wang S and Li H. 2012. Progress of genome wide association study in domestic animals. *Journal of Animal Science and Biotechnology* **3**:26.