Genomic Selection: Concept, Methodologies and Application in Animal Science (Theory & Practical Session)

Chapter · January 2013

DOI: 10.13140/RG.2.1.4929.8401

CITATIONS
0

READS
1,540

2 authors:

C. S. Mukhopadhyay
Guru Angad Dev Veterinary and Animal Sciences University
122 PUBLICATIONS **314** CITATIONS

SEE PROFILE

Dinesh Kumar
Indian Agricultural Statistics Research Institute
163 PUBLICATIONS **1,807** CITATIONS

SEE PROFILE

Genomic Selection: Its Prospects and Challenges

C. S. Mukhopadhyay[+] and Dinesh Kumar [++]

[+] School of Animal Biotechnology, GADVASU, Ludhiana, Punjab, (csmscience(at)gmail(dot)com)

[++] Senior Scientist, NBAGR, Karnal, Haryana, (dineshkumarbhu(at)gmail(dot)com)

Introduction and Concept:

Molecular genetics has traversed commendable way starting from single locus marker aided selection process to genome-wide association study (GWAS) or whole genome association study (WGAS) during the last three decades. The molecular markers (single or a few taken together) aimed at screening the individuals with desirable (economic trait loci) or deleterious (disease carriers) alleles in order to select the best individuals to propagate their genome to the next generation. The advent of high throughput sequencing technologies has revolutionized the marker assisted selection process. Now, the whole genome of the individual is screened for the loci contributing to the trait of interest. In contrast to methods which specifically test one or a few genomic regions, the GWAS investigates the entire genome. The approach is therefore said to be genome-wide, non-candidate-specific in contrast to candidate-gene approach. GWA studies identify SNPs and other variants in DNA which are associated with a trait; however, it cannot specify the causal genes on their own. The availability of high-throughput, whole-genome genotyping arrays viz. Illumina's Infinium® HD DNA Analysis BeadChips have enabled the researchers to efficiently screen for associations between variations in the genome and phenotypes of interest (http://www.illumina.com/pagesnrn.ilmn?ID=89). RNA samples can be custom sequenced commercially by a number of companies, like, Illumina, Helicos, LC Sciences, Oceanridgebio, seqwright, macrogen, CoFactor Genomics. For genomic selection studies, DNA samples are sent to commercial agencies like Geneseek (Lincoln, Nebraska). GWA studies have been able to identify the quantitative trait loci (QTL) affecting many common complex diseases, demonstrating the utility of this approach for dissecting the genetic basis of polygenic traits (McCarthy *et al*., 2008). The first GWA study was conducted on patients with age-related macular degeneration, which was published in 2005 and reported two SNPs between case and control individuals (Klein *et al*., 2005).

The present discussion will cover two broad sections: (a) Genomic selection in livestock and (b) Network analysis using the genome wide association studies.

What is genomic selection?

Genomic selection refers to selection decisions based on genomic breeding values (GEBV). The GEBV are calculated as the sum of the effects of dense genetic markers, or haplotypes of these markers, across the entire genome, there by potentially capturing all the quantitative trait loci (QTL) that contribute to variation in a trait. The QTL effects, inferred from either haplotypes or individual single nucleotide polymorphism markers, are first estimated in a large reference population with phenotypic information. In subsequent generations, only marker information is required to calculate GEBV.

How GS is different from MAS/GAS?

Genomic selection (GS) is a form of marker assisted 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. This approach has become feasible due to revolution in SNP discovery method like deep sequencing and throughput SNP genotyping on DNA chip.

Marker assisted selection or marker aided selection (MAS) is a process whereby a marker (morphological, biochemical or one based on DNA/RNA variation) is used for indirect selection of a genetic determinant or determinants of a trait of interest (i.e. productivity, disease resistance, abiotic stress tolerance, and/or quality). This process is used in plant and animal breeding. When the gene sequence itself becomes an identifiable marker it's called GAS (Gene Assisted Selection).

Considerable developments in biotechnology have led animal breeders to develop more efficient selection systems to replace traditional phenotypic-pedigree-based selection systems.

Marker assisted selection (MAS) is indirect selection process where a trait of interest is selected, not based on the trait itself, but on a marker linked to it. For example if MAS is being used to select individuals with a disease, the level of disease is not quantified but rather a marker allele which is linked with disease is used to determine disease presence. The assumption is that linked allele associates with the gene and/or quantitative trait locus (QTL) of interest. MAS can be useful for traits that are difficult to measure, exhibit low heritability, and/or are expressed late in development.

Problems associated with QTL/MAS?

There are three types of markers used in MAS viz. genes, linkage equilibrium (LE), which are located very close to the gene, and linkage disequilibrium (LD), which are located farther from the gene. LD markers are easiest to find but hard to use. In selection, LE markers are markedly inferior to genes. There were no markers for low heritability traits as these require a large amount of data for estimation. Finally, benefits from commercial applications were hard to assess (Misztal, 2006).

What are the advantages of genomic selection?

Genomic selection may result in higher rates of genetic gain over traditional selection because genomic EBVs have higher reliabilities than BLUP EBVs, especially for young animals, and secondly because young animals with high genomic EBVs become attractive to be selected as parents, which reduces the generation interval (Meuwissen *et al*., 2001). The advantages of genomic selection may be highest for dairy cattle breeding programs because the generation interval in traditional progeny testing schemes is large and selection of young bulls for progeny testing is inaccurate (Schaeffer (2006)) Furthermore, thousands of bulls that have been progeny tested in the last decades are available as a reference population with very reliable phenotypes, leading to genomic EBVs with high reliabilities (VanRaden *et al*., 2009). For these reasons, the up-take of genomic selection in animal breeding in recent years has been very high. Genomic selection may decrease the rate of inbreeding because Mendelian sampling effects can be estimated more accurately, which reduces the co-selection of relatives (Daetwyler *et al*., 2007).

Genomic selection in combination with a reduced generation interval may double the rate of genetic gain while keeping the rate of inbreeding per generation constant. Young bulls will be superior to proven bulls, and the number of progeny test bulls can be greatly reduced drastically reducing the cost of a progeny testing programme.

Challenges in implementing GS programme in India

The requirements to implement genomic selection in breeding programmes are relatively simple. Generally there will be a discovery dataset where a large number of SNP have been assayed on a moderate number of animals who have phenotypes for all the relevant traits. A prediction equation that uses markers as input and predicts BV is derived from this data. There should then be a validation sample (which can be smaller than the derivation sample) where a larger number of animals are recorded for the traits and genotyped at least for the markers that are proposed to be used commercially. The prediction equation is tested to assess its accuracy on this independent sample. Then selection candidates are genotyped for the markers and the prediction equation estimated in the discovery data used to calculate GEBV, but their accuracy is assumed to be that found in the validation sample. In practice, the process may be more complex but the distinction between discovery, validation and selection candidates is still useful. For instance, it makes clear that the estimation of QTL effects can be carried out on animals that are completely separate from the selection candidates. In fact the selection candidates do not need to have phenotypes recorded at all. This technology has potential to fetch large changes in the structure of the livestock breeding programmes. The combined

discovery and validation datasets are referred as ‘reference’ population (Goddard & Hayes, 2007).

In India the challenges are double sworded. Firstly, the “phenotype gap” due to limited phenome data generation on sheep and goat, secondly, lack of integrated DNA bank. This holds true for almost all domestic animal species. The genetic infrastructure required in implementing GWAS plan and sheep has to be started from scratch. The data generation and curation with integrated DNA sample is a herculean task. The predictor and predicted reference population of more than 4000 sample size at least, is required in order to establish the gEBV in each major breed. This is going to take very long time thus gestation period for GWAS implementation in sheep and goat is no doubt going to take much longer time.

Genome Wide Association Study:

Genome-wide association study (GWAS) is a process for inspection and screening of detectable common genetic variants (single-nucleotide polymorphisms) in individuals to identify the variant(s) associated with the trait under study. The GWAS compares the DNA profiles of individuals having altered trait (viz. disease, improved production, reproduction or growth parameters) with the control ones (healthy ones or with normal or below average parameters). The DNA specimen from each of the individuals is subjected to microarray analysis for detection of specific SNPs that are more prevalent in any one group. The associated SNPs mark a region on the genome.

As of 2011, thousands of individuals have been tested, over thousand of GWAS have examined over 200 diseases and traits in human, which revealed 4,000 SNP associations. The introduction of biobanks in the western countries, to conserve the rare repository of human genetic materials, and the initiation of International HapMap project with an aim to identify the human SNPs, had transformed the GWAS from conceptual framework to a practical application in human and animal sciences.

Linkage analysis, as has been discussed earlier has proved to be very useful and was the only way to identify causal genes for monogenic traits. Howbeit, the efficacy of linkage studies is far away from acceptability for polygenic traits and economic trait loci. Thus, an alternative approach was genetic association study which opts statistical tools to determine whether an allele of a genetic variant is found more often than expected in individuals with the phenotype of interest. The major drawback of this study is the number of loci considered is far less than the actual. For instance, milk production traits in dairy cattle was estimated to be controlled by 150 QTL (Hayes *et al*. 2006), but there are even QTLs more because the power to detect these QTL was not 100% (Goddard and Hayes, 2007).

The GWAS has got several basic applications in molecular animal breeding:

1. To associate between the variations in genotypes and phenotypes to identify the causal genetic mechanism.
2. To identify QTL underlying many common, complex disease.
3. To associate a trait with a region in the genome, in order to map the clinically and/or economically important QTLs.

The traditional GWAS scheme incorporates the expression QTL (eQTL) approach, where the trait being investigated is associated with a region in the genome. The mRNA abundance is considered as a trait, i.e., the expression profile of thousands of transcript is a trait. In this regard, the eQTL can be associated and positioned with the transcript in two different ways:

1. **cis-acting eQTL**: the locus that affects the mRNA abundance (i.e. the eQTL) position overlaps the location of the affected gene (i.e. transcript). It results from a variant in the regulatory region of the gene that affects its level of abundance. The cis-eQTL has been used for identification and validation of quantitative trait genes (QTG) in candidate gene approach.
2. **trans-acting eQTL** or eQTL hotspots: the expression of multiple genes (or transcript) map to the same SNP (or eQTL). Hence, it is also termed as “eQTL hotspots”. The pivotal mechanisms behind trans-eQTL could be abundance or activity of a transcription factor or variation in a component of a signaling cascade resulting in abundance of the mRNA. The eQTL hotspots are the coordinately regulated subsets of loci resulting from differential mRNA (transcript) abundance of those loci due to variants of a unique eQTL. Such gene sets may map to eQTL trans-bands, or “eQTL hot spots,” that overlap a QTL for the clinical trait (Drake *et al*., 2006).

Candidate Gene Filtering with cis-eQTL:

The eQTL approach to identify the candidate genes associated with a specific trait is considerably challenged by identification and validation of the quantitative trait genes (QTGs) containing the causative genetic variant (DiPetrillo *et al*., 2005), due to lack of biological evidence supporting the results generated from functional pathway analysis. Experimental evidence suggests that association between a local genetic variant and phenotypic variation could establish expression signatures that can be specific to a unique QTG candidate. This approach can be applied to identify the candidate genes associated with the underlying trait (Stranger *et al*., 2005). Identification of candidate QTG warrants validation of the same in unrelated samples.

GWAS methodology:

In human, the GWA studies compares two groups (healthy versus diseased) of individuals. All individuals in each group are genotyped, using microarray, for a million of SNPs. In the simplest approach, the healthy and diseased groups are tested for the odds ratio of the allele frequencies for each of these SNPs (Clarke *et al*., 2011). The odds ratio measures the ratio between the proportions of the two groups. If the two groups are not differing significantly, the odds-ratio is one. P-value for the significance of the odds ratio is calculated using chi-squared test. One of the basic requirements of GWAS is a sufficiently large sample to increase the accuracy of detecting the SNPs associated with the trait of interest. The most important aspect is functional enrichment of the genes found in the modules in order to identify the genes governing the trait of interest. This also requires validation of the generated data with biological samples.

The GWAS has been successfully extended towards animal science, for detecting the differentially expressed genes as well as identification of the central key gene(s) underlying the trait(s) of interest, followed by construction of network. It encompasses disease tolerance/susceptibility, production vis-à-vis reproduction traits and growth traits, as well. The intermediate phenotypes (viz. enzyme concentration, fatty acid content in meat, specific casein content in milk etc) are measured (Danesh and Peppys, 2009) and subjected to analysis for determining the gene-clusters contributing to the traits or detecting the differentially expressed genes in two groups.

Presently, several approaches are followed to find out the differentially expressed genes/ transcripts between the two groups, namely, cluster analysis (Eisen *et al*., 1998), weighted gene co-expression network analysis (WGCNA) (Zhang and Horvath, 2005), partial correlation information theory (PCIT) (Reverter and Chan, 2008) etc. Bioinformatics software such as PLINK, R etc. are being extensively used for doing calculations and generating the graphs. The detail of the analysis part will be discussed in the practical demonstration session.

Limitations of GWAS:

GWAS can effectively map loci underlying the phenotypes of interest; however, it does not enable one to determine the causative genetic variation that confers its effect. This issue can be addressed by integrating multiple forms of data into a single analysis. Another, serious concern is the amount of phenotypic variance explained by the GWAS. Most of the SNP variations found by GWAS are associated with only a small amount of heritable variation, and have only a small predictive value (Ku *et al*., 2010).

Weighted Gene Co-expression Network Analysis (WGCNA):

Weighted gene co-expression network analysis (WGCNA) is a systems biology method for identifying the correlation patterns among genes across expression data generated from deepseq data/ microarray samples. The correlation between the expression profile of thousands of genes/ transcripts can be used for finding clusters (or modules i.e. highly interconnected genes with high correlation coefficient), for summarizing the clusters into topological overlap matrix plot, using the module eigengene (the first principal component of a given module) or an intramodular hub gene, for relating modules to one another and to external sample traits (using eigengene network methodology), and for calculating module membership (eigengene based connectivity) measures (Langfelder and Horvath, 2008, Zang and Horvath, 2005). Correlation networks has been applied in network based gene screening methods to identify candidate biomarkers or therapeutic targets, in various biological contexts (e.g. cancer, mouse genetics, yeast genetics, and analysis of brain imaging data). Correlation networks can be used as a data exploratory technique (similar to cluster analysis, factor analysis, or other dimensional reduction techniques) and as a screening method (Zhang and Horvath, 2005).

The detail of the correlation network construction (with practical data vis-à-vis simulation study) and conceptual discussion of the weighted gene correlation network analysis is available in the papers by Horvath *et al*., 2006, Zhang and Horvath, 2005 and Langfelder and Horvath, 2008. Readers are requested to go through the papers cited above to get detail insight of this approach. The basics of the WGCNA will be discussed here.

Correlation network methodology has a potential application in the Gene Co-expression networks that describes the pair-wise relationships among gene transcripts. Correlation networks are undirected, weighted gene networks. The nodes of such a network correspond to gene expression profiles. The nodes are connected if the corresponding genes are significantly overexpressed across appropriately chosen tissue samples (Zhang and Horvath, 2005, Horvath *et al*., 2006). The 'edges between genes' are determined by the pairwise correlations between gene expressions. The absolute value of the correlation coefficient is magnified by increasing the power to $\beta \geq 1$ (soft thresholding). The low correlation coefficients are thereby awarded penalty and high correlations are emphasized. WGCNA has been worked out using signed ($a_{ij} = (|cor(x_i, x_j)|)^{\beta}$) as well as unsigned ($a_{ij} = (|(1 + cor(x_i, x_j))/2|)^{\beta}$) correlation coefficient. The formulae given in parentheses represent adjacency between two nodes, where, x_{ij} is the quantitative measurements that can be described by an $n \times m$ matrix $X = [x_{il}]$.

Some of the salient features and definitions pertaining to WGCNA has been given below (Langfelder and Horvath, 2008):

1. The rationale behind correlation network methodology is to use network semantic to describe the pairwise relationships (correlations) between the rows of X.

2. The correlation networks can be used to find clusters (modules) of interconnected nodes.

3. WGCNA summarizes the node profiles of a given module by a highly connected hub node, which is centrally located in the module.

4. Module significance (for modules with high or average node significance) is determined as the average absolute gene significance measure for all genes in a given module. “Gene significance” is defined as the correlation of gene expression profiles with an external trait y. The measure tends to be highly related to the correlation between the module eigengene and the trait y.

5. Eigengene significance: The eigen-vectors of a square matrix correspond to the non-zero vectors that, after being multiplied by the matrix, either remain proportional to the original vector (i.e., change only in magnitude, not in direction) or become zero. For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector changes when multiplied by the matrix (http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors).

The “eigengene significance” is the correlation coefficient between the microarray sample trait y and the module eigengenes.

6. Annotation of network nodes: It is the module membership, in terms of the closeness of the nodes to the identified modules. The module membership accounts for the membership of the i^{th} gene with regard to the respective module, which is determined by a fuzzy measure to identify nodes that lie intermediate between and close to two or more modules. Highly connected intra-modular hub genes tend to have high module membership values to the respective module.

7. Gene significance (GS): It measures the biological significance of a particular gene to incorporate external information into the co-expression network. Gene significance of 0 indicates that the gene is not significant with regard to the biological question of interest.

8. Comparing two independent networks: a differential network analysis can be used to identify changes in connectivity patterns or module structure between different conditions. Consensus module analysis identifies the shared modules between two or more networks. The consensus modules are the building blocks in multiple networks, which represent fundamental structural properties of the network.

Conclusion:

Widespread use of DNA markers will have a major impact on the structure of the breeding programmes and a significant impact on production systems more generally. Breeding animals will be reared cheaply with minimum recording of phenotypes and pedigree. Selection will be based on a prediction equation derived from a reference population that has extensive

phenotypic recording and genotype data. To take maximum advantage of the genomic selection, generation intervals will be shortened as much as reproductive technology will allow. Genomic selection approach is still in its inception. It has been observed that genomic selection is more applicable in large animals than in the small one due to the difference in the generation interval that directly affects the economic feasibility of the project. Moreover, the Bayesian approach holds promise in estimating the molecular breeding values of the animals undergoing selection process. The accuracy of the predicted GEBV is also affected by the type of trait (viz. monogenic or polygenic), the initial values of pi and variances considered in the model and coverage of markers in the model. The time has come and we have to gear up before it gets further delayed or never!

##

Practical Session:

We will discuss two different applications of genome wide marker detection, namely, (a) Genomic Selection using GenSel software, and (b) Studying differential expression patterns using gene-expression data by Weighted Gene Co-expression network analysis (WGCNA).

Genomic Selection using GenSel software:

The GenSel program is extensively used for estimation of molecular breeding values of animals under selection, based on the SNP (or marker data) for the phenotype of interest. The software was written by Rohan Fernando and implemented by Dorian J. Garrick in the project Bioinformatics to Implement Genomic Selection (BIGS), at Iowa State University, Ames, Iowa, USA. The maiden version of GenSel was developed on MAC platform, using GNU compiler collection (GCC) along with libraries from GNU scientific libraries (GSL), MatVec and Boost. The latest version GenSel 4.0 utilizes MatVec, Boost and STL libraries, however, excludes the GSL. The gnuplot is required to display the graphics of posterior. GenSel is not permitted for public distribution; however, research collaborators and trainees can utilize the software through remote log-in.

GenSel can be operated through command line (MAC or Unix interface of BIGS) or through user-friendly menu driven approach (any operating system). The jobs are submitted in queue for analysis. The software can be used for estimation of marker effect of training data-set using Bayesian approach (Bayes A, B, C and C-pi) or least squares, linkage disequilibrium estimation an prediction of molecular breeding value using validation or unrelated marker data-set (without phenotypic records). Bayesian statistics uses all information surrounding the

likelihood of an event not only the data collected experimentally. If the variances are unknown, Bayes C is the best approach to estimate the variances and then these can be fit in Bayes B model which is more sensitive to the variance components.

In the practical session we will discuss the GenSel program in the following sections:

1. How to create the input data files
2. Job submission and job control
3. Analyzing the result files

Creating Input Data Files:

GenSel requires three types of input files: Genotype or marker data file, phenotype file and Map file.

Genotype Data File: The genotype data file is written in text format, as space delimited unix file, containing one header line (for the description of parameters) and matrix of covariates (genotypic values). For windows system, if the marker data is in Excel format (let, the file is example.xls), the same can be first saved as ".prn" format (i.e. example.prn), then the file format can be manually changed to ".txt", by renaming the file name from ".prn" to ".txt". One should be cautious that the columns have distinguishable distance and the data are left aligned, before converting to ".prn" format, else the columns may overlap with each other when the data is read by GenSel.

The genotype/marker data file contains animal Ids (in alpha-numeric format) in the first column followed by SNP-ids with substitution effect for each animal in subsequent column. The rows correspond to animals and the columns (excluding the first column, which is for animal Identifiers) harbors the covariate substitution effects for each marker or SNP. It is assumed that each marker position has got two SNPs (let, A and B), hence three different genotypes (AA, AB and BB) are assigned values like -10, 0 and +10, respectively. One relevant point is that, the effect of the markers is quantifiable, hence the heritability of the marker is also considered in the analysis.

Missing value(s): there are options to handle missing values in marker data file:

1. Calculate the average value for that genotype in the specific breed and replace the blank space with the average value.
2. Predict the genotypic value by using some haplotype estimation software, and replace the missing value. We can use PHASE software for haplotype analysis.
3. GenSel cannot handle missing values, hence these should be deleted otherwise. The file should have same column number for all the rows.

If map-file is not available, the SNP-locations should be arranged according to genomic location.

File-Size: The text-file format demands space, viz. for 50K SNP data file with 1000 animals will consume 1 GB of space. The alternative is to convert the file to binary format by gen2bin software, which reduces the space requirement to half. The binary file has "*.newbin" format.

Phenotype Data File: This file is lighter than the marker data file and one file contains only one phenotypic trait data. It contains Animal Identifier (exactly in the same format, as in the marker data file) in the first column, strictly followed by the phenotype column containing trait value. The following columns may contain the fixed effects (both covariate and categorical), although there is no specification for the effects.

There is a convention for defining the covariate and class variables:

1. Class variables: Column-name ending with $
2. Covariates: Column-name ending with no special symbol (only alphabetical)
3. Columns to be ignored during analysis: Column-name ending with
4. Weighted Analysis: Column headed by "rinverse", which means the inverse of a diagonal matrix of residual variance, is used to weigh the parameter with some weighing values.

It is not necessary to have exactly same animals present in genotype and phenotype files. The software matches the Ids and considers only the common animal identifiers for further analyses.

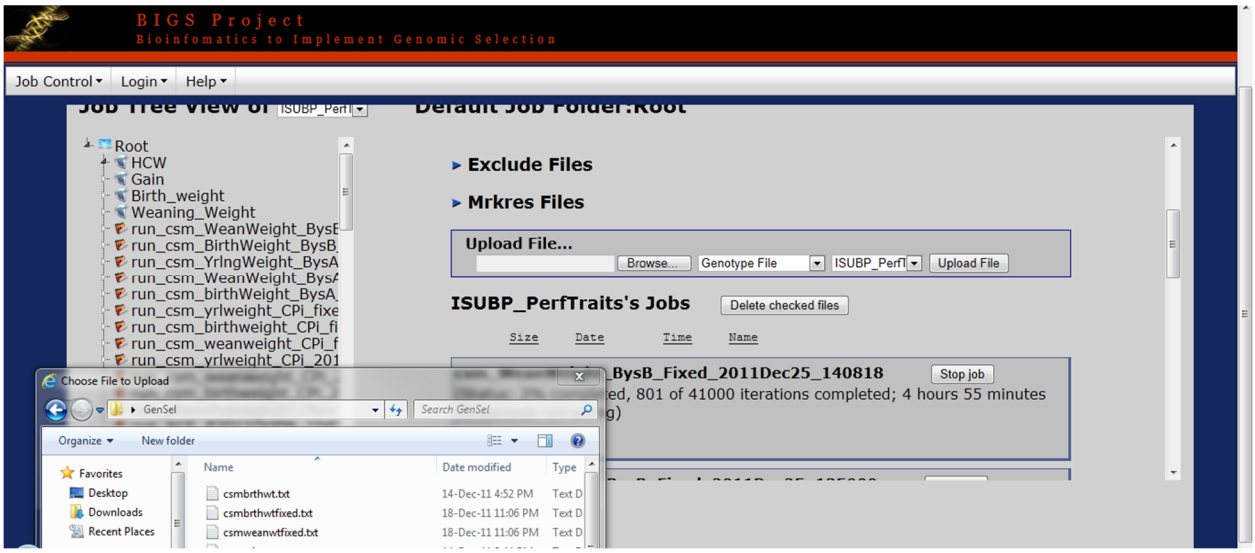
Map File: It provides chromosome and base-pair position (start and end position in the chromosome) information for at least one build to define the location of each SNP. Moreover, all the marker names in the genotype file must exist somewhere in the map file, while vice versa is not required.

Job Submission and Job Control:

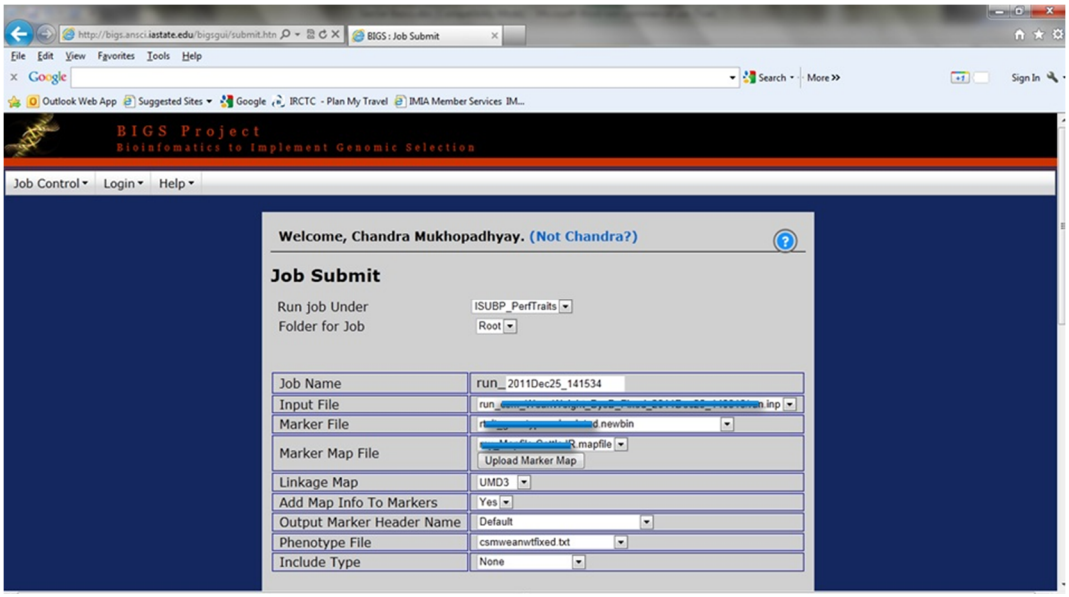
Job Control: First we need to sign in the BIGS GenSel program using own user Id and password, allotted by the GenSel Adminstrator(s). There are two options to run the program (under "Job Tree View of"), namely, ISUBP_PerfTraits (this is accessed and visible by all users) and your user name (this is only visible to you and cannot be accessed by others). It is better to use the first option, if the results are to be checked and discussed with users throughout the world.

There is a drop-down menu option on the left-hand side top for “Job Control”. Browse the pointer on this, then click on the drop-down option, “Gensel Job Control” to upload a new file through “Upload File” option. The new file will then be set through “Gensel Job Submit” option under the same drop down menu “Job Control”

.



Job Submit: It sets the parameters before running a new job. The phenotype file should be loaded from job control before setting the “Job Submit”. Otherwise, we can also use an already submitted phenotype file (if a new file is not to be uploaded).



Now, we need to set the parameters according to our requirement. The important points are discussed below:

1. Run job Under: ISUBP_PerfTraits or your personal domain, select any one according to your need.

2. Job Name: a distinguishable job name with trait-name should be given (the suffix with date should not be altered).

3. Input File: Select the file which has similar type (with your present analysis) of Analysis already run.

4. Marker File: Upload your own marker file or use previous one if a different phenotype is to be analyzed.

5. Marker Map File: To be uploaded afresh, if the marker file is changed.

6. Linkage Map: Depends on species and the reference genome used.

7. Phenotype File: Select your uploaded file.

8. Analysis type Bayes, LD (calculates linkage disequilibrium), LS (step wise analysis of the markers in forward and reverse submodels), Predict (for prediction of molecular breeding value) and Generate Data (generates simulated genotypes and phenotypes)

9. Bayes-type: Bayes A, B, C and C-pi.

The GenSel software has four Bayesian Approach: Bayes A, B, C and Cπ:

a) Bayes A: It follows Bayes B method without incorporating the “pi” value, i.e. “pi” is kept zero. It fits all the covariates in the model. All the markers are included in the model, assuming equal variance for all covariate.

b) Bayes B: Bayes B is run after setting the value of pi. It yields QTL Variance i.e. Variance of SNP effect in specific region of window (Window is a row of data that represents a single 1 Mb chromosomal region). All the markers are not included in the variance, and the genetic variance for the markers are assumed different. The Bayes B is very sensitive to prior assigned. The Vg and Vp is estimated by Bayes B and Bayes C. If the prior for Vg and Vp are unknown, it is better to run Bayes C (as it is less sensitive). For polygenic trait Bayes C is equally good as Bayes B.

c) Bayes C: To estimate the genotypic and phenotypic variances (V_G and V_P, respectively), without incorporating the “pi” value. Bayes C is less sensitive. For Bayesian analysis we need to use prior, hence, if we don’t know how much is the genetic and residual variance, the best method is Bayes B. It is a kind of generalized BLUP. Every fitted locus has the same variance, but a known fraction of loci have no effect. It uses a mixed model to estimate whether a locus should have a fitted effect or no effect. It is not influenced greatly by the priors for the genetic and residual variances as Bayes B which is sensitive to priors.

d) Bayes Cπ: Put the values of the V_G and V_P in the Bayes Cπ and run to obtain the pi which is fraction of markers having zero effect. Bayes Cπ is BayesC with “pi” estimation.

10. Chain Length: Total number of realizations/iterations to be set in the Markov chain. In general, chain length of 50,000 yields quite acceptable result, while increasing the length to 1,00,000 or more does not make any significant difference, since the genotypic and error variance converge over the iterations up to 50,000. Chain length and iteration are not different. In a single analysis there are some thousands of iteration or chain length, but there is a single chain.

11. Burnin: the number of cycles that are ignored while calculating the posterior genotypic and residual variances.

12. The parameters, probability fixed, genotypic and residual variances and degrees of freedom as well as the output frequency and random seeds are set by default values. It is not required to change unless the variance components are known (depending on traits).

13. Select the GenSel verion as “gensel4”.

14. Any comment can be put if required to mark the analysis.

Analyzing the result files:

GenSel yields many result files along with the zipped (*.tgz) file is to be downloaded and checked.

The important files are *.mrkRes1 (of marker results), *.winQTL1 and *.out1.

In the *.winQTL1 file (opened in a separate tab) the results we look for:

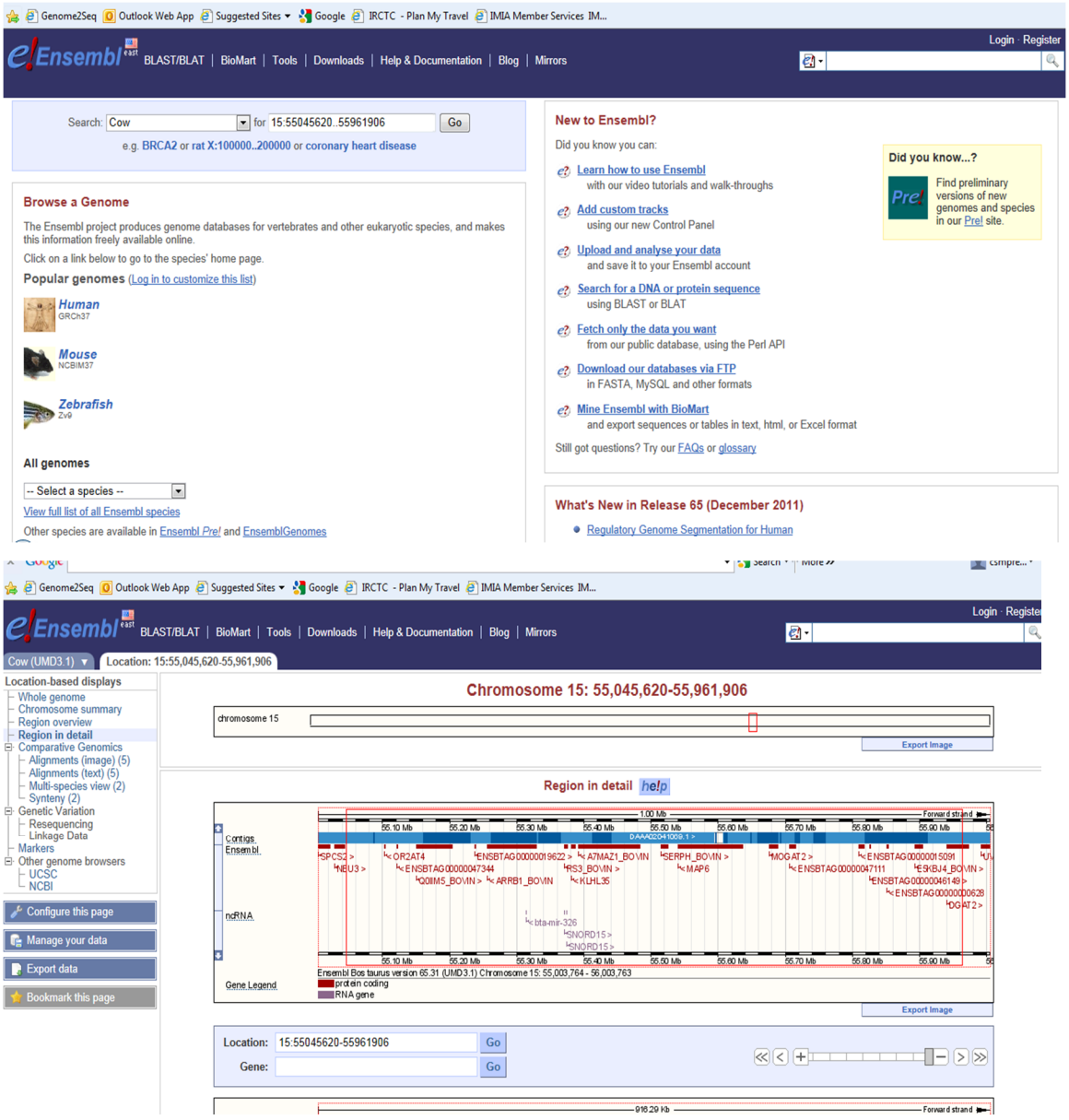
e) Higher %Var among the different 1 MB windows along with the number of SNPs available for that window.

f) Select the Map_Pos0 and map_posn and paste it on Ensembl (Choose species, put in the search box as “Chr: start..stop; illustration has been given below):

g) We need to look at every genes of the “Region in Detail” window. The basic idea is to study the function(s) of the genes in varioius species, with an aim to connect the traits. For this purpose we may need to go through the relevant literatures.

h) If more than one phenotypes are studied together, the files can be merged for the %var, SNPs, Map_Pos0 and map_posn, and p>0 to study all the phenotypes together. Now, select for those windows having higher %var vis-à-vis SNPs.

Some cautions: The results of the Bayes Cπ may not be useful due to weird values, like all pi=1.0 or specific other values, all windows belonging to the same chromosome. The range of genotypic and error variance should be acceptable. If such type of erroneous result(s) is(are) encountered, we need to adjust the specific parameters (variance, pi etc) to get acceptable results.



For the associated terminologies In the mark.Res1 file the GenSel manual published by Fernando and Garrick (2009) may be consulted. Here, the terms have been explained from the manual cited above:

Marker: It stands for the integer covariates used in the genotypic data file. In other words, the numerical values assigned to each marker loci along the columns of the marker data file. The Bayesian models differ in the fraction of markers included in the model, i.e. whether, all the markers in the dataset will be in the model (Bayes A) or a fraction of the marker (denoted by "1-pi") will be included in the model (Bayes B and C); while Bayes Cpi calculates the pi value, i.e., what proportion of markers are to be excluded from the model.

Effect column: Effect designates the effect of the covariates included in the model, the effect column values are calculated as the posterior mean of the covariate effect averaged

across the post-burnin chain. In the case of analyses with pi greater than 0 these effects are averaged across the chains that include and exclude the particular effect in the model.

EffectVar: It is the average of the variance used for that locus across the post-burnin chains including the iterations that excluded that covariate from the model.

Again, we can assume same genetic variance for all the markers (Bayes A, Bayes C, Bayes Cpi) or the markers included in the model exhibit different genetic variance (Bayes B).

ModelFreq: Every iteration in the Markov chain does not include the specific covariate marker in the model, rather, the number of times it is included determines the significance of the marker. The "ModelFreq" estimates the proportion of post-burnin iterations which included that particular covariate in the model. For the models that include "pi" value (1>p1>0), the model frequency is found to be highly correlated with the estimate of the absolute value of the effect, or the effect variance.

GeneFreq: It defines the frequency of the fitted allele, rather than the minor allele frequency.

Window: It is 1 MB in size. In GenSel 4, the Window is kept fixed to 1 MB and the number of SNPs in 1 MB window is studied, however, in older versions of GenSel, 5 SNPs were studied in one set, no matter what is the size of the Window.

Window Frequency: It is convenient to study the number of times the SNPs are appearing in the specific Window, rather than taking 1 MB region flanking a single SNP(s). Window frequency tells us how many times the different SNPs are coming in the model.

GenVar: it reflects the contribution of that locus to the genetic variance, calculated as 2p(1-p) times the square of the mean effects for that locus.

Effectdelta1: reports the posterior mean effect for only those chains that included the effect in the model, ie Effect/ModelFreq.

SDDelta1: is the standard deviation of the posterior distribution of effects for only those chains that included the effect in the model. It is the standard deviation of that effect in those chains where the model frequency is "0" (Fernando and Garrick, 2009).

t-like: is the ratio of the absolute value of EffectDelta1/SDDelta1. This latter statistic is highly correlated with ModelFreq but may have some additional utility in distinguishing informative markers with consistent versus inconsistent estimates of effects. In models with pi=0, modelFreq will be constant for all markers

Validation of marker results using Predict:

The posterior means obtained through training or test data set now can be used to predict the genomic expected breeding value (GEBV) of the validation data-set. We need the genotypes (marker file) of the later samples and the posterior means of the previous analysis. The markers

in both the cases should exactly align. The output file has the suffix ".ghat1" (g-hat). The file contains animal-Id in the first column, followed by genomic prediction of animal (g-hat value) followed by contents of the phenotypic file input. It is required for cross validation purpose also.

Weighted Gene Co-expression network analysis

In the practical session the data available from the paper by Ghazalpour *et al* (2006), Integrating Genetics and Network Analysis to Characterize Genes Related to mouse body weight will be used. R codes along with detailed explanatory notes are available in the "Tutorials for the WGCNA package" by Langfelder and Horvath (2008) at http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/ . The R codes will also be distributed (electronic copies) during practical session for explaining the code writing syntax.

'R' is a freely available integrated suite of software that includes effective data (including matrix and arrays) handling and storage facility, graphical display, conditionals and looping facility vis-à-vis user defined recursive function calling. The R software packages are comprehensive collection of R functions for performing various aspects of WGCNA. The package includes functions for network construction, module detection, gene selection, calculations of topological properties, data simulation, visualization, and interfacing with external software. The package will be studied using the test-data (http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA) on genes expressed in liver of female and male mice for finding clusters /modules of highly correlated genes, summarizing gene clusters using the module eigengene/ intramodular hub gene and for relating modules to one another and to external sample traits (for enrichment of the genes of interest).

Functions in the WGCNA package can be divided into the following categories (Zhang and Horvath, 2005):

1. network construction;
2. module detection;
3. module and gene selection;
4. calculations of topological properties;
5. data simulation;
6. visualization;

7. interfacing with external software packages.

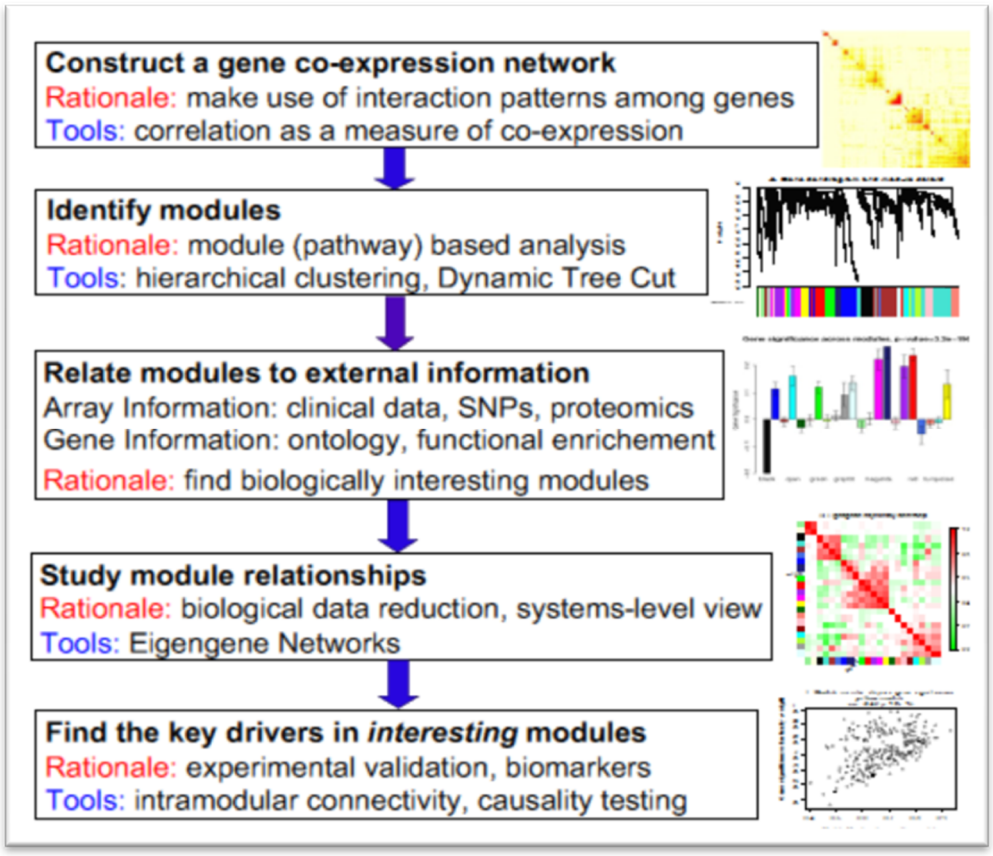


Figure: WGCNA methodology in a nutshell; (Langfelder and Horvath, 2008)

Requirements for Hands on Practice:

The participants are requested to download and install the following software and packages in their respective PCs, in order to save time during demonstration:

- Latest version of R (R-2.14.1 for Windows: 32/64 bit) can be downloaded from: cran.r-project.org/bin/windows/base/
- Expression, gene annotation and phenotype data sets (namely, LiverFemale3600.csv, GeneAnnotation.csv, ClinicalTraits.csv, respectively) on female mouse liver sample, available at: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/
- Please install the R packages: “impute", "dynamicTreeCut", "cluster", "flashClust", "Hmisc", from the Comprehensive R Archive Network (CRAN) site: : http://cran.r-project.org/

- An exhaustive tutorial on WGCNA using the data sets and R is available at: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/
- WGCNA User Manual: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Softwares/WGCNA/WGCNA%20User%20Manual%201.0.doc

References:

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., Zondervan, K.T. 2011. Basic statistical analysis in genetic case-control studies. *Nat Protoc*., **6 (2)**: 121–33.

Daetwyler, H. D., Villanueva, B., Bijma, P. and Woolliams, J.A. 2007. Inbreeding in genome-wide selection. *J. Anim. Breed. Genet*., **124**: 369-376.

Danesh, J. and Pepys, M.B. 2009. C-reactive protein and coronary disease: is there a causal link?. *Circulation*, **120 (21)**: 2036–9.

DiPetrillo, K., Wang, X., Stylianou, I.M. and Paigen, B. 2005 Bioinformatics toolbox for narrowing rodent quantitative trait loci. *Trends Genet*., **21(12)**: 683–92.

Drake, T.A., Schadt, E.E. and Lusis, A.J. 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mamm Genome*, **17(6)**:466–79.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci*., **95(25)**, 14863–14868.

Fernando, R. and Garrick, D. 2009. GenSel- User Manual for a portfolio of Genomic Selection related Analyses. (http://taurus.ansci.iastate.edu/Site/Welcome_files/GenSel%20Manual%20v2.pdf)

Ghazalpour, A., Doss, S., Zhang, B., Plaisier, C., Wang, S., Schadt, E.E., Thomas, A., Drake, T.A., Lusis, A.J. and Horvath. S. 2006. Integrating genetics and network analysis to characterize genes related to mouse weight. *PloS Genetics*, **2(8)**: e130, 2006.

Goddard, M. and Hayes, B. 2007, Genomic selection. Journal of Animal Breeding and *Genetics*, **124**: 323–330.

Goddard, M.E. and Hayes, B.J. 2007. Genomic selection. *J. Anim. Breed. Genet*., **124**: 323–330.

Hayes, B.J., Chamberlain, A., Goddard, M.E. 2006 Use of linkage markers in linkage disequilibrium with QTL in breeding programs. Proc. 8th World Congr. Genet. Appl. Livest. Prod.. Belo Horizonte, Brazil, 8, pp. 30–06.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M. , Laurance, M.F., Zhao, W., Shu, Q., Lee, Y., Scheck, A.C., Liau, L.M., Wu, H. , Geschwind, D.H. , Febbo, P.G. , Kornblum, H.I., Cloughesy, T.F. , Nelson, S.F. and Mischel, P.S. 2006. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a novel molecular target. *Proc. Natl. Acad. Sci. USA*., **103(46)**: 17402-17407.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh, J. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, **308**: 385–9.

Ku, C.S., Loy, E.Y., Pawitan, Y. and Chia, K.S. 2010. The pursuit of genome-wide association studies: where are we now?. *J. Hum. Genet*., **55 (4)**: 195–206.

Langfelder, P. and Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, **9**:559:560.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., *et al*. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Genet*., **9**: 356–369.

Meuwissen, T., Hayes, B. and Goddard, M. 2001. Prediction of total **genetic** value using genome-wide dense marker maps. *Genetics*, **157**: 1819-1829.

Misztal, I. 2006. Challenges of application of marker assisted selection – a review. *Animal Science Papers and Reports*. **24(1)**: 5-10

Reverter, A. and Chan, E.K.F. 2008. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**, 2491–2497.

Schaeffer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.,* **123**: 218-223.

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., *et al*., 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet*., **1(6)**: e78.

VanRaden, P.M., Van Tassel, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. and Schenkel, F.S. 2009. Reliability of Genomic Predictions for North American Holstein Bulls. *J. Dairy Sci*., **92**: 16-24.

Zhang, B. and Horvath, S. 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.*, **4(1)**: 1:45. http://www.bepress.com/sagmb/vol4/iss1/art17