14th Quantitative Trait Locus and Marker Assisted Selection Workshop

Poznań, Poland

17-18 May 2010

Poznań University of Life Sciences

The book of abstracts
14th QTL – MAS Worksop

The book of abstracts

14th QTL – MAS Workshop

The 14th QTL-MAS workshop has been organized by **Poznań University of Life Sciences, Department of Genetics and Animal Breeding**, in collaboration with **Wroclaw University of Environmental and Life Sciences** and **The Wielkopolska Regional Centre of Animal Breeding and Reproduction in Poznań, based in Tulce Ltd.**

The core of the **Organizing Committee** consists of:

Paweł Krajewski	Krzysztof Moliński
Tomasz Strabel	Tomasz Szwaczkowski
Joanna Szyda	Maciej Szydłowski

The Workshop is organized with support from:

Alicja Borowska	Dagna Kręgielska
Sebastian Mucha	Paulina Paczyńska
Marcin Pszczoła	Katarzyna Rzewuska
Anna Wolc	Kacper Żukowski

This book of abstracts has been edited by the Organizing Committee.

HONORARY PATRONAGE

14th QTL-MAS 2010 workshop has been organized under the honorary patronage of:

Minister of Agriculture and Rural Development

MARSHAL
OF THE WIELKOPOLSKA REGION
Marek Woźniak

Marshal of the Wielkopolska Region

POZnań
*Eastern energy, Western style

The Mayor of the City of Poznań

The following sponsors have provided financial support to the 14th QTL-MAS workshop:

Breeding and Insemination Centre "SHIUZ" in Bydgoszcz

The Wielkopolska Regional Centre of Animal Breeding and Reproduction in Poznań, based in Tulce Ltd.

Illumina Inc.

Cobb-Vantress, Inc.

Mazowieckie Centrum Hodowli i Rozrodu Zwierząt Sp. z o.o. w Łowiczu

Polish Federation of Cattle Breeders and Dairy Farmers

P. H. KONRAD in Łomża

SAS Institute

The Małopolskie Biotechnology Centre company

CONTENTS

PROGRAM

Monday May 17th **Morning**

Chair: Tomasz Strabel **9:00 – 12:30**

Session I Genetic architecture and QTL mapping of quantitative traits

09:00	Organizing Committee	Opening and introduction
09:15	**William Hill**	**Genetic architecture of quantitative traits (invited)**
09:45	Mats Pettersson	Re-examining epistatic interactions in Virginia line chickens
10:00	Carl Nettelblad	Haplotype inference based on Hidden Markov Models in the QTL-MAS multi-generational dataset
10:15	Maciej Szydłowski	QTL-MAS 2010: Simulated Dataset
10:30 - 11:00	**Coffee break**	
11:00	Aniek Bouwman	A Bayesian approach to detect (pleiotropic) QTL affecting a simulated binary and quantitative trait
11:15	Albart Coster	Partial least square regression applied to the QTL-MAS 2010 dataset
11:30	Mario Calus	Genomic breeding value estimation and QTL detection using univariate and bivariate models
11:45	Xia Shen	Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps
12:00	Burak Karacaören	Association analyses of the QTL-MAS data set using grammar, principal components and Bayesian network methodologies
12:15	Kacper Żukowski	Testing association of genotypes with discrete and continuous traits
12:30 - 13:30	**Lunch**	

Monday May 17th **Afternoon**

Chair: Paweł Krajewski **13:30 – 17:00**

Session II Methods for QTL mapping detection and association analysis

13:30	**Karl Broman**	**The genetic dissection of complex traits in model organisms (invited)**
14:00	Animesh Acharjee	Integrating genetic markers with ~omics data using genetical genomics and modern regression methods
14:15	Ghyslaine Schopen	Genome wide association study using single and multiple SNP analysis
14:30	Simon Teyssèdre	Robustness and power of single-SNP analysis in related populations
14:45	Lucy Crooks	Comparison of an improved method for calculating line origin probabilities against GridQTL using simulated data

15:00 - 15:30 Coffee break

15:30	**Yurii Aulchenko**	**Tutorial on *ABEL package (invited)**
16:00	Robin Wellmann	Genome wide evaluation using dominance
16:15	Anna Johansson	Genome wide effects of divergent selection for body weight in chickens
16:30	Saber Qanbari	Using genome scans of DNA polymorphism to identify regions exhibiting positive selection
16:45	Organizing Committee	Comparative analysis of submitted results on QTL mapping and applied methods

18:30 – 21:00 Dinner

Tueasday May 18th **Morning**

Chair: Joanna Szyda **9:00 – 12:30**

Session III Methods for estimation of genomic breeding values

09:00	**Małgorzata Bogdan**	**Genome-wide scans for quantitative trait loci in experimental populations - issues of multiple testing and model selection (invited)**
09:30	Matthew Cleveland	Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals
09:45	Joseph Ogutu	A comparison of random forests, boosting and support vector machines for genomic selection with SNP markers
10:00	Zhe Zhang	Genomic selection using Best Linear Unbiased Prediction with a trait specific relationship matrix
10:15	Javad Nadaf	Applying different genomic selection approaches on QTL-MAS 2010 data
10:30	Torben Schulz-Streeck	Pre-selection of markers for genome-wide selection
10:45 - 11:15	**Coffee break**	
11:15	**Yurii Aulchenko**	**Methods for genome-wide association analyses of quantitative trait loci in human genetically isolated populations (invited)**
11:45	Ronald Nelson	A new R package for QTL analysis
12:00	Kacper Żukowski	The estimation of SNP effects on a binary and a quantitative trait
12:15	Magdalena Malina	Logic regression based methods in application to detection of gene-gene interactions
12:30 - 13:30	**Lunch**	

Tuesday May 18th **Afternoon**

Chair: Maciej Szydłowski 13:30 – 16:45

Session IV Application of genomic selection

13:30	**Alain Charcosset**	**Genetic architecture of yield and related traits in European maize: insights into the effects of linkage and allelic series. Consequences for marker assisted selection (invited)**
14:00	Weronica Ek	Mapping systemic scleroderma genes in a cross between UCD200 and jungle fowl chickens
14:15	Joanna Szyda	Genomic selection in Polish Holstein
14:30	Jan-Thijs van Kaam	Validation experiences in Italian Holstein genomic selection
14:45	Organizing Committee	Analysis of submitted breeding values and applied methods
15:00 - 15:30	**Coffee break**	
15:30	**Karl Broman**	**Tutorial on R/qtl (invited)**
16:15	Paweł Krajewski	POLAPGEN-BD: a project on biotechnology for breeding cereals with increased resistance to drought
16:30	Organizing Committee	QTL-MAS 2010 Closing remarks

ABSTRACTS

Genetic architecture of quantitative traits

William G. Hill[1*]

[1]Institute of Evolutionary Biology, University of Edinburgh

[*] Presenting author: William Hill, email: w.g.hill@ed.ac.uk

Understanding the 'black box' of the genetics of quantitative traits has been a long standing objective. Important unknowns are the numbers of loci involved, and the distribution of their frequencies, effects on the trait and their interactions These depend both on the basic architecture and on the selective and other forces whereby variation is lost or maintained. Some information has come at the quantitative level from covariances among relatives, inbreeding and selection experiments which indicate many loci are responsible. The availability of dense markers provides the opportunity for deeper study, but although some genes of large effect have been detected, genome wide association studies of traits in humans also indicate many loci are contributing to variation.

I shall discuss what we may expect on the basis of population genetics theory and knowledge of parameters to be the distributions of effects and frequencies, how these influence inferences that can be drawn from GWAS and similar studies, and how these relate to observations.

Re-examining epistatic interactions in Virginia line chickens

Mats E. Pettersson[1*]

[1] Computational Genetics, Department of Animal Breeding and Genetics, Swedish Univeristy of Agriculture
* Presenting author: Mats Pettersson, email: mats.pettersson@hgen.slu.se

It has previously been shown that a substantial fraction of the phenotypic difference between the high growth and low growth lines of Virginia lines chicken, which show a nine-fold difference in body-weight after 50 generations of differential selection, can be attributed to epistatic interactions between different QTL[1].

The present study is based on data from an advanced intercross line, with a pedigree reaching down to an F_8 generation, that is an extension of the F2 population originally used. In this data set, which also has a higher density of markers, the epistatic interactions, as well as single QTL effects, have been re-evaluated. I employ a bootstrapping method to improve the reliability of QTL detection, and use data stratification to show epistatic interactions.

The results are that several of the epistatic interactions previously found are recovered in this analysis, reaffirming the importance of epistasis in the system. However, there are also some differences; not all interacting pairs are replicated and the total effect of epistasis is somewhat lower. Nevertheless, the overall picture is largely the same.

1. Carlborg, Ö., Jacobsson, L., Åhgren, P., Siegel, P. & Andersson, L. Epistasis and the release of genetic variation during long-term selection. Nature Genetics 38, 418-419 (2006).

Haplotype inference based on Hidden Markov Models in the QTL-MAS multi-generational dataset

Carl Nettelblad[1*]

[1] Department of Information Technology, Uppsala University
∗ Presenting author: Carl Nettelblad, email: carl.nettelblad@it.uu.se

Background. In previous work (Nettelblad et al., LNCS 6462, 2009, Springer Verlag) we demonstrated an approach for efficient computation of genotype probabilities, and more generally probabilities of allele inheritance in inbred as well as outbred populations. This work also included an extension for haplotype inference, or phasing, using Hidden Markov Models. The code was able to phase > 99.0 % of all heterozygous markers in the dataset from the 12th QTL-MAS correctly.
Computational phasing of multi-thousand marker datasets has not become common as of yet. In this work, we elaborate on the differences in allele probabilities, as well as QTL detection performance, when considering an identical workflow with and without phasing based on expectation-maximization.

Results. Using phasing for 20 iterations, almost all heterozygous markers in all generations converged. This makes actual allele origin traceable, back to the founder generation, in over 99 % of all locus-individual pairs. Some specific regions turned out to be hard to track, but the artifacts were minor in a dense marker set of the current type. Haldane mapping distances were also computed during the phasing.
A litter/parental effect was fitted for each pair of parents for the scalar phenotype in the dataset, explaining about 27.7 % of the total phenotypic variance in that phenotype. Obviously, some genetic variance is also included in this number.
The residuals from the litter fitting were fitted against loci, with an indicator vector for each individual indicating the allele origin probability for each of the 40 founder alleles. When using consecutive forward selection, 5 QTL were identified above a threshold derived from random permutations, accounting for 14.30 % of the phenotypic variance.
Corresponding results for an identical workflow without phasing, unable to distinguish between founder alleles resulted in a total explainable variance of only 6.97 % for the 5 highest ranking QTL, although not all of those were found significant.

Conclusions. Using inferred phasing can greatly increase power and accuracy in multi-generational QTL detection, at least when simple models are employed.

QTL-MAS 2010: Simulated Dataset

Maciej Szydlowski[1*]

[1] Poznań University of Life Sciences, Poznan, Poland

[*] Presenting author: Maciej Szydłowski, email: mcszyd@jay.au.poznan.pl

Hypothetical pedigree data set was simulated for a population consisting of 3226 individuals in 4 generations with large full-sib groups (about 30 offspring per mating).

Genome size was about 500 mln bp and it consisted of 5 chromosomes, each about 100 mln bp long. The alleles for 20 founders were generated by the use of *mh* software (Hudson 2002) and then dropped down the pedigree assuming incomplete interference. The SNP markers were a random sample from all biallelic sites with MAF>0.1. For each individual a set of 10031 unordered genotypes was available.

Two traits were generated: a quantitative trait (Q) and a binary trait (B). The Q trait was simulated assuming 30 random biallelic additive QTLs, two pairs of epistatic QTLs and 3 imprinted loci (active paternal allele) and the narrow-sense heritability of 33% (30 add loci). The B trait was generated under pure additive threshold model, assuming 22 functional SNPs and heritability of 50%. The two traits shared 22 additive QTLs. Genotypes for 10 functional SNPs were available in the data set. No systematic environmental effect was generated, residuals were normal and uncorrelated. Phenotypes for generation 4 was removed from the dataset.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337-8.

A Bayesian approach to detect (pleiotropic) QTL affecting a simulated binary and quantitative trait

Aniek C. Bouwman[1*], Luc L.G. Janss[2], Henri C.M. Heuven[3]

[1] Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[2] Aarhus University, DJF Department of Genetics and Biotechnology, P.O. Box 50, 8830 Tjele, Denmark
[3] Clinical Sciences of Companion Animals, Faculty of Veterinary Medicine, Utrecht University, P.O. Box 80163, 3508 TD Utrecht, the Netherlands
∗ Presenting author: Aniek Bouwman, email: Aniek.Bouwman@wur.nl

Background. The simulated dataset of the QTL-MAS 2010 workshop in Poznan (Poland) was analyzed. The data contained 2,326 individuals phenotyped for two traits: a quantitative trait and a binary trait. The aim of this paper was to detect QTL and to recover the genetic architecture of the traits. A genome-wide association was performed for each trait considering all SNPs simultaneous using a Bayesian algorithm.

Results. Applying a bivariate animal model in ASReml showed heritabilities of approximately 53% for the quantitative trait and approximately 22% for the binary trait. The genetic correlation between both traits was 0.66. The genome-wide scan revealed eight significant and one putative QTL affecting the quantitative trait. The genome-wide scan revealed four significant and six putative QTL affecting the binary trait. The QTL differed in size. All SNPs together explained 26% of the phenotypic variation of the quantitative trait, and 38% of the phenotypic variation of the binary trait. None of the QTL appeared on chromosome 5. Two QTL on chromosome 2 and one QTL on chromosome 3 had an effect on both traits, indicating pleiotropy, which explains the genetic correlation between the traits.

Conclusions. Several QTL have been detected for both traits. The three pleiotropic QTL explain the genetic correlation between the traits.

Partial least square regression applied to the QTL-MAS 2010 dataset

Albart Coster[1], Mario P. L. Calus[2*]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands
[2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands
* Presenting author: Mario Calus, email: mario.calus@wur.nl

Background. Partial least square regression (PLSR) was used to analyze the data of the QTL-MAS 2010 workshop. The objectives of this analysis were to identify genomic regions affecting either one of the two simulated traits and to accurately predict breeding values for the two traits of the simulated individuals. PLSR was especially appropriate for the analysis of these data because it enables to simultaneously fit several traits to the large number of markers in the data.

Results. A preliminary analysis showed that the two simulated traits were phenotypically and genetically correlated. Consequently, the data for the two traits were analyzed jointly in PLSR for each chromosome independently. Regression coefficients estimated for the markers were used to calculate the variance of each marker and QTL inference was based on local maxima of a smoothed line traced through these variances. In this way, 23 QTL for the continuous trait were found and 30 for the discrete trait. There was clear evidence for pleiotropic QTL on chromosomes 1 and 3. Bootstrapping was used to calculate empirical standard errors of the regression coefficients and the 2000 most significant markers were used in a second PLSR model which was used to estimate breeding values of the individuals based on the markers. Breeding values estimated with this method correlated well (correlation > 0.50) with the observed phenotypes of both traits.

Conclusions. Results of this analysis show the viability of PLSR for QTL analysis of multivariate models and for estimating breeding values using markers. The methods used in this study will need to be compared to the other methods used in this workshop.

Genomic breeding value estimation and QTL detection using univariate and bivariate models

Mario P.L. Calus[1*], Han. A. Mulder[1], Roel F. Veerkamp[1]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, Netherlands
* Presenting author: Mario Calus, email: mario.calus@wur.nl

Background. The provided simulated dataset, including a quantitative and a binary trait, was analyzed with four univariate and bivariate linear models to predict breeding values for animals without phenotypes. Two models were used that estimated variance components with REML using a numerator relationship matrix (A), or a SNP based genomic relationship matrix (G), as well as two SNP based Bayesian models with one (BayesA) or two distributions (BayesC) for estimated SNP effects. The bivariate BayesC model sampled QTL probabilities for each SNP conditional on both traits. Genotypes were permutated 2000 times against the phenotype and pedigree data, to obtain significance thresholds for the posterior QTL probabilities.

Results. Estimated breeding values had correlations with phenotypes in the reference population ranging from 0.75 and 0.89 for the quantitative trait, and from 0.58 to 0.67 for the binary trait.
Correlations were calculated between all different estimated breeding values, across models and traits, for animals without phenotypes. Correlations between the different SNP based models were greater than 0.93 (0.87) for the quantitative (binary) trait. Correlations between both traits ranged from 0.48 to 0.77 for the SNP based models, and from 0.36 to 0.61 for model A. Correlations between both traits were on average 0.78 (0.55) for the bivariate (univariate) models. Estimated genetic correlations were 0.71 (0.66) for model G (A).
The bivariate BayesC model detected 17 significant SNPs at the genome-wide level and 24 significant SNPs chromosome-wide. Those SNPs clustered into 14 different windows of 2Mb, suggesting that 14 QTLs were detected.

Conclusions. Estimated breeding values of three different SNP based models, both in their univariate and bivariate forms, were in good agreement. Correlations between estimated breeding values of both traits indicated that bivariate models made better use of the data. Permutating the genotypes against phenotype and pedigree data in the BayesC model provided an effective way to derive significance thresholds for the posterior QTL probabilities.

Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps

Xia Shen[1,2*], Lars Rönnegård[2,3], Örjan Carlborg[1,3]

[1] The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden
[2] Statistics Group, Dalarna University, Borlänge, Sweden.
[3] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.
∗ Presenting author: Xia Shen, email: xia.shen@lcb.uu.se

Background. Genome-wide dense markers can be used to detect genes and estimate polygenic effects. Among many methods, Bayesian techniques have been shown to be powerful in genome-wide breeding value estimation and association studies. However, computation is known to be intensive in the Bayesian framework, and specifying a suitable prior distribution for each parameter is difficult. We propose to use hierarchical likelihood as an alternative statistical approach to map genes and estimate genomic breeding values in datasets with dense marker maps, to solve such problems. Using double hierarchical generalized linear models, estimation of marker-specific variance is unified in one model and estimated using a fast iterative algorithm solely based on weighted least squares.

Results. We analyzed the dataset distributed for the QTL-MAS conference 2010 using double hierarchical generalized linear models and report breeding value estimates and detected QTL. The estimated breeding values obtained using double hierarchical generalized linear models were quite similar to those obtained using generalized linear mixed models (Pearson correlation of 0.990 for the quantitative trait and 0.987 for the binary trait). Using a smoothed version of the double hierarchical generalized linear model that we propose, QTL were clearly mapped by estimating marker-specific variances.

Conclusions. Hierarchical likelihood enables estimating marker-specific variances under a non-Bayesian framework. Double hierarchical generalized linear models can be estimated using an iterative algorithm, which greatly shortens the execution time comparing to the Bayesian computation. There is furthermore no need to specify any priors. Smoothing by defining spatial correlation reduces noises at zero-effect markers and is powerful in localizing QTL. Estimating epistatic effects is also possible by such a unified analysis via hierarchical likelihood.

Association analyses of the QTL-MAS data set using grammar, principal components and Bayesian network methodologies

Burak Karacaören*, José M. Álvarez-Castro, Chris S. Haley, Dirk Jan de Koning

The Roslin Institute and R(D)SVS, University of Edinburgh, EH25 9PS, Roslin, UK
* Presenting author: Burak Karacaören, email: burak.karacaoeren@roslin.ed.ac.uk

Background. In this study we want to apply association analyses with machine learning methods to the workshop data. It has been shown that if genetic relationships among individuals are not taken into account for genome wide association studies, this may lead to false positives. To address this problem, we used Genome-wide Rapid Association using Mixed Model and Regression (Aulchenko et al, 2007) and principal component stratification analyses (Price et al, 2006). It has been shown that using principal components loadings obtained from top markers as covariate may be useful to choose most significant SNPs based on correction for linkage disequilibrium (Pant et al, 2010). Estimation of Bayesian networks may also be useful to investigate linkage disequilibrium among SNPs and relation with environmental variables.
For the quantitative trait we first estimated residuals while taking polygenic effects into account. We then used a single SNP approach to detect most significant SNPs and applied principal component regression to take linkage disequilibrium among SNPs into account. For the categorical trait we used principal component stratification methodology with first 10 principal components. For correction of linkage disequilibrium we used principal component logit regression. Bayesian networks were estimated to investigate relationship among SNPSs. Using the natural and orthogonal interactions model we estimated the effects of the detected SNPs from previous approaches.

Results. Using the Genome-wide Rapid Association using Mixed Model and Regression and principal component stratification approach we detected around 100 of significant SNPs for the quantitative trait ($p<0.05$) and 318 of significant (FDR=0.05) SNPs for the categorical trait; with additional principal component regression we reduced the list to 16 and 50 SNPs for the quantitative and categorical trait, respectively. Estimated SNP effects were similar using the linear model and the natural and orthogonal interactions model. Although deviation from normality in residuals may give advantage to the orthogonal model.

Conclusions. By combining a number of existing approaches, we performed a comprehensive analysis of the workshop data. Principal component methods can be used to deal with population structure as well as linkage disequilibrium among significant SNPs.

Testing association of genotypes with discrete and continuous traits

Kacper Żukowski[1*], Anna Macierzyńska[1], Heliodor Wierzbicki[1]

[1] Department of Genetics and Animal Breeding, Wrocław University of Environmental and Life Sciences, Kożuchowska 7, 51-631 Wrocław, Poland

[*] Presenting author: Kacper Żukowski, email: kacper.zukowski@up.wroc.pl

Background . The aim of this study was to predict breeding values of continuous and binary trait and to investigate associations between genetic markers and both traits. Different approaches were used to investigate both traits – some of them considered each trait separately and the other considered traits simultaneously.

Methods. Three groups of models were used. First group, named BLUP, was used for investigating associations between SNPs and continuous and binary trait. SNP's effects were treated as random whereas mean and gender were treated as fixed effects. The second group of models, named gBLUP, was applied to predict genomic breeding values, which were considered as random effects. The fixed effects were analogical as in the first group of models. In both a. m. groups, two models considered both traits simultaneously. Then one trait was put as the fixed effect in the model for the second trait. The third group of models, named Foulley's methods, considered both traits simultaneously. Then the fixed and the random effects for both traits were estimated in the same time (Foulley et al., 1983). Foulley's methods were used to predict breeding values and to estimate SNPs effects, which were treated as random effects. Results received from Foulley's method were compared with results obtained using two first groups of models.

Results and Conclusions. The correlation between EBV and GEBV for the quantitative trait was 0.67 for animals with phenotype and for all animals. For the binary trait accuracy was 0.78 for all animals and 0.84 for animals with phenotype. Results obtained suggest that additive effect of SNPs in genome were not the same for the quantitative and the binary trait.

Foulley J.L., Gianola D., Thompson R. (1983) Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. Genet. Sel. Evol., 15(3), 401 – 424

The genetic dissection of complex traits in model organisms

Karl W. Broman[1*]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison
* Presenting author: Karl Broman, email: kbroman@biostat.wisc.edu

Both agricultural and biomedical researchers seek to dissect the genetic architecture of complex phenotypes, but their goals are somewhat different. Consequently, the appropriate analysis strategies have important differences. I will discuss some of these differences and will further describe recent advances in complex trait analysis in model organisms.

Integrating genetic markers with ~omics data using genetical genomics and modern regression methods

Animesh Acharjee[1,2,*], Bjorn Kloosterman[1], Chris Maliepaard[1,3], Ric de Vos[4], Christian Bachem[1,3], Richard GF Visser[1,3]

1Wageningen UR Plant Breeding, Wageningen University and Research Center, PO Box 386, 6700 AJ Wageningen, The Netherlands
2 Graduate School Experimental Plant Sciences
3 Center for BioSystems Genomics
4 Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands
*Presenting author: Animesh Acharjee, email: animesh.acharjee@wur.nl

Utilization of the natural genetic variation in traditional breeding programs remains a major challenge in crop plants. In the post genomics era, high throughput technologies give rise to data collection in fields like transcriptomics, metabolomics and proteomics and as a result, large amounts of data have become available. We have screened a diploid potato population for gene-expression and obtained LC-MS data resulting in the identification of many expression and metabolite QTL's across the genome. However, the integration of these data sets with phenotypic and marker data is still problematic. Here we present novel approaches to study the various ~omic datasets to allow the construction of networks integrating gene expression, metabolites and markers. We used univariate regression and modern regression methods like lasso, elastic net and sparse partial least squares regression to select a subset of the metabolites and transcripts which show association with potato tuber flesh colour. The selected subset of metabolites and transcripts shows high significant (p value<2.2e-16) to the flesh colour trait and variance explained by regression model is about seventy one percent.

Genome wide association study using single and multiple SNP analysis

G.C.B. Schopen[1*], M.P.L. Calus[2], M.H.P.M. Visker[3], J.A.M. van Arendonk[3], H. Bovenhuis[3]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, P.O. Box 338, 6700 AH Wageningen, the Netherlands

[2]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, the Netherlands

∗ Presenting author: Ghyslaine Schopen, email: Ghyslaine.Schopen@crv4all.com

Background. The objective of this study was to compare the SNPs showing the most significant effects, the location and the fraction of variance explained by these SNPs between single SNP analysis and multiple SNP analysis in the Dutch Holstein-Friesian population. The comparison was performed for the relative concentrations of the six major milk proteins. In total, 1713 cows with genotypes and phenotypes were available. In total, 45,999 SNPs distributed across 29 bovine autosomes were used in the single and multiple SNP analyses.

Results. The same main four chromosomal regions on BTA5, 6, 11, and 14 were detected in the single and multiple SNP analysis. The proportion of genetic variance explained by each of the SNPs in the single SNP analysis was higher compared to the SNP with the highest posterior probability in the multiple SNP analysis, except for β-casein. Summing up the proportion of genetic variance explained by adjacent SNPs next to the SNP with the highest posterior probability in the multiple SNP analysis, resulted in an increase of the genetic variance explained similar to the SNP most significantly associated in the single SNP analysis, except for β-casein. There was one additional region on BTA7 detected in the multiple SNP analysis. The number of SNPs with effects is considerably lower in the multiple SNP analysis as compared to the single SNP analysis.

***Conclusions*.** Multiple SNP analysis result in higher power and in higher mapping precision to detect QTL as compared to the single SNP analysis.

Robustness and power of single-SNP analysis in related populations

S. Teyssèdre[1*], J-M. Elsen[1] and A. Ricard[2]

[1] INRA, UR 631, 31326 Castanet-Tolosan, France
[2] INRA, UMR 1313, 78352 Jouy-en-Josas, France
∗ Presenting author: Simon Teyssedre, email: simon.teyssedre@toulouse.inra.fr

Background. The availability of the SNP array in many species reinforced the idea of using linkage disequilibrium (LD) for QTL detection. Quite often, LD methods for QTL detection made the hypothesis of unrelated animals and this is not the case in most animal species. The aim of this study was to formulate algebraically and evaluate in practical example theoretical robustness and power of some single-SNP test such as regression or GRAMMAR (Aulchenko *et al.,* 2007) when animals are related and/or when paternal half sib's families are used.

Results. The obtained formulae clearly demonstrate that Regression analysis gave a false discovery rate (FDR) which was higher when the heritability of the trait increased. Moreover, FDR also increased with the number of progeny per sire and the variability of relationships in the sample : With a sample size of 600 and h^2=0.5, FDR was 5% for unrelated population, 8% for random sample with relationships parameters of French Trotters and 9% (resp. 26%) for 5 (resp. 60) progeny per sire. In spite of increase of FDR, the power decreased with increase of variability of relationships (12% less for a $0.20\sigma_p$ SNP effect and 60 progeny per sire). GRAMMAR analysis gave lower FDR and power when relationships existed: FDR was 2.5% at the nominal 5% and the decrease of power was 12% for 60 progeny per sire with a sample size of 600 and h^2=0.5. Moreover, the estimate of the SNP effect was strongly biased downwards when h^2 increased: bias was -58% with the above sample.

Conclusions.Regression and Grammar were not robust against population structure, Regression analysis was very anticonservative, while Grammar was conservative. Moreover, results showed in both cases a loss of power when the heritability and population structure increased. This study provides the theoretical formulas of type I and type II errors for Regression and Grammar analysis. These formulas can be used for any trait and pedigree of a QTL detection specific protocol.

Aulchenko,Y.S., de Koning, D-J and Haley, C. (2007) Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics.177:577-585

Comparison of an improved method for calculating line origin probabilities against GridQTL using simulated data

Lucy Crooks[1*], Carl Nettelblad[2], Sverker Holmgren[2], Örjan Carlborg[3]

[1] Department of Cell and Molecular Biology, Uppsala University, Sweden
[2] Department of Information Technology, Uppsala University
[3] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences
∗ Presenting author: Lucy Crooks, email: Lucy Crooks@icm.uu-.se

A fundamental step in QTL analysis of line crosses is estimating the probability that chromosome regions originate from each of the founder lines. GridQTL (Seaton *et al.*, 2006) is a popular tool for calculating these probabilities. However, the algorithm employed by GridQTL is computationally inefficient and not designed for the large numbers of markers that are now being genotyped. We have presented an alternative method, cnF2freq (Nettelblad *et al.*, 2009) that achieves far greater efficiency using a hidden Markov model approach. Here, we compare the performance of the two methods on a series of replicates from simulated data. The simulation is based on dense SNP genotypes from a chicken intercross; the aim being to accurately reflect the patterns of marker informativeness that occur in real data. Each replication comprised one chromosome of about 1,500 SNPs in nearly 800 F_2 individuals derived from the same set of grandparents. The grandparent haplotypes were determined by approximate haplotyping of the actual founders and the pedigree structure of the real data was used. GridQTL states that it can analyse datasets containing up to around 1500 markers on a chromosome. With our data, it was unable to analyse a complete set of 10 replicates, even when the number of markers was reduced to 1,200. For 1,200 markers it was around 80 times slower than cnF2freq. In the analyses that worked, GridQTL failed to estimate probabilities for large sections of the chromosome, with about a third of individuals having missing values for most of the genome. For the positions where line origin probabilities were estimated by GridQTL, the accuracy of the results from GridQTL and cnF2freq were compared. The performance of cnF2freq was also evaluated over a more extensive set of 1000 replicates. Our results provide a validation of cnF2freq, show that GridQTL is unable to deal with datasets of the size generated by SNP chips and demonstrate that cnF2freq can handle substantially larger datasets than GridQTL.

Nettelblad C., Holmgren, S., Crooks, L., Carlborg, Ö. *Proc. 1st BICoB* (2009): 307-319.
Seaton G., Hernandez J., Grunchec, J. A., White, I. *et al. Proc. 8th WCGALP* (2006).

Genome wide evaluation using dominance

Robin Wellmann[1*], Jörn Bennewitz[1]

[1] Universität Hohenheim, Institut für Tierhaltung und Tierzüchtung, D-70599 Stuttgart, Germany
∗ Presenting author: Robin Wellmann, email: r.wellmann@uni-hohenheim.de

Background. The inclusion of dominance effects into models for the prediction of genomic breeding values could increase the accuracy of the predictions and dominance effects could be used to choose mating pairs with good combining ability. The effects of different amounts of dominance variance and inbreeding depression on the accuracy of predicted breeding values and dominance values is studied by simulation in a bottlenecked population.
The distribution of the QTL effects was heavy tailed. Simulated dominance degrees were normally distributed and independent from the additive effects. Multiple regression was used to predict the accuracy of breeding values and dominance values with the heritability, the dominance variance and the inbreeding depression as covariates.

Results and Conclusions. Inbreeding depression decreased the accuracy of predicted breeding values but it could well be utilized to predict dominance values. The inclusion of dominance increased the accuracy of the breeding values.

Genome wide effects of divergent selection for body weight in chickens

Anna M. Johansson[1*], Mats E. Pettersson[1], Paul B. Siegel[2], Örjan Carlborg[1]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-75007 Uppsala, Sweden
[2]Virginia Polytechnic Institute and State University, Department of Animal and Poultry Sciences, Blacksburg, VA 24061-0306, USA
*Presenting author: Anna Johansson, email: Anna.Johansson@hgen.slu.se

To understand the genetic mechanisms leading to phenotypic differentiation, it is important to identify regions in a genome that are under selection. Here, we perform a genome wide scan using a 60k SNP chip in two chicken lines from a single trait selection experiment, where 40 generations of selection have resulted in a nine-fold difference in body weight. We analysed individuals after 40 and 50 generations of selection in both the high body weight line and the low body weight line. We show that the effect of selection is as dramatic on the genome as on the phenotype. More than 60 regions were identified where selection has fixed alternative alleles in the two lines. Many more regions with highly significant large allele frequency differences were present all across the genome. Another 10 regions displayed strong evidence for ongoing selection during the last 10 generations.

Using genome scans of DNA polymorphism to identify regions exhibiting positive selection

S. Qanbari[1*], D. Gianola[2], B. Hayes[3], F. Schenkel[4], S. Miller[4], S. Moore[5], G. Thaller[6], H. Simianer[1]

[1] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

[2] Department of Animal Sciences and Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA

[3]Animal Genetics and Genomics, Primary Industries Research Victoria, 475 Mickleham Rd, Attwood, VIC 3049, Australia.

[4] Centre for Genetic Improvement of Livestock, Animal and Poultry Science Department, University of Guelph, Guelph, Ontario, N1G 2W1 Canada

[5] Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada

[6] Institute of Animal Breeding and Animal Husbandry, Christian-Albrechts-University, 24098 Kiel, Germany

∗ Presenting author: Saber Qanbari, email: sqanbar@gwdg.de

In this study, two different and complementary methods were applied to identify traces of decades of intensive artificial selection for traits of economic importance in modern cattle. We scanned the genome of a diverse set of dairy and beef breeds from Germany, Canada and Australia with a 50K SNP panel. In the first study we employed the integrated Haplotype Homozygosity Score (|iHS|) for tracing on-going sweeps. Across breeds, a total of 109 extreme |iHS| values exceeded the empirical threshold level of 5% with 19, 27, 9, 10 and 17 outliers in Holstein, Brown Swiss, Australian Angus, Hereford and Simmental, respectively. Annotating the regions harboring clustered |iHS| signals to the genome revealed significant enrichment for functional genes like SPATA17, MGAT1, PGRMC2 and ACTC1, COL23A1, MATN2, respectively, in the context of reproduction and muscle formation. In the second analysis a new Bayesian F_{ST}-based approach was applied with a set of geographically separated populations including Holstein, Brown Swiss, Simmental, Canadian Angus and Piedemontese for detecting differentiated loci. The algorithm used is able to process a large battery of marker information and allows inference based on the posterior distribution of F_{ST} values. In total, 127 regions exceeding the 2.5 per cent threshold of the empirical posterior distribution were identified as extremely differentiated. To a substantial proportion (56 out of 127 cases) the extreme F_{ST} values were found to be positioned in poor gene content regions which deviated significantly ($p < 0.05$) from the expectation assuming a random distribution. However, significant F_{ST} values were found in some gene regions, like SMCP and FGF1. A remarkable observation in this study is a selection signal confirmed by both |iHS| and F_{ST} analyses in the vicinity of Sialic acid binding Ig-like lectin 5 gene on BTA18 which recently was reported as a major QTL on productive life and fertility traits in Holstein cattle. Overall, based on the results of this study we conclude that high-resolution genome scans are capable to identify outlier regions that potentially contain genes contributing to within and inter-breed phenotypic variation.

Genome-wide scans for quantitative trait loci in experimental populations - issues of multiple testing and model selection

Małgorzata Bogdan[1*]

[1] Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland
* Presenting author: Małgorzata Bogdan, email: Malgorzata.Bogdan@pwr.wroc.pl

There exist a variety of methods for locating multiple interacting quantitative trait in experimental population. Most of these methods are based on linear models, relating trait values to the genotypes of potential QTL. The most difficult part in selecting the best linear is the estimation of the number of QTL. This task could in principle be addressed by the application of one of the model selection criteria, like e.g. AIC or BIC. However, as observed in [1], BIC (and so AIC) has a strong tendency to overestimate QTL number. In [2] and [3] it is explained that this phenomenon is related to the well known issue of multiple testing. Also, in the Bayesian context, the overestimation can be explained by the undesired properties of the uniform prior on the class of possible models, implicitly used by BIC. To address the problem of overestimation, some modified versions of BIC have been proposed. Two of these versions, modified BIC (mBIC,[4]) and extended BIC (EBIC,[5]), are based on the Bayesian principles. BIC is extended by using different prior distributions on the class of possible models. As shown in a series of papers, the priors used in mBIC and EBIC allow to overcome the problem of overestimation and both these criteria have very good properties as applied for QTL detection.

In this talk we will give a brief overview of the problems of high dimensional model selection and their relationship to multiple testing. We will also present mBIC and EBIC and the results of their application to the problem of localizing QTL influencing the count data. If time permits, we will also briefly discuss the application of model selection tools in the context of Genome Wide Association Studies in human populations.

[1] Broman K, Speed T. A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Roy. Stat. Soc. B., 2002, 64: 641--656.

[2] Bogdan M,Ghosh JK,Żak-Szatkowska M. Selecting explanatory variables with the modified version of the Bayesian Information Criterion. QREI, 2008, 24:627--641.

[3] Bogdan M,Frommlet F, Biecek P, Cheng R, Ghosh JK,Doerge RW. Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping. Biometrics, 2008, 64: 1162--1169.

[4] Bogdan M, Ghosh JK, Doerge RW. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics, 2004, 167: 989-999.

[5] Chen J, Chen Z. Extended Bayesian Information criteria for model selection with large model spaces. Biometrika, 2008, 95: 759--771.

Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals

Matthew A. Cleveland[1*], John M. Hickey[2,], Brian P. Kinghorn[2]

[1] Genus/PIC, 100 Bluegrass Commons Blvd. Suite 2200, Hendersonville, TN 37075, USA
[2] School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia
∗ Presenting author: Matthew Cleveland, email: matthew.cleveland@pic.com

Background. There is wide interest in calculating genomic breeding values (GEBVs) in livestock using dense, genome-wide SNP data. The general framework for genomic selection assumes all individuals are genotyped at high-density, which may not be true in practice. Methods to add additional genotypes for individuals not genotyped at high density have the potential to increase GEBV accuracy with little or no additional cost. In this study a long haplotype library was created using a long range phasing algorithm and used in combination with segregation analysis to impute dense genotypes for non-genotyped dams in the training dataset (S1) and for non-genotyped or low-density genotyped individuals in the prediction dataset (S2). Alternative low-density scenarios were evaluated for accuracy of imputed genotypes and prediction of GEBVs.

Results. In S1, females in the training population were not genotyped and prediction individuals were either not genotyped or genotyped at low-density (evenly spaced at 2, 5 or 10cM). The accuracy of imputed genotypes for training females did not change with the addition of genotypes in the prediction set, as expected, whereas the number of correctly imputed genotypes in the prediction set increased slightly. S2 assumed the complete training set was genotyped for all SNPs and the prediction set was not genotyped or genotyped at low-density. The number of correctly imputed genotypes increased with genotyping density in the prediction set. Genomic breeding values for the prediction set in each scenario were correlated with predicted GEBVs when all animals were genotyped for all SNPs to evaluate the potential loss in accuracy with reduced genotyping. For both S1 and S2 the correlations with the high-density GEBVs were similar when the prediction set was not genotyped and increased with the addition of low-density genotypes, where the increase was larger for S2 than S1.

Conclusions. Genotype imputation using a long haplotype library and segregation analysis has promise for application in sparsely-genotyped pedigrees. The results of this study suggest that dense genotypes can be imputed for selection candidates with some loss in GEBV accuracy compared to the high-density case, but genotyping strategies for the training set may be needed to maintain accuracy when using low-density SNP panels for genomic selection.

A comparison of random forests, boosting and support vector machines for genomic selection with SNP markers

Joseph O. Ogutu[1*], Hans-Peter Piepho, Torben Schulz-Streeck[1]

[1] Bioinformatics Unit, Institute for Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

∗ Presenting author: Joseph Ogutu, email: jogutu2007@gmail.com

Background. Genomic selection (GS) is a method for estimating breeding values using molecular markers spanning the entire genome. It involves jointly estimating the effects of all genes or chromosomal segments and combining the estimates to predict the total genomic breeding value (GBV). Accurate prediction of GBVs thus represents a central challenge to contemporary plant and animal breeders. However, the existence of a wide array of approaches for predicting breeding values using markers makes it essential to evaluate and compare their relative predictive performances to identify approaches able to accurately predict breeding values. Here, we compare the predictive performances of three key machine learning methods, namely random forests; boosting and support vector machines, for predicting GBV using SNP-marker data. We also explore the utility of random forests for importance ranking of markers prior to mapping their chromosomal positions.

Methods. We predicted GBV for one quantitative trait in a simulated data set of 3226 individuals spanning five generations and an associated genome encompassing five chromosomes with 10031 biallelic SNP-marker loci generated for the 14th QTL-MAS workshop. Of the 3226 individuals 2326 in the first four generations were phenotyped and genotyped and were used to train the random forest, boosting and support vector machine regression models. The models were used to predict GBV for the remaining 900 individuals in the fifth generation without phenotypic records. A 5-fold cross-validation was used to evaluate the predictive performance of each method. Predictive accuracy was measured by mean squared error and the correlation between GBV and the simulated values. Random forest was used to rank the importance of markers and positions of the most important markers mapped on the pertinent chromosomes.

Results. The correlation between the predicted and simulated breeding values for the 5-fold cross validation for random forests was modest and ranged from 0.393 to 0.5878 and the mean squared error from 65.47 to 70.71. For support vector machines the mean squared error for the 5-fold cross-validation ranged from 196.4599 to 236.4377.

Conclusions. The support vector machine predicted GBV better than random forests and boosting but our evaluation of the performance of boosting is still ongoing.

Genomic selection using Best Linear Unbiased Prediction with a trait specific relationship matrix

Zhe Zhang[1,2*], Xiangdong Ding[1], Jianfeng Liu[1], Dirk-Jan de Koning[2], Qin Zhang[1]

[1] Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China

[2] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, EH25 9PS, UK

∗ Presenting author: Zhe Zhang, email: zhangzhecau@126.com

Background. Using the common dataset published from the XIV QTL-MAS workshop, three methods for estimating genomic breeding values (GEBVs) of a quantitative trait were compared: best linear unbiased prediction with a trait specific relationship matrix (TA-BLUP), ridge regression best linear unbiased prediction (RR-BLUP), and BayesB.

Methods. TA-BLUP is a method that is identical to the conventional BLUP method except that the numeric relationship matrix is replaced with a trait specific relationship matrix (TA). The TA matrix is constructed based on both marker genotypes and their estimated effects on the trait of interest. The marker effects were estimated in a reference population consisting of 2,326 individuals using RR-BLUP and BayesB and the GEBVs of individuals of the reference population as well as 900 young genotyped individuals were estimated using the three methods. Subsets of markers were selected to perform low-density marker genomic selection for TA-BLUP method.

Results. The correlations between GEBVs from different methods are over 0.95 in most cases. The correlation between BayesB and TA-BLUP using 200 selected markers to construct the TA matrix is 0.98 in the candidate population.

Conclusions. TA-BLUP provides the possibility to use low-density markers for estimating GEBVs in candidate populations. Using only a small proportion of the total markers, selected on the basis of their estimated effects, TA-BLUP gives GEBVs that are almost equivalent to that from RR-BLUP or BayesB. TA-BLUP is therefore a promising method for genomic selection in the candidate populations because of the remarkably reduced cost for genotyping, even though there might be a little loss of accuracy.

Applying different genomic selection approaches on QTL-MAS 2010 data

Javad Nadaf [1*], Ricardo Pong-Wong[1]

[1] The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK
* Presenting author: Javad Nadaf, email: javad.nadaf@roslin.ed.ac.uk

Background. Four different genomic evaluation methods were applied on the QTL-MAS 14th dataset which includes 3226 individuals across 4 generations, all genotyped for 10031 SNPs on 5 chromosomes. The genomic selection analyses included two Bayes B type of methods (BB): using only SNP information (GBB) or SNP and Pedigree information (GPBB); and two GBLUP and GPBLUP. When using BB methodology, the probability of SNP having an effect on the traits (which include a quantitative and a binary trait) was also estimated. Supplementary analyses were also done, including association analysis and QTL mapping, to have a better interpretation of the results.

Results. The polygenic heritability of the traits was estimated at 0.55 and 0.42, for the quantitative and the binary trait respectively. The correlation of EBVs between the two traits (a good estimation of genetic correlation) was estimated at 0.58. A good consistency was obtained between different methods in the estimation of EBVs: the correlation between EBVs of 900 juveniles (the 4th generation), estimated by different approaches were at least 0.90, for the two traits; when considering all individuals, this number was as high as 0.94. The correlation was less affected by adding polygenic effect in the model than changing the methodology (BB vs. GBLUP); however, adding polygenic effect was still important, at least for the quantitative trait. The percentage of SNPs with an effect on the traits was estimated at about 5 and 10% for the quantitative and binary trait respectively, using the GBB method. The results of BB analyses were consistent with association and QTL mapping analyses: The SNPs with the most important effects were common and positioned on chromosome 1 and 3. There was also good consistency regarding the SNPs with moderate effects.

Conclusions. An important part of additive genetic effects, estimated for both traits, could be captured by genomic information. Using this source of information, the results obtained by the 4 different approaches were quite consistent. Good consistency was also observed between genomic selection (BB) and association and QTL mapping analyses, concerning the effect of SNPs on traits.

Pre-selection of markers for genome-wide selection

Torben Schulz-Streeck[1*], Joseph O Ogutu[1], Hans-Peter Piepho[1]

[1]Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

∗ Presenting author: Torben Schulz-Streeck, email: torben.schulz-streeck@uni-hohenheim.de

Background. Accurate prediction of genomic breeding values (GEBVs) requires numerous markers. However, prediction accuracy can be enhanced by excluding markers with no effects or markers with inconsistent effects among crosses that can adversely affect the prediction of GEBVs.

Methods. We present three different approaches for pre-selecting markers prior to predicting GEBVs and assess the extent to which pre-selection of markers improves prediction accuracy.

1. Testing each SNP using a linear regression model similar to Macciotta et al. (2009)
2. Analyzing each SNP for consistency among crosses using mixed models
3. Analyzing each SNP for consistency among generations using mixed models

We predicted GEBVs for the single quantitative trait in the common dataset provided for the 14th QTL-MAS workshop using four different BLUP methods, including ridge regression and three geostatistical models. Performances of the models were evaluated using four different model selection criteria plus 5-fold cross-validation.

Results and conclusions. Ridge regression and the geostatistical models gave almost similar fits. Pre-selecting markers was evidently beneficial since excluding markers with inconsistent effects among crosses increased the correlation between GEBVs and observed values in validation datasets from 0.530 (using all markers) to 0.584 (using pre-selected markers).
However, extension of the ridge regression model to allow for heterogeneous variances between the n (n =5, 10, 50, 100, 250) most significant markers and the remaining markers only marginally increased the accuracy of prediction (from 0.584 to 0.587).
Results produced using models selected by AIC and GCV were nearly consistent with those for models selected by the 5-fold cross-validation, implying that model selection criteria such as AIC and GCV may be used instead of cross-validation to reduce computing time.
We submit results from the following two final models:

1. Linear spatial model with the 1000 most significant markers selected by *method 2*
2. Ridge regression with the 1000 most significant marker selected by *method 2* and heterogeneous variance between the 50 most significant markers and the remaining markers

Macciotta NPP, Gaspa G, Steri R, Pieramati C, Carnier P, Dimauro C: Pre-selection of most significant SNPS for the estimation of genomic breeding values. BMC Proc 2009, 3(Suppl 1):S14

Methods for genome-wide association analyses of quantitative trait loci in human genetically isolated populations

Yurii Aulchenko[1*]

[1] Department of Epidemiology, Rotterdam, The Netherlands
[*] Presenting author: Yurii Aulchenko, email: i.aoultchenko@erasmusmc.nl

Humans are one of the worst genetic model objects in the sense of possible applicable experimentation techniques, maintainability and generation time. On the other hand, there is a huge interest in studying the determinants of the health and disease of humans, leading to large investments and possibilities to apply top technology to challenge the questions about humans.

In this talk, I will discuss the problem of identification of genomic loci whose variation is responsible for the variance in complex traits in human populations. In particular, I will give an overview of methodology, implementation of different variants of linear mixed models (FASTA, GRAMMAR, etc.). The uses and limitations of this approach will be outlined, and practical examples provide.

Implications of previous knowledge to new generation of molecular data (e.g. HT re-sequencing, -'omics'), general uses of linear mixed models, and discussion of what are the ways to go further with the genetics of complex (human) traits will conclude the talk.

A new R package for QTL analysis

Ronald Nelson[1*], José Álvarez-Castro[1], Lucy Crooks[2], Francois Besnier[1], Carl Nettelblad[3], Lars Rönnegård[1], Xia Shen[2], Örjan Carlborg[1]

[1] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden
[2] Department of Cell and Molecular Biology, Uppsala University, Sweden
[3] Department of Information Technology, Uppsala University, Sweden
∗ Presenting author: Ronald Nelson, email: ronnie.nelson@hgen.slu.se

The availability of high-throughput methods for molecular characterization of individuals means that many individuals can be typed for thousands of markers. It is therefore necessary to develop new algorithms to handle these large datasets when analyzing complex traits. We are developing a computationally efficient, easy to use, well-documented, freely available and flexible software package for the analysis and mapping of complex traits in inbred and outbred line-crosses. The package will be developed with a front-end in the R-language. Several modules have already been developed and tested. These include: cnF2freq for calculating genotype probabilities from line-crosses (including outbred lines) typed for thousands of markers; FIA for variance component based mapping of QTL including epistasis; MCIBD, for determining identity by descent from pedigree data; and NOIA, for modeling epistasis. We also have modules to analyze and haplotype individuals from deep pedigrees as well as optimized user-friendly algorithms implemented in the R-language for regression based analysis. The software will enhance QTL analysis by allowing the inclusion of substantially more markers than current software can handle, increasing the accuracy of mapping single and interacting QTL by including variance component analysis and extracting additional information from deep pedigree crosses, if available. This package will be an important resource for utilizing current, and future, data within the framework of genetic analysis of complex traits.

The estimation of SNP effects on a binary and a quantitative trait

Kacper Żukowski[1,*], Joanna Szyda[1,2]

[1] Department of Genetics and Animal Breeding, Wrocław University of Environmental and Life Sciences, Kożuchowska 7, 51-631 Wrocław, Poland

[2] Institute of Natural Sciences, Wrocław University of Environmental and Life Sciences, Norwida 25, 50-375 Wrocław, Poland

∗ Presenting author: Kacper Żukowski, email: kacper.zukowski@up.wroc.pl

Background **.** The analysis concerns on estimating the effects of Single Nucleotide Polymorphisms (SNP) on a binary and a quantitative trait, based on the simulated QTL-MAS workshop data. Moreover, the association between the two traits is considered.

Methods. For a quantitative trait a gBLUP model was applied $y = \mu + Za + \varepsilon$, where: y is a quantitative trait, Z is an incidence matrix for SNP effects, a is a vector of random SNP effects assuming $a \sim N\left(0, I\frac{\sigma_\alpha^2}{n}\right)$ with n representing the number of SNPs, ε is a vector of random errors assuming $e \sim N\left(0, I\sigma_\varepsilon^2\right)$. This model was applied for all animals, for affected animals only, and for unaffected animals only. Further on, a modification of the gBLUP model with SNP effects estimated separately for affected and unaffected animals (CgBLUP) was used: $y = \mu + Za_{unaff} + Za_{aff} + \varepsilon$, where a_{unaff} and a_{aff} are random vectors for unaffected and affected animal respectively, both effects were assumed to be uncorrelated. For a binary trait a logistic regression model (LOGIT) was used: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_i Z_i$, where p is a probability of being affected, β_0 is an intercept and β_i represents an additive effect of SNP i with a corresponding column in incidence matrix Z_i.

Results. The most significant SNPs for both traits were found on chromosome 1 and chromosome 3. Two most significant SNPs influencing both, a binary and a quantitative trait: #4480 (22 030 629 bp) and #4491 (22 520 568 bp) were located on chromosome 3.
Generally, similar locations were significant for both traits, but no significant SNPs effects were reported by CgBLUP model.

Conclusions. The analysis showed many significant SNPs, one gene located on 22,0-22,5 Mb on chromosome 3 could have influence on a quantitative and a binary trait simultaneously.

Logic regression based methods in application to detection of gene-gene interactions

Magdalena Malina [1*], Małgorzata Bogdan[2]

[1]Mathematical Institute, University of Wrocław
[2]Institute of Mathematics and Computer Science, Wrocław University of Technology,
∗ Presenting author: Magdalena Malina, email: malina@math.uni.wroc.pl

Background. We consider a biological problem of locating multiple interacting QTLs. Logic regression introduced in [2] by Ruczinski, Kooperberg, LeBlanc is a regression method, specifically aimed at detecting high order interactions in SNP data. The method attempts to construct predictors as Boolean combinations of dummy variables coding marker genotypes. The method searches for these combinations of predictors in the entire space of such combinations, which are represented by logic binary trees. In [1] Schwender and Ickstadt propose a procedure called logicFS (logic feature selection) in which the "best" model of a fixed dimension is estimated with simulated annealing, aimed at maximizing the likelihood. The instability of binary trees is overcome by using bootstrap aggregation. Based on bootstrap replications, for each interaction an importance measure is calculated. The method can be applied for both binary and quantitative responses.

Results. We applied LogicFS for QTL mapping. We analyzed two real QTL data sets (see [3], [4]). The data by Lyons et.al ([3]) contains different phenotypes related to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. We searched for interactions that influence gallbladder bile characteristics (binary variable). The second data set ([4]) considers different phenotypes related to obesity detected in an intercross of SM/J and NZB/BINJ inbred mice. Here we searched for interactions that influence gonadal fat pads weight. For both data sets logicFS was able to detect several interactions which were not reported in original papers.

Conclusions. Logic regression models allow to present very complex linear models with many parameters in a much simpler form. Increasing number of components that may be included to models of this form causes also that the problem of multiple testing become more complex. In such models however it is easier to get to the higher order interactions without addition of new parameters. Interactions included in such model are also much easier to interpret. According to our results, logicFS can be successfully applied to identification of interacting QTLs.

[1]Schwender, H. and Ickstadt, K. (2008). Biostatistics, 9. 187-198
[2] Ruczinski I., Kooperberg C., LeBlanc M. (2003). J. Comput. Graphical Statist. 12 (3) 474-511
[3] Lyons MA, Wittenburg H, Li R, Walsh KA, Leonard MR, Churchill GA, Carey MC, Paigen B. Physiol Genomics. 2003 Aug 15;14(3):225-39. PMID 12837957
[4] Stylianou IM, Korstanje R, Li R, Sheehan S, Paigen B, Churchill GA. Mamm Genome. 2006 Jan;17(1):22-36. PMID 16416088

Genetic architecture of yield and related traits in European maize: insights into the effects of linkage and allelic series. Consequences for marker assisted selection

A. Charcosset [1*], G . Blanc [1], Y.-F. Huan [1], A. Gallais [1], L. Moreau [1]

[1] INRA-CNRS-UPS-AgroParisTech Station de Génétique Végétale, Ferme du Moulon, 91190 Gif sur Yvette, France

∗ Presenting author: Alain Charcosset, email: charcos@moulon.inra.fr

Background. QTL mapping experiments conducted so far in plants generally considered populations derived from the cross between two inbred lines. They have proven efficient to detect QTL involved in trait variation. However, due to a limited number of effective recombination generations, they provide a poor resolution, with QTL confidence intervals often exceeding 20 cM for QTL of small effect and usual population sizes (between 100 or 200 individuals). They also address only a limited fraction of the genetic variability available in a breeding program. Alternative QTL experimental designs such as advanced intermated populations or multiparental populations, have been proposed to overcome these drawbacks.

Results. To evaluate the first strategy, we conducted a QTL detection experiment in (i) an advanced intermated F_3 population of 322 lines and (ii) a conventional F_3 population of 300 lines. Both were derived from the same parental maize inbred lines, with four additional generations of intermating for (i), and jointly evaluated in testcross progeny for dry grain yield and related traits. QTL confidence intervals were on average 2.31 shorter in the intermated population than in the conventional population. However, fewer QTLs were detected in the intermated population and less than 50% of the detected QTLs were common to the two populations. Cross-validation showed that selection bias was more important in the intermated population. Results suggest that the a substantial fraction of the QTLs detected for Grain moisture at harvest and Grain Yield in the conventional population were actually due to coupling phase between small effect QTL. Regarding the second strategy, we conducted a QTL detection experiment in six F2 populations (150 individuals each) derived from a half-diallel between four European maize inbred lines. The joint analysis of the whole design increased the power of QTL detection and its resolution. It revealed stronger epistatic effects for yield than for other traits. It also revealed allelic series with gradual effects for most QTL and allowed us to identify the parental origin of the most favorable alleles. This information was used to conduct rapid cycles of selection based only on marker evaluation. Phenotypic evaluation of plants selected at each generation showed a significant genetic gain through cycles of selection. Interest of applying marker assisted selection to such multiparental designs was further supported by simulations.

Conclusions. These results illustrate that yield related traits in maize have a complex genetic architecture, involving several tens of loci, linkage situations affecting apparent QTL effects and possibly genetic variance, and allelic series with gradual effects. They also show that strategies, such as genomic selection, addressing a broad diversity and using present high-throughput genotyping technologies should prove highly beneficial to accelerate genetic gain for such traits.

Mapping systemic scleroderma genes in a cross between UCD200 and jungle fowl chickens

Weronica Ek[1*], Anna-Stina Sahlqvist[2], Roswitha Sgonc[3], Hermann Dietrich[3], Georg Wick[3], Olov Ekvall[4], Leif Andersson[1,5], Örjan Carlborg[1], Olle Kämpe[2], Susanne Kerje[5]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.
[2]Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden
[3]Division of Experimental Pathophysiology and Immunology, Biocenter, Innsbruck Medical University, Innsbruck, Austria.
[4]Department of Rheumatology och Inflammation Research, The Sahlgrenska Academy, Gothenburg, Sweden
[5]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden.
*Presenting author: Weronica Ek, email: weronica.ek@hgen.slu.se

Background. Systemic Scleroderma (SSc) is a rare autoimmune disease characterized by fibrosis of skin and internal organs, such as gastrointestinal tract, lungs kidneys and heart. Scleroderma is a complex disease and evidence suggests that genetic factors may be involved and that immune-regulation plays a major role. University of California at Davis (UCD) line 200 chickens develops a heritable systemic scleroderma, similar to the one in humans, and is therefore important as an animal model for the study of human scleroderma.

Method. We have generated a backcross between UCD200 and the red jungle fowl (control) chickens and performed a genome-wide QTL scan to identify loci underlying the disease. Two approaches were used to analyze the data. Generalized linear regression on line-origin QTL genotype probabilities calculated using cnF2freq (Nettelblad *et al*, 2009) and Flexibel Intercross Analysis (FIA), a variance component model that allows for variation within lines (Rönnegård *et al*, 2007). Scleroderma was scored as a binary trait, affected or not.

Results. The analysis confirmed one significant QTL on chromosome 14 (OR 2.7, $p_{Genome\text{-}wide} < 0.05$) and one suggestive QTL on chromosome 2 (OR 2.8, $p_{Genome\text{-}wide} = 0.06$). Birds phenotyped for early comb necrosis, scored at 3-weeks of age, showed a significant peak on chromosome 12 (OR 2.9, $p_{Genome\text{-}wide} < 0.05$). Scleroderma was more frequent in males (94% affected) then in females (25% affected) at 175 days of age.

C. Nettelblad, S. Holmgren, L. Crooks, O. Carlborg (2009) cnF2freq: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations using Markov Models. *Proceedings to BICoB* 2009, New Orleans, LNBI 5462, pp. 307-319, Springer Verlag Berlin

Rönnegård, L., Besnier, F., Carlborg, Ö. 2007. An improved method for quantitative trait loci detection and identification of within-line segregation in F2 intercross designs. *Genetics* 178: 2315-2326.

Genomic selection in Polish Holstein

Joanna Szyda[1*], Andrzej Żarnecki[2], Stanisław Kamiński[3]

[1]Department of Animal Genetics, Institute of Natural Sciences, Wrocław University of Life and Environmental Sciences, Wrocław, Poland
[2]National Research Institute for Animal Production, Balice, Poland
[3]Department of Animal Genetics, University of Warmia and Mazury, Olsztyn, Poland
∗ Presenting author: Joanna Szyda, email: joanna.szyda@up.wroc.pl

Background. Recently many countries incorporated the genomic information, in the form of thousands of Single Nucleotide Polymorphisms (SNP) genotypes originating from a microarray technology, into their genetic evaluation systems. We describe the results of fitting a genomic evaluation model to the Polish population of Holstein Friesian cattle, as well as the future activities related to genomic selection in this breed.

The data set, used as a training data set for the estimation of additive effects of SNPs in the genomic evaluation model, consisted of 1,227 Polish Holstein-Friesian bulls, born between 1987 and 2003. Genotypes were generated by the use of the Illumina BovineSNP50 Genotyping BeadChip, which consists of 54,001 SNPs. For the estimation of Direct Genomic Values (DGV) 46,267 SNPs were selected based on the minor allele frequency and call rate, giving 56,502,470 bull-SNP genotypes in total for milk yield. Direct genomic values (DGV) were calculated for 28, comprising three production traits: milk-, fat- and protein- 305-day yields, somatic cell score, three fertility traits, and 21 traits describing conformation. For this purpose a mixed linear model with a random effect of additive SNP effect was applied to deregressed national proofs (EBV) based on daughter production data available in 2009.

Results. For bulls from the training data set correlations between EBV and DGV are high, varying from 0.98 for milk yield to 0.83 for rear leg rear view - the trait with the lowest heritability (0.04). The average reliability of DGV also varied across the analyzed traits, ranging from 0.573 for foot angle to 0.940 for milk yield. The average reliability of DGV for young selection candidates is however lower: 0.228 for milk- and 0.216 for fat- and protein yields.

Conclusions. Thousands of SNP genotypes densely distributed along the genome for many individuals with well defined phenotypes and familial relationships provide very valuable piece of information. From the point of view of animal breeders, the genomic selection is a powerful, internationally recognized selection tool. Moreover, from a geneticist perspective estimation of SNP effects densely distributed along the genome provides a unique opportunity for detection of causal mutations and interplay between them.

Validation experiences in Italian Holstein genomic selection

J.B.C.H.M. van Kaam[1*], R. Finocchiaro[1], F. Canavesi[1], G.B. Jansen[2]

[1]ANAFI - Italian Holstein Association, Via Bergamo 292, 26100 Cremona, Italy
[2]Dekoppel Consulting, Casale Rovera 10, 10010, Chiaverano, Italy
* Presenting author: Jan-Thijs van Kaam, email: jtkaam@anafi.it

Background. In Italy, several projects (SelMol, ProZoo and Elica) have been set up, which contribute to the creation of a reference data set. SNP marker data were recoded, merged or rejected when needed and then checked to determine which SNPs and which bulls to retain. Data editing resulted in a reduction from 2694 samples on 2613 Italian Holstein bulls to 2568 bulls and from 54,001 SNPs to 39,259 SNPs. Deregressed proofs were obtained by full pedigree deregression. Individual SNP effects were estimated with genomic BLUP using residual updating. Direct genomic values resulted from summing SNP allelic effects. Various validation runs with 12 traits have been undertaken.

Results. For random effects in a mixed model a variance ratio, between residual and explanatory effect, is needed. Rough estimates of Ve and Vg were obtained by partitioning the variance of deregressed proofs in residual and genetic components, using heritabilities and average effective daughter contributions derived from average reliabilities. For the marker effects, a variance ratio of Ve/Vm with Vm=Vg/sum2pq, i.e. marker effects are proportional to the informativity of the marker, was used. Results of regressing direct genomic values on deregressed proofs show that the regression coefficient was clearly below 1 for all traits. Using larger variance ratios, i.e. regressing marker estimates to lower values, resulted in larger regression coefficients with slightly lower R^2. A variance ratio multiplied by 5 gave regression coefficients near 1.

Conclusions. Results indicated that R^2 depends mainly on the number of phenotypes, which was expected due to the excess of explanatory variables compared to response variables. More lax or stringent selection of SNPs only affected the 3rd decimal of the R^2. The variance ratio needed to be higher than expected to obtain regression coefficients near 1.

Tutorial on R/qtl

Karl W. Broman[1*]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison
∗ Presenting author: Karl Broman, email: kbroman@biostat.wisc.edu

R/qtl is an extensible, interactive environment for mapping quantitative trait loci (QTL) in experimental crosses. It is implemented as an add-on package for the freely available and widely used statistical language/software R (www.r-project.org). The development of this software as an add-on to R allowed us to take advantage of the basic mathematical and statistical functions, and powerful graphics capabilities, that are provided with R. Further, the user benefits by the seamless integration of the QTL mapping software into a general statistical analysis program. Our goal was to make complex QTL mapping methods widely accessible and allow users to focus on modeling rather than computing.

In this tutorial, we will briefly demonstrate a portion of the features of R/qtl (for those that wish to follow along, see http://www.rqtl.org/qtlmas2010).

POLAPGEN-BD: a project on biotechnology for breeding cereals with increased resistance to drought#

Paweł Krajewski[1*], Anetta Kuczyńska[1] and POLAPGEN Consortium[2]

[1]Institute of Plant Genetics, Polish Academy of Sciences, Strzeszyńska 34, 60-479 Poznań
[2]POLAPGEN Consortium for Applied Genetics and Genomics, Strzeszyńska 34, 60-479 Poznań, www.polapgen.pl
∗ Presenting author: Paweł Krajewski, email: pkra@igr.poznan.pl

The subject of the project is drought resistance in cereals, investigated in spring barley treated as both a model and economically important plant. Increasing desiccation of the environment, reflecting the water deficit in the soil, requires tools that would enable the breeders to carry out selection of genotypes resistant to drought. The project consists of 23 research tasks carried out under POLAPGEN Consortium whose partners are 10 research units and 2 breeding companies. Tasks were formulated in such a way that a wide range of characteristics determining plants' resistance to drought will be studied. All tasks will be realized on the same plant material, which enables integration of many research teams around the same problem solved in a modern way. Comprehensive approach to the problem of cereal resistance to shortage of water will enable to evaluate interdependence of various parameters determining that characteristic. System approach will be achieved by adopting a model of tolerance of plants to drought stress containing ecophysiological, morphological, anatomical, metabolic, proteomic, and molecular levels considered in the context of genetics. Selected parameters of the model and interdependence between them will be evaluated. The project will allow to set new markers for resistance to drought, both molecular and morphological, as well as new methods of evaluation of resistance based on physiological and physical indicators. Data obtained will also allow the creation of the ideotype of resistant varieties, characterized by the complex of characteristics and properties at the whole plant level and at the cellular and molecular level.

Project carried out under Innovative Economy Programme 2007-2013, Action 1.3, Subaction 1.3.1. within the subject „Biological progress in agriculture and environment protection".

Genomic estimated breeding values and QTL positions: Common dataset of QTL-MAS 2010 Workshop

X. Sun[1*], J. C. M. Dekkers[1]

[1]Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, Iowa 50011

*Presenting author: Xiaochen Sun, email: xsun1120@iastate.edu

We used a linear model with sex as fixed effects and random SNP effects to fit the dataset provided by QTL-MAS 2010 Workshop and applied Bayesian analysis to obtain genomic estimated breeding values (GEBVs) and QTL positions.

GEBV of each individual in Generation 5 was the sum of genetic effects of all SNPs it carried, where effects of SNPs were estimated from phenotypic and genotypic records in the first four generations. The accuracy of our Bayesian method was evaluated by the correlation of phenotypes of individuals in the fourth generation with their GEBVs and divided by the square root of estimated heritability from REML, which is equivalent to the correlation of GEBVs with true breeding values. The accuracy estimated this way could be as high as 0.8. However, since the individuals in the fifth generation did not have phenotypic records, the accuracy of their GEBVs was not available.

The bins spanning every 10 adjacent SNPs that explained genetic variances higher than a predefined threshold were selected as candidate genome regions containing QTLs. Within each bin, the SNP explaining the highest genetic variance was used as the position of QTL. To set a threshold for genetic variances of bins, we simulated genotypes without QTL for individuals in the first four generations using the same pedigree and SNP panel as the common dataset. Simulated SNP effects and genetic variances of bins were estimated from simulated genotypes and original phenotypes using the same Bayesian method. The threshold of certain significance level was determined from the distribution of genetic variances explained by the bins across the whole genome. Since the threshold was arbitrary, we reported a primary QTL list together with a secondary list, where the threshold of the former was stringent and the latter more liberal.

SPONSORS

illumina®