Livestock Science 166 (2014) 232–248

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities ☆

Haja N. Kadarmideen *

Section of Animal Genetics, Bioinformatics and Breeding, Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 7, 1870 Frederiksberg C, Denmark

ARTICLE INFO

Keywords:
sgBLUP
Genomic prediction
Systems genetics
Systems biology
eQTL
RNAseq

ABSTRACT

Livestock genomics has gone through a paradigm shift since the advent of genome sequencing that includes Genome-Wide Association Studies (GWAS), Whole Genome Predictions (WGP) and Genomic Selection (GS). Beginning with a brief review of current progress and challenges in livestock GWAS, WGP and GS, opportunities for next generation methods are introduced that unravel the underlying systems genetics of complex traits and provide biologically meaningful and accurate predictions. Genome-Wide Epistasis Association (GWEA) and Weighted Interaction SNP Hub (WISH) network methods are introduced here to unravel complex trait genetics. These methods effectively address the problems of GWAS that have no ability to model and analyze genome-wide genetic interactions and thus do not capture any epistatic variance that could explain part of the missing heritability. Further, the *Systems genomic* BLUP (*sg*BLUP) prediction method is introduced in this paper as a next generation WGP or GS tool that can account for and differentiate SNPs with known biological roles in the phenotypic or disease outcomes and potentially increase the accuracy of prediction. It is emphasized that tools that link genetic variants to their functions, pathways and other biological roles will become even more important in the future. These tools include *FunctSNP*, *Postgwas and NCBI2R* which are briefly discussed. Genome-Wide Gene Expression (Transcriptomics) analyses using RNAseq technology are briefly discussed with some examples including results from our own pig experiments. In the last part of this review, systems genetics and systems biology approaches are introduced that involve joint modeling and analyses of multi-omics data types from genomics through transcriptomics (microarray and RNAseq), metabolomics to proteomics. It is shown using published studies that these systems approaches are valuable and powerful compared to stand-alone genomic methods in identifying key causal and highly predictive genetic variants for complex traits as well as in building up complex genetic regulatory networks. In all sections, some applications of next generation/-omics methods in livestock species (e.g. feed efficiency, growth, weight gain, fertility and disease resistance in cattle, pigs and sheep) are provided with references to relevant software and tools. In conclusion, this paper reviewed the current progress, lessons and challenges in livestock genomics and its ongoing transition to and opportunities for integrative systems genetics and systems biology in animal and veterinary sciences. Most of these integrative systems genetics and systems biology tools and methods presented here are equally applicable to plant and human genetics and systems biology.

☆ This paper is part of the special issue entitled: Genomics Applied to Livestock Production, Guest Edited by Jose Bento Sterman Ferraz.

* Tel.: +45 35333577.
E-mail address: hajak@sund.ku.dk

http://dx.doi.org/10.1016/j.livsci.2014.04.028
1871-1413/

1. Introduction

Food production from livestock will be the primary driver in alleviating the concerns raised by a rapidly increasing human population's demand for food of animal origin. For efficient animal production and reproduction, there are challenges to be overcome for a better understanding of how animal production can contribute more effectively to the bio-economy. Animal breeding and genomics play a critical role in producing animal raw materials (meat, milk, eggs and their products) to meet current and future demands of food security for all human beings, while ensuring sustainable use of natural resources and less environmental impact. The genomic revolution in livestock was an aftermath of the human genomic revolution vis-a-vis genome sequencing projects. In the last 20 years, we have seen an astonishing development in livestock genomic technologies. Quantitative Trait Loci (QTL) mapping in the early 1990s spurred a lot of enthusiasm that saw several hundred research projects identifying QTLs in livestock species. The Animal QTL database (http://www.animalgenome.org/QTLdb/) reported several thousands of QTLs for major livestock species; however, most of these QTLs were detected using sparse microsatellite markers with large confidence intervals covering several megabases of the genome containing dozens to hundreds of genes and variants. Therefore, it was difficult to detect genes causing substantial quantitative trait variation which, in turn, initiated re-mapping and fine mapping of the initially mapped QTLs (see review by Georges, 2007). Subsequently, the release of whole genome sequences of major livestock species like cattle (The Bovine Genome Sequencing and Analysis Consortium et al., 2009), sheep (The International Sheep Genomics et al., 2010) and pig (Groenen et al., 2012) have led to a paradigm shift in availability of high-throughput genetic markers ranging from 10,000 to 50,000, 100,000 and up to one million Single Nucleotide Polymorphisms (SNP) markers today. These markers are genotyped using high-throughput Affymetrix or Illumina genotyping platforms (DNA arrays or SNPchips). In the first phase, the high-density SNP genotype data were mainly used in conducting Genome-Wide Association Studies (GWAS) that match genetic variants, with or without pedigree records, with the observed phenotype and provide estimates for hundreds of thousands of markers on each phenotype considered. In humans, GWAS has identified hundreds of associations of common genetic variants with over 100 diseases and traits (http://www.genome.gov/gwastudies). A consistent quest for variants that explain more of the disease or trait heritability has resulted in assaying increasingly higher-density SNP arrays with more than one million SNPs and dramatic increases in population sample sizes. In human genetics, the focus had been on precise delineation of causal variants that alter human phenotypes, particularly diseases, and on those variants that provide crucial insights into the biology connecting genotype and phenotype. In livestock species, the use of GWAS has been limited in the context of how it can be applied to breeding for improved performance and disease resistance. In both humans and animals, there are hundreds of success stories and the hype in GWAS is still unprecedented.

Functional genomics or transcriptomics studies that are based on microarray gene expression profiling (MGEP) has been and still is popular in livestock species. MGEP uses high-throughput transcriptomic arrays containing up to 30,000 transcripts to reveal underlying genetic (co) regulation in a set of biological conditions that clearly relate to phenotypic differences or disease states. These hybridization-based approaches typically involve incubating fluorescently labeled cDNA with custom-made microarrays or commercial high-density oligo microarrays. The focus of MGEP studies have been on those transcripts that provide holistic insights into the functional biology connecting genes throughout the genome and phenotypic or disease outcomes and eventually provide drug targets or biomarkers.

The SNP chip or microarray-based genomics and transcriptomics studies in livestock are being rapidly replaced by next- generation sequencing (NGS) technologies as robust genome/transcriptome sequencing technology platforms are widely and cheaply available and rapidly parallel development in statistical- and computational-biology and bioinformatics methods and tools to analyze NGS data. The NGS technology provides enormous opportunities for livestock sciences to move forward and make a transition to systems biology but also pose formidable challenges.

In summary, the sheer volume of genomic and transcriptomic data from hundreds of thousands of breeding animals in cattle, sheep, pigs and poultry and the availability of large-scale phenotyping for a range of complex and economically important traits has resulted in major challenges and opportunities for livestock production. This review paper is organized as follows: In the first part, I will briefly outline Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) methods with some existing challenges and examples from our groups' own studies. In the second part, which is one of the two main focus areas of this paper, I will outline next generation genomics wherein GWAS and genomic selection are based on complex statistical–computational genetic methods and on the next generation sequencing technologies for both genomic and transcriptomic profiling. In the third part, the other main focus of this paper, I will introduce livestock systems genetics and systems biology where networks construction (on genomic and transcriptomic datasets) and an integration of multiple data types from genomics through transcriptomics, metabolomics, and proteomics are shown to be valuable approaches to identifying key causal and highly predictive genetic variants for complex traits. In all sections, I have highlighted some of the software and tools that can be used.

2. Genome-wide association studies and genomic selection

2.1. GWAS—Single SNP and haplotype approaches

GWAS rely on a natural phenomenon of population-wide linkage disequilibrium (LD) between genetic (SNP) markers and causal variants, quantitative trait loci (QTL) or nucleotide (QTN). GWAS require larger samples of individuals than those required for linkage-QTL studies, because (1) population-wide

LD between SNPs and causal variants are expected to be low in livestock populations and (2) reliable association signals must show a very high level of significance (e.g. $p < 1 \times 10^{-7}$) to survive the multiple testing corrections for nearly a million association tests. There are hundreds of published GWAS in livestock species and it is not the intention to review all GWAS results here, but to highlight typical results from livestock based on our own GWAS on Danish production pigs. We conducted GWAS for feeding behavioral traits (Do et al., 2013b) and feed efficiency (Do et al., 2014) using Illumina Porcine SNP60 BeadChip. In the feeding behavior GWAS, we considered six behavioral traits observed on 1130 boars. The regions: 64–65 Mb on SSC 1, 124–130 Mb on SSC 8, 63–68 Mb on SSC 11, 32–39 Mb and 59–60 Mb on SSC 12 harbored several genome-wide significant SNPs. Fig. 1 shows a Manhattan plot for feeding behavior (number of visits per day to feeder) and residual feed intake 1 (RFI1) in pigs. This is the first GWAS to identify genetic variants and biological mechanisms for feeding behavior in pigs. Further, this study (Do et al., 2013b) conducted pig-human comparative gene mapping that revealed some important genomic regions and/or genes on the human genome that may influence human eating behavior and consequently affect the development of obesity and metabolic syndromes.

In another companion paper, we conducted GWAS for residual feed intake (RFI) in the same pig resource population (Do et al., 2014) where we defined RFI as the difference between the observed feed intake and the expected feed intake with two sub-definitions RFI1 and RFI2. Residual feed intake was the residual in the regression of daily feed intake (DFI) on average daily gain with initial body weight as a covariate in the model (RFI1) and the same definition, but with additional regression on backfat (RFI2) (Do et al., 2013a). Using deregressed estimated breeding values as response variables in GWAS, we detected 15 and 12 loci that were significantly associated ($p < 1.52 \times 10^{-6}$) with RFI1 and RFI2, respectively. Four and three linkage disequilibrium blocks were found on the two most interesting chromosomal regions: 30.5–31.5 Mb on porcine chromosome (SSC) 1 and 120.5–121.5 Mb on SSC 9 for both RFI. The SNPs within *MAP3K5* and *PEX7* on SSC 1, *ENSSSCG00000022338* on SSC 9 and *DSCAM* on SSC 13 might be interesting markers for both RFI measures.

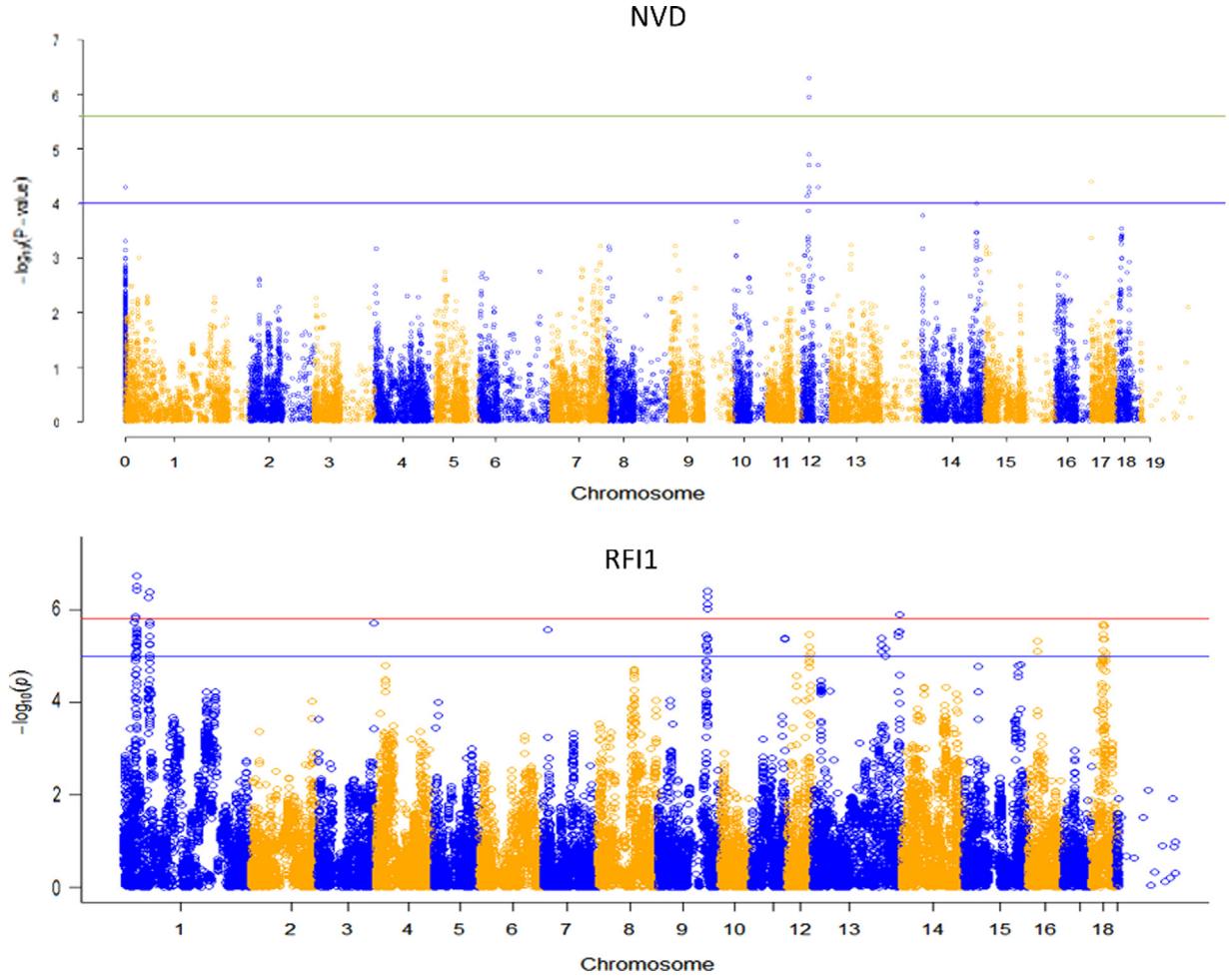


Fig. 1. Manhattan plot of genome-wide *p*-values of association for residual feed intake 1 (RFI1). The horizontal red and blue lines represent the genome-wide significance threshold at $p < 1.52 \times 10^{-6}$ and $p < 5 \times 10^{-5}$, respectively. From (Do et al., 2014). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

While GWAS has been typically performed using one SNP at a time, it has also become important to conduct GWAS using SNP haplotypes, because in reality SNPs often do not segregate independently and are transmitted in "haplotype blocks" due to high LD or physical proximity, for instance, as a cluster of 3–5 SNP markers. Instead of using single SNPs for predicting performance or disease risks, one may use a haplotype block formed by tightly linked/co-segregating SNPs. Once QTL regions harboring significant SNPs are identified by GWAS, QTL regional haplotypes can be defined and their effects estimated and used in predictions for complex diseases. An example of such an approach can be found in Mogensen et al. (2012) where haplotype-based prediction of risks for developing disk calcifications was made in wire-haired Dachshunds using 36 SNP markers within a susceptibility locus that was previously identified by GWAS at the genomic position: CFA12: 36,750,205–38,524,449.

Ideally, haplotype-based estimations and predictions should be performed on a genome-wide scale rather than on a limited number of QTL/genomic regions. For such a whole genome-based prediction and selection to be applied successfully, there is a need to understand the extent and distribution of LD across the entire genome in a population. In particular, we need to know how LD varies from one population to another, so that predictions in one breed can be valid in another breed. To facilitate patterning SNP haplotypes in the whole genome and understanding its diversity, we (Goodswen and Kadarmideen, 2011) developed *SNPpattern*—a generic bioinformatic tool for finding SNP allele patterns in populations (Goodswen and Kadarmideen, 2011). *SNPpattern* does this by grouping, counting, and comparing SNP allele patterns of various block sizes and statistically tests the differences in SNP allele block frequency as a measure of haplotype diversity within and between groups, defined by the user. We have demonstrated in another study how *SNPpattern* can be used to examine the patterns and extent of LD within and between four Australian sheep breeds on Ovine 60k SNPchip data (Goodswen et al., 2009). *SNPpattern* is implemented in Perl and supported on Linux and MS Windows. All scripts are freely available from the author or downloaded from http://systemsgenetics.dk/pages/resources.php. He et al. (2011) also developed an efficient approach to haplotype-based analysis in GWAS by using a reference panel where they showed that their method accelerated the phasing process and reduced the potential bias generated by unrealistic assumptions in the phasing process. Their haplotype-based approach delivered more power and less type I error inflation for GWAS.

Regardless of what type of GWAS is performed (either single SNP- or haplotype-based), there are still major technical and analytical challenges in GWAS. They include multiple test corrections leading to very conservative thresholds and thus missing biologically relevant loci or a block; inability or low power to detect loci of small effects; the risk of finding spurious association due to population stratification; overestimation of SNP (haplotype) effects; poor model fit (e.g. unaccounted epistatic and genotype-environmental interaction effects); insufficient sample sizes; low-density SNP coverage; excluded rare variants and undetected CNV effects. Perhaps one of the most significant limitations of GWAS is its inability to explain the full genetic variation in complex traits, more due to statistical issues than biological issue. This problem was first noted as "missing genetic variation or heritability" after GWAS on complex traits (Manolio et al., 2009; Clarke and Cooper, 2010; Gibson, 2010). Several papers have demonstrated that a mixed or random model GWAS can capture a much larger proportion of "missing heritability or genetic variation" (e.g. for human height Yang et al., 2010, 2011; Eichler et al., 2010) and there were indications that an SNPchip that only has common variants but not rare variants may be the cause of hidden heritability (Gibson, 2012).

2.2. Whole genomic prediction (WGP) and genomic selection (GS)

While it is widely accepted that standard GWAS statistical methods suffer from multiple testing and are underpowered to capture all genetic variation and overestimate SNP effects, an alternative GWAS statistical method that reduces these problems is needed. This is possible in a GWAS method that treats SNP effects as random and simultaneously fits family polygenic effects to account for stratifications as in usual best linear unbiased prediction (BLUP) methods (mixed model GWAS). While mixed model GWAS was the way to go to study genetic architecture of complex traits, the focus has now shifted to improving the prediction of unobserved phenotypes amongst populations (Makowsky et al., 2011; Wray et al., 2013). The term 'whole genomic prediction' (WGP) was coined in the landmark paper by Meuwissen et al. (2001). WGP methods based on BLUP models enable us to predict the unobserved performance of animals given their genotypes at SNPs without ever recording a phenotypic observation. This dramatically changed traditional progeny testing schemes in cattle and other species, because it only requires a smaller proportion of animals to be recorded for their performance, while the rest of the animals would only be genotyped. This has been quickly adopted, because traditional genetic evaluation schemes require a longer time span to prove genetic merit of animals as well as costs involved in progeny- or sib-testing schemes. The general consensus is that the genetic gain (ΔG) is increased by genomic selection (GS) which is defined as: $\Delta G = [i \times r \times \sigma_g^2]/G$ where i is the intensity of selection, r is the reliability of predictions, σ_g^2 is the genetic variance and G is the generation interval. GS favorably affects each one of these components in ΔG, by increasing i, r and σ_g^2 and by reducing G. Genomic selection has been thoroughly reviewed in several papers (e.g. recently by Meuwissen et al., 2013) and in special issues of major animal science journals (e.g. http://g3journal.org/site/misc/GenomicSelection.xhtml). However, the preceding discussion was intended as a summary of existing approaches and background for later introducing new methods of genomic selection in the systems biology context.

2.2.1. Basics of SNP-BLUP, GBLUP and ssBLUP

The main concept of WGP or GS in the simplest scenario can be explained in the form of a *SNP-BLUP* model

or Random Regression-SNP model that fits all markers simultaneously

$$\mathbf{y} = 1\mu + \mathbf{Mg} + \mathbf{Zu} + \mathbf{e}$$

where **y** is the phenotype, **1** is a vector of ones, **μ** is the mean, **M** is the genotype matrix with *m* number of SNP genotypes coded as 0, 1, or 2, **g** is the effect of each SNP, **Z** is a design matrix for random animal polygenic effect and **u** is the vector of polygenic effect. Distributional assumptions are often with $\mathbf{g} \sim N(0, \mathbf{I}\sigma_g^2)$, $\mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2)$ and $\mathbf{e} \sim N(0, \sigma_e^2)$. The term **Zu** can be dropped if *m* is sufficiently large (e.g. $>50{,}000$) such that markers capture most genetic variation that is present in the trait. Here, fitting **g** as *random regression* shrinks estimates back to 0 to account for the lack of information. Since this model fits *all SNP markers simultaneously* and the estimation is in one step, there is no problem of multiple testing and (high) False Discovery Rates. The genomic estimated breeding values (GEBVs) for each *i*th animal is then calculated in the second step as weighted sum of estimated SNP effects, $\widehat{g}$, with weights $\mathbf{M_j}$ being the genotype code for *i*th animal for the *j*th SNP for $j=1$ to *m*: $GEBV_i = \sum_j^m M_j \widehat{g_j}$.

The above SNP-BLUP model has been shown to be equivalent to what is called *genomic BLUP* or GBLUP method (VanRaden, 2008). The most important difference from the SNP-BLUP model is that the dimension of SNP genetic effects, $\mathbf{M'M}$ matrix with $m \times m$ dimension (where *m* is the number of SNP markers), is reduced to $\mathbf{MM'}$ matrix in GBLUP with $n \times n$ dimension (where *n* is the number of animals) in a mixed model equation (MME). The $\mathbf{MM'}$ is a standardized matrix with respect to allele frequencies, in that it behaves like a numerator relationship matrix **A** in a regular BLUP. The standardized $\mathbf{MM'}$ matrix is often called **"**genomic relationship matrix or GRM" (notation here is: **G**). GBLUP provides one solution or EBVs for animals (no summing up SNP effects, because there are no individual SNP effects estimated in GBLUP).

Other approaches, such as those based on Bayesian methods, use the prior distribution of QTL effects and allow some markers to shrink towards zero (zero variance explained by some markers). These methods use different shrinkage factors depending on the informative level of loci. It has some implications when we want to use only QTLs of moderate size in predicting genomic breeding values (GEBVs). In fact, we (Do et al., 2014, *unpublished*) investigated Bayesian Power LASSO (BPL) models with different power parameter to investigate genetic architecture, to predict genomic breeding values, and to partition genomic variance for RFI and daily feed intake (DFI) in Danish pigs. A total of 1272 Duroc pigs had both genotypic and phenotypic records for these traits. The BPL based gene mapping detected significant SNPs were detected on chromosome 1 (SSC 1) and SSC 14 for RFI and on SSC 1 for DFI. BPL models had similar accuracy and bias as GBLUP method but use of different power parameters had no significant effect on predictive ability of the models. Partitioning of genomic variance results showed that SNP groups either by position (intron, exon, downstream, upstream and 5′UTR) or by function (missense and protein-altering) had similar average explained variance per SNP, except that 3′UTR had a higher value.

The H-BLUP or single-step BLUP (ssBLUP) method proposed and further developed by Misztal et al. (2009), Christensen and Lund (2010), Forni et al. (2011), Legarra and Ducrocq (2012) includes both non-genotyped and genotyped animals in GEBV calculations. The main difference from GBLUP is that the genomic relationship (**G**) matrix is replaced by an **H** matrix that has the relationship computed for both genotyped and non-genotyped animals, based on both the marker genotypes and the pedigree. The inverse of the **H** relationship matrix can then be used in the traditional BLUP animal model to obtain the GEBVs of all animals (genotyped or not) where both sets of animals benefits from the exchange of phenotypic information via relationships, thus producing highly accurate GEBVs. A comparison of different statistical methods of WGP and GS is given in Koivula et al. (2012), Misztal et al. (2013) with species-specific implementation and reviews in dairy cattle (Pryce and Daetwyler, 2012; Bouquet and Juga, 2013) and pigs (Lillehammer et al., 2013; Tribout et al., 2013; Wellmann et al., 2013; Tribout, 2014).

In summary, many animal breeding organizations and companies routinely use either two-step or single-step methods (GBLUP or ssBLUP) to compute GEBVs for which many animal/plant breeding software packages can be used (e.g. ASReml package Gilmour et al., 2009) or the DMU package (Madsen et al., 2006)) and include them in a total merit index for selection of breeding animals.

3. Next-generation livestock genomics

The main method used to identify genes associated with the disease or trait of interest has been GWAS which focuses on identifying single SNP effects and has no ability to fit genome-wide genetic interactions. Thus, ignoring any contribution of (additive) epistatic interactions to additive genetic variance could explain the missing heritability. This has been postulated as one of the reasons for missing "heritability" in human studies (Manolio et al., 2009; Yang et al., 2010, 2011). In reality, the underlying model of association between genetic variants and many complex traits and diseases in humans and animals is non-additive meaning that single locus GWAS methods do not reflect the true associations. There is growing evidence that gene-gene and gene-environment interactions contribute to complex diseases and traits rather than single genes (Marchini et al., 2005; Kadarmideen et al., 2006a; Shao et al., 2008). Several models for epistasis (i.e. gene-gene interactions) have been proposed (Marchini et al., 2005; Shao et al., 2008), including models in which the genes alone have no effect on disease etiology, but where their interaction modifies disease risk. The genetic contribution might even include higher-order interactions between genetic and non-genetic factors in complex biological pathways. The genes involved in these complex underlying pathways will probably not be picked up using traditional single-locus analyses, and different methods are needed to extract this information from high-throughput genotype (HTG) datasets. Despite the appreciation of the key contribution of epistasis to genetic variation of complex traits, it has been ignored in practice

for some (good) reasons: genome-wide epistasis study using HTG data is simply a difficult task due to statistical complexity (e.g. multiple testing) and computational burden (Marchini et al., 2005; De Lobel et al., 2010). For example, even for the low-density Bovine SNP Chip which comprises of more than 50k SNPs, the number of SNP combinations would be 1.25×10^9 for testing two SNPs at a time; this analysis could take several days or even weeks on a standard workstation. On a high-density SNP panel (such as 777k), the problem is 225 fold. Several studies have detected the gene × gene interactions in two-stage models; however, only two SNPs are taken into account at a time. The number of possible interactions involving more than two loci will also be exponentially higher, referred to as the curse of dimensionality (Moore and Ritchie, 2004). While statistical–computational (optimization) strategies still take a long way to resolve this problem, the sample size to estimate billions of interaction effects would be astronomical. Hence, there has to be another strategy which reduces dimensionality, the multiple-testing problem and takes interaction effects into account when analyzing HTG data.

Two methods to fit genome-wide epistasis association (GWEA) analyses based on our recently published work (Ali et al., 2012, 2013; Kogelman and Kadarmideen, 2014) are described below. The first is based on a more practical strategy that examines a subset of SNPs that could have an influence on a trait of interest. We (Ali et al., 2012, 2013) developed a two-stage approach for analyzing genome-wide epistasis association (GWEA) and applied it to carcass traits in beef cattle. The second approach is called the Weighted Interaction SNP Hub (WISH) network method that develops genetic interaction networks based on HTG data and relates to phenotypes of interest (Kogelman and Kadarmideen, 2014). In brief detail, the two methodologies that go well beyond standard GWAS to capture additional genetic variation and underlying systems genetics of complex traits are described.

3.1. Genome-wide epistasis association (GWEA)

In GWEA, the first step is an additive association model where SNP effects are estimated by a single-trait-single-SNP association analysis (i.e. GWAS). From the first step of GWAS, a subset of significant SNPs at a very lenient threshold of p-value ≤ 0.001 or some other lower cut-off value can be selected from the additive association model. The lower thresholds are due to the fact that those variants with true epistatic effects will have small main-effect size and may not have survived the genome-wide stringent cut-off values. At the nominal or suggestive threshold, the number of SNPs will be a few hundred, which is much different from using the entire SNP dataset. Then, an analysis model for pair-wise (e.g. $SNP_1 \times SNP_2$) epistasis will be the following linear mixed model

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{fix} + \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_1 \times \mathbf{g}_2 + \mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is the phenotypic measurement, $\boldsymbol{\mu}$ is the overall mean, **fix** are significant fixed effects specific to each trait, $\mathbf{g_1}$ and $\mathbf{g_2}$ are three-level factors for genotypes at two SNPs (e.g. AA, AB, and BB) at $\mathbf{g_1}$ and $\mathbf{g_2}$, $\mathbf{g_1} \times \mathbf{g_2}$ is the interaction between SNP_1 and SNP_2 genotypes as an indicator for epistasis effect, $\mathbf{u}$ is the polygenic effect of animals and to account for the effect of relatedness, and $\mathbf{e}$ is the random error. Note that while the SNP effect for the additive association model may be treated as a covariate (to maximize power of association detection), for estimating interactions it is necessary to treat the effects of SNPs as factors. A separate model is then fitted for each pair of selected SNPs from the first stage GWAS and is run for all pair-wise combinations. To account for multiple testing, false discovery rates can be estimated for GWEA results using packages such as the q-value package in **R** (http://master.bioconductor.org/packages/release/bioc/manuals/qvalue/man/qvalue.pdf). These two-stage models can be used to ease multiple testing problems and computational demand.

We have successfully applied these two-stage epistatic models in 583 heifers of Brahman breed that were measured for carcass, growth traits and serum insulin-like growth factor-1 (IGF-1) (Ali et al., 2012, 2013). Our epistatic GWAS revealed key "hot spots" throughout the bovine genome for fat depth at 12/13th rib (RIB) and rump at P8 site. Fig. 2 shows genome-wide epistasis association on several chromosomes (see red spots in the heat map); the strongest epistatic signals are on BTA8/BTA12, BTA8/BTA 14 and BTA8/BTA15 for fat depth at P8 site. SNP annotation using dbSNP *Bos Taurus* genome revealed protein coding genes of *SNTG1*, *RAPGEF2*, *TMEM13D* and NNT as candidate genes. We have also applied epistatic GWAS models to the same cattle resource population to analyze serum insulin-like growth factor-1 (IGF-1), an indicator hormone for growth and reproduction (Ali et al., 2013). Using epistatic GWAS models, the most significant (p-value $= 10^{-15}$, q-value $= 10^{-12}$) epistatic signals were detected between rs29022513 on BTA 10 (86,513,542 bp) and rs29020759 on BTA 16 (7817810 bp); and between rs29016126 and rs29013864 on BTA 17 (44,319,169 bp and 18,809,070 bp, respectively). We identified 19 genes that had epistatic effects on the expression of *IGF-1* in this resource population.

3.2. Weighted interaction SNP Hub (WISH) network method

The WISH network method can be applied to HTG data using two different ways of detecting the interaction patterns between SNPs: (1) based on genomic correlations and (2) based on their epistatic interactions. The full methodological details are described in Kogelman and Kadarmideen (2014).

The WISH method based on genomic correlations (correlation between SNP genotypes in a group of individuals with extreme phenotypes) proceeds by describing relationships between SNPs by specifying an $n \times n$ dimensional adjacency matrix $\mathbf{A} = \mathbf{A}_{ij}$, where $\mathbf{A}_{ij}$ states the connection strength between SNP_i and SNP_j. The connection strength between the SNPs is defined by the absolute Pearson's correlation between the number of allele copies of pairs of SNPs for all SNP pairs in the data, creating a weighted network with values for $\mathbf{A}_{ij}$ between 0 and 1. This adjacency matrix is then raised to the power γ (soft thresholding) to ensure scale free topology. The

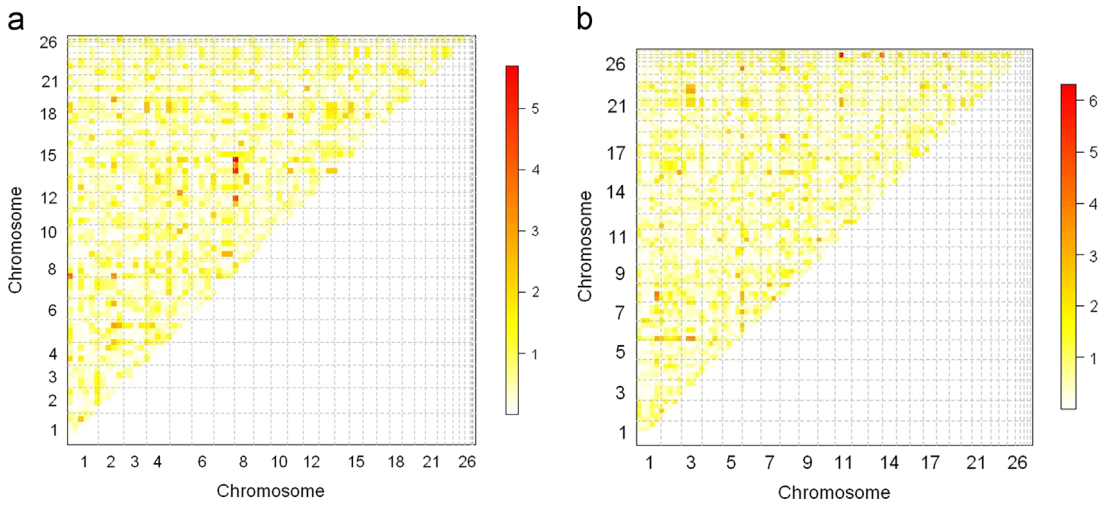


Fig. 2. Heat map image of genome-wide epistasis association. The heat map legend scale (a) left) is on −log10 (*p*-value) scale. (b) (Right): fat depth at P8 site. Right: fat depth at RIB. *Note:* red spots indicate epistatic signals (Ali et al., 2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

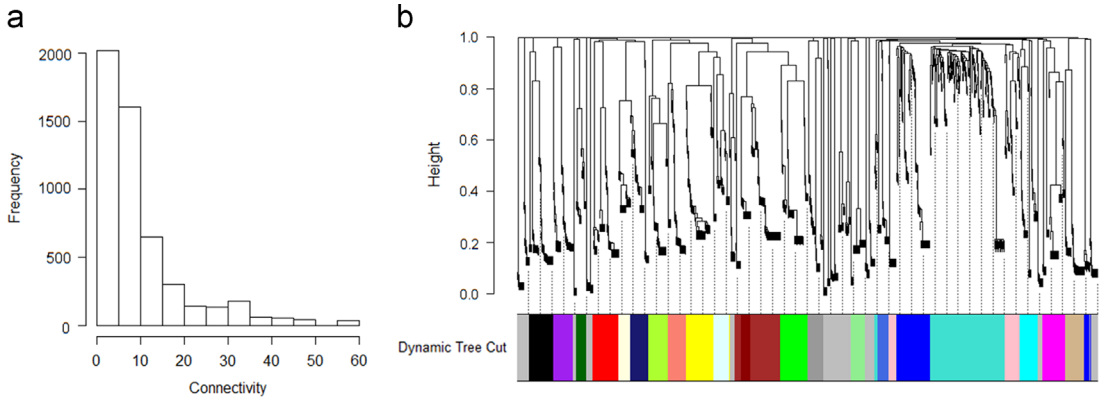


Fig. 3. Visualization of a WISH network construction based on genomic correlations. The histogram of the connectivity (a) shows many SNPs with low connectivity and a small number of SNPs with high connectivity (hubSNPs, potentially of biological importance). The SNP dendrogram (b) shows the clustering of SNPs based on the Topological Overlap Measure, whereby modules are differently colored. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

connectivity of a SNP (*c*) can concordantly be calculated by taking the sum of connection strengths between a SNP and all other SNPs. The adjacency matrix (**A**) that is created by calculating the correlations between the SNP genotypes and tested for scale free topology is the core object in WISH methodology. Results are visualized as "SNP dendrograms" and as "Topological Overlap Matrix" (TOM), showing highly interconnected SNPs in the form of large and small modules. The eigenvalues of each of this module are then associated to the phenotype contrasts and the test is called *"Genome-wide Module Association Test (GMAT)"*.

The WISH network method based on epistatic interactions is another approach on the same pipeline, but it deliberately targets known epistatic SNPs and builds networks based on them. The procedure is that the adjacency matrix is now based on the epistatic interactions instead of genotype correlations. The epistatic interactions between a pair of SNPs for all pair-wise combinations can be estimated using several methods (including our two-stage epistatic models described earlier as GWEA (Ali et al., 2012, 2013) which provides estimated regression coefficients for each pair of epistatic (SNP_i*SNP_j) interactions. Those regression coefficients represent the connection strength between SNPs and are therefore used as input for the adjacency matrix. The regression coefficients are normalized to create a data matrix with values between 0 and 1. From here on, the methods are comparable to the methods used in the WISH based on genomic correlations. We (Kogelman and Kadarmideen, 2014) applied the WISH network methods to an F2 pig resource population measured for carcass weight and have shown how the WISH network method, based on both genomic correlations and epistatic interactions, detected potential biologically relevant modules for carcass weight observations on individual pigs. The WISH network based on genomic correlations showed a clear scale-free network,

with many SNPs with low connectivity and a limited number of SNPs with high connectivity (Fig. 3A). Based on the TOM, we detected a total of 23 modules with at least 30 SNPs each (Fig. 3B). Of these 23 modules, three were appropriate for further downstream analysis based on their GMAT with the EBVs for carcass weight: the Blue module (GMAT=0.62, 62 SNPs), the Cyan module (GMAT=−0.62, 35 SNPs) and the Turquoise module (GMAT=0.42, 171 SNPs). Significant gene ontology (GO) terms and pathways in this module were detected using the NCBI2R R-package. Several of those GO terms and pathways were related to carcass weight, e.g. *actin filament processes* (Biological Process, Turquoise module) and *transforming growth factor (TGF) beta-activated receptor activity* (Molecular Function, Cyan module). Overall, the WISH method (Kogelman and Kadarmideen, 2014) is an advanced systems genetics/systems biology method to analyze GWAS-HTG data in the network context.

3.3. Next-generation genome sequencing

While I discussed at length the use of SNP chips in GWAS and Genomic Selection, recently, 'next-generation sequencing (NGS)' of the whole-genome has provided an unprecedented means to construct comprehensive maps of genetic variation that includes several million single nucleotide variants (SNVs), hundreds of thousands of small insertions or deletions, and thousands of structural variants (Cooper and Shendure, 2011). The NGS of DNA essentially includes chopping up the sample DNA (genome) into millions of small pieces (short reads of e.g. 50 or 100 bp) and aligning them back to a reference genome. Sequence depth, a measure of the number of reads covering a specific nucleotide position and averaged across all nucleotides, is often used to indicate how well DNA can be mapped to reference genome (if available for the species). If a genetic difference between sample DNA reads and that of the reference genome is identified, then they are reported as "genetic variant" ("Variant Calling"). However, since a whole genome consists of billions of nucleotides, one could expect millions of genetic variants (mostly SNPs) genotyped by NGS technology ("genotyping-by-sequencing"; (Glaubitz et al., 2014)). There are numerous resources available via the internet, articles and books, so this will not be explained any further here. There are some world-wide initiatives of whole genome sequencing in livestock; for instance in cattle, the 1000 bull genome project (www.1000bullgenomes.com).

Invariably, NGS has enabled 'next-generation GWAS and Genomic Selection' that also poses a great challenge to quantitative genetics. Meuwissen et al. (2013) argued and demonstrated that GBLUP or HBLUP may not profit much from the use of sequence data, because it merely uses the SNPs to estimate genetic relationships between the animals. They suggested that those GS methods that explicitly assume large numbers of variants with no effect and a small number of variants with large effect on the phenotype (QTN), such as BayesB, BayesC and BayesR, will give more accuracy in predicting the genetic merit. Using sequence data in a simulation study, Meuwissen et al. (2013) found an accuracy of GBLUP of ~0.5, whereas BayesB yielded accuracies of 0.83–0.97, depending on the number of simulated QTLs. They also found that whole-genome sequence data were substantially more accurate than a typical dense SNP-chip with 1000 SNPs per chromosome. They suggested that GS with whole-genome sequence data is still possible if we sequence the most influential founder animals from the current population, densely genotype the training population, and use genotype imputation to impute the missing genotypes based on those available. This approach can yield whole-genome sequence data for many thousands of training animals, whereas only relatively few founder animals are actually sequenced.

4. Systems genomic BLUP (*sg*BLUP) predictions

A new method and terminology, *Systems genomic BLUP (sgBLUP) predictions* is introduced here as one of the variants to existing the WGP and GS methods (as reviewed in Meuwissen et al., 2013). In whole-genomic prediction, it is possible to "assume" QTLs or QTN effects in the predictions such as BayesB, BayesC, and BayesR. However, the existing methods can be extended to deliberately model SNPs that have a *known* biological or functional role in the trait of interest, rather than assuming. Currently, entire genomic information (be it genotype or genome sequence data) is used only to build genomic relationship between animals and biological relevance of genetic variants is not captured or used in genomic predictions. This would unfortunately be a waste of investment and efforts in genomics. In order to accommodate biology into genomic prediction, we need to separate SNPs with known biological roles in the phenotype of interest based on information such as: SNP chromosomal location, exon/intron status, synonymous/non-synonymous effect, whether or not SNPs are in QTL regions, whether or not they are represented in KEGG/biological pathways, enriched in relevant GO terms and protein products for related genes, etc. Such an approach will differ from some of the existing methods that neither investigate biological functions of variants nor explicitly differentiate SNPs with *known* biological roles as different from those that have *unknown* biology.

We (Goodswen et al., 2010) developed an R package called *FunctSNP*, which is the user interface to custom built species-specific databases (human, pigs, sheep, poultry and cattle) containing SNP data together with functional annotations, further described later in this paper. However, there are other similar software packages that are freely available; for instance—*postgwas* (Hiersche et al., 2013), SNAP (Johnson et al., 2008) and NCBI2R (http://NCBI2R.wordpress.com). Moreover, there are SNPs known to be *expression QTLs* (eQTLs) or *expression Quantitative Trait Nucleotide (eQTN)* that influence the expression levels of genes (or gene transcripts). Such SNPs can be considered to be functional variants having regulatory effects (either in *cis-acting* or in *trans-acting* mode) on many candidate genes. Full reviews of detection and mapping of eQTLs or eQTNs applied to various traits in livestock species are available in Kadarmideen et al. (2006b), Kadarmideen and Reverter (2007),Kadarmideen (2008). So with the knowledge

of systems genetics of these genetic variants (and with tools such as *FunctSNP, postGWAS, SNAP*), one can categorize SNPs as functionally relevant and not relevant (residual genome). Then it is possible to explicitly model these two types of SNPs in a GBLUP model.

The GBLUP model can be extended by including two sets of SNPs, one with known biological functions **(M_B)** and the other with unknown functional role (M_U) as random effects in the GBLUP model. These random effects for the two different groups M_B and M_U may have different genetic variances and may also have different distributional assumptions.

Hence, the model will look like

$$y = \mu 1_n + \mathbf{M}_B g_B + \mathbf{M}_U g_U + e$$

where y is the vector of observations; μ is the intercept, 1_n is the vector of ones, $\mathbf{M}_B$ and $\mathbf{M}_U$ are genotype matrices corresponding to random genetic effects $\mathbf{g}_B$ and $\mathbf{g}_U$, respectively and **e** is the residual error. Note that the scaled GRM (**G**) matrix should be constructed separately for each one of the SNP groups, depending on the number of SNPs in each group.

Then the solutions of the mixed model equations (MME) are

$$\begin{bmatrix} \hat{\mu} \\ \hat{g}_B \\ \hat{g}_U \end{bmatrix} = \begin{pmatrix} \mathbf{1'1} & \mathbf{1'M}_B & \mathbf{1'M}_U \\ \mathbf{M}_B\mathbf{'1} & \mathbf{M}_B\mathbf{M}_B' + \mathbf{I}\lambda_{g_B} & \mathbf{M}_B\mathbf{M}_U' \\ \mathbf{M}_U\mathbf{'1} & \mathbf{M}_B\mathbf{M}_U' & \mathbf{M}_U\mathbf{M}_U' + \mathbf{I}\lambda_{g_U} \end{pmatrix}^{-1} \begin{bmatrix} \mathbf{1'y} \\ \mathbf{M}_B\mathbf{'y} \\ \mathbf{M}_U\mathbf{'y} \end{bmatrix}$$

where $\lambda_B = (\sigma_e^2)/(\sigma_B^2)$ and $\lambda_U = (\sigma_e^2)/(\sigma_U^2)$ Setting $\mathbf{G}_B = \mathbf{M}_B\mathbf{M}_B'$ and $\mathbf{G}_U = \mathbf{M}_U\mathbf{M}_U'$ corresponding to two GRMs (scaled as per allele frequencies as in VanRaden, 2008), then variances of two groups of SNP random effects are

$$\mathrm{Var}\begin{bmatrix} \mathbf{g}_B \\ \mathbf{g}_U \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_B\sigma_B^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_U\sigma_U^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_{\mathbf{e}}^2 \end{bmatrix}$$

where σ_B^2 and σ_U^2 were genetic variances associated with $\mathbf{g}_B$ and $\mathbf{g}_U$ effects, respectively and $\boldsymbol{\sigma}_{\mathbf{e}}^{\mathbf{2}}$ is the residual variance. For this univariate *sg*BLUP model, $\mathbf{g}_B$, $\mathbf{g}_U$ and **e** can be assumed to be normally distributed with means zero and (co) variances as specified above or a choice of distribution could be different between these two SNP sets, e.g. one with *t*- or χ^2 distribution and the other with normal distribution. It is important to have as accurate information on σ_B^2, the genetic variances associated with $\mathbf{g}_B$ effects and it is based not only variances explained but also on the expected role of SNPs in underlying biology of trait. For simplicity we assume zero covariance between SNPs in $\mathbf{M}_B$ and $\mathbf{M}_U$ ($\sigma_{BU}=0$), but a non-zero covariance can be modeled due to the existence of LD between SNPs in two SNP subsets $\mathbf{M}_B$ and $\mathbf{M}_U$ ($\sigma_{BU} \neq 0$). It is reasonable to expect that SNPs that are grouped as being "biologically relevant" in $\mathbf{M}_B$ is indeed a robust assumption. This is because the grouping is done by well-established systems genetics software such as *FunctSNP* or *SNAP* or *postgwas or* suite of *Bioconductor* genomic annotations tools; they have an in-built biological data validation methods and stringent statistical tests to control false positives which in turn ensures that these biologically relevant SNPs are not a random set by chance alone (see Section 4.1).

As in GBLUP, genotype codes 0, 1, or 2 are standardized using their allele frequencies to have a mean of 0 and standard deviation of 1, for instance as $\mathbf{G}_B = \mathbf{M}_B\mathbf{M}_B'/2\Sigma p_i(1-p_i)$ where elements in column i of $\mathbf{M}_B$ are $0-2\mathrm{p}_i$, $1-2\mathrm{p}_i$ and $2-2\mathrm{p}_i$ for genotypes AA, AB and BB genotypes, respectively, and p_i is allele frequency of B allele. $\mathbf{G}_U$ will be scaled similarly. So the *Systems genomic BLUP* is a regular GBLUP model except that genome portioning is made with respect to known and unknown biological or functional information of SNPs. Hence, two GRMs, $\mathbf{G}_B$ and $\mathbf{G}_U$, allows better capturing of biologically relevant information for the phenotypes and replaces the traditional **G** matrix.

Then this follows the straightforward GBLUP approach, with two types of estimated breeding values: $\widehat{g_B}$ corresponding to biologically relevant SNP effects and $\widehat{g_U}$ corresponding to random SNP effects of unknown biology function. Then the total genomic breeding value ($\widehat{GEBV}$) of an individual is calculated as the weighted sum of breeding values estimated with two SNP subsets

$$\widehat{\mathbf{GEBV}} = \mathbf{w} \times \widehat{\mathbf{g}}_B + (1-\mathbf{w}) \times \widehat{\mathbf{g}_U}$$

where, the weight **w i**s given arbitrarily so as to match the biological or functional importance of the SNPs in $\mathbf{M}_B$ in relation to the phenotype. Note that if $w=1$, then the entire prediction is based on biologically important SNPs only; if $w=0.5$, equal weights are placed on both sets of SNPs and if $w=0$, then entire prediction is based on non-biologically important or residual SNPs. The default option would be to give equal weights. Overall, **sgBLUP** is very appealing, because it addresses a valid limitation that **GBLUP** or **ssBLUP** (single step BLUP) methods do not explicitly capture underlying biology. In other words, I argue that the ***sgBLUP*** matrix would capture true biological relationships better than the **GBLUP or ssBLUP** matrix, provided the annotations and pathway information on tested SNP markers are adequate enough to link phenotypes. It must be noted that $\mathbf{G}_B$ can be replaced by $\mathbf{I}_B$ if there is only a small number of SNPs (e.g. <100), as the genomic relationship information in $\mathbf{G}_U$ containing rest of the SNPs will be sufficiently high. Further, it should be noted that **sgBLUP** would become statistically and computationally more demanding if multiple trait genomic prediction fitting several traits simultaneously is performed. This is because the two GRMs, $\mathbf{G}_B$ and $\mathbf{G}_U$, will be different for each trait.

4.1. Tools for sgBLUP for linking SNPs to their functions

In most GWA studies, an associated SNP is likely part of a larger region of linkage disequilibrium (containing several hundreds of SNPs), making it difficult to precisely identify the genetic variants that are biologically linked with phenotypes. It has been shown that analyzing biological pathways using a systems approach could therefore potentially complement efforts to identify causal loci for complex traits. For instance, SNPs within coding regions of genes, which are involved directly in producing proteins, metabolites or hormones affecting phenotypes, would need to be given higher weights than those that are peripherally involved. Finding the potential biological functions of such SNPs can be an important step towards

further use in human and agricultural populations (e.g. for identifying genes related to susceptibility to complex diseases or genes playing key roles in development or performance). In the context of the newly proposed method *sgBLUP*, one needs to allocate SNPs to two different SNP groups: one with *known* biological functions ($\mathbf{M}_B$) and the other with *unknown* functional role ($\mathbf{M}_U$). The current challenge is that the information holding the clues to SNP functions is distributed across many different databases and researchers do this on an ad hoc basis. We identified the need for efficient bioinformatics tools to seamlessly integrate up-to-date functional information on SNPs. Many web services have arisen to meet the challenge, but most work only within the framework of human medical research. As mentioned before, we released an R package called *FunctSNP*, which is the user interface to custom build species-specific databases for 5 species: cattle, pigs, chicken, sheep and human (Goodswen et al., 2010). The *FunctSNP* functions provide access to information such as: SNP chromosomal location, exon/intron status, synonymous or non-synonymous effect, SNPs in Quantitative Trait Loci (QTL) regions, biological pathways, GO terms, and protein products for related genes. Multiple databases (one for each species) can be queried in the same R session. The FunctSNP software can be obtained from the author or downloaded from http://functsnp.sourceforge.net/.

The more recent "*Postgwas*" package (Hiersche et al., 2013) is specifically aimed at post-processing, visualization and advanced analysis of GWAS results. It attempts to unify and simplify several procedures that are essential for the interpretation of GWAS results, including the generation of advanced Manhattan and regional association plots with rare variant display as well as novel interaction network analysis tools for the investigation of systems-biology aspects.

NCBI2R is another R package that annotates lists of SNPs and/or genes, with current information from NCBI, including LD information. R functions in this package will provide annotation of the results from GWAS to provide a broader context of their meaning. For instance, it can generate candidate SNP/gene lists that are created from keywords, such as specific diseases, phenotypes or gene ontology terms. The output of this package produces text fields and web links to more information for items such as gene descriptions, nucleotide positions, OMIM, pathways, phenotypes and lists of interacting and neighboring genes. Please see the website at http://NCBI2R.wordpress.com for more information.

Annotations of genomic data, in general, can be carried out in R programs like Bioconductor (http://www.bioconductor.org/help/workflows/annotation/annotation/). Bioconductor has extensive facilities for mapping between microarray probe, gene, pathway, gene ontology, homology and other annotations. This is facilitated via GO, KEGG, vendor and other annotations, and R interface can easily access NCBI, Biomart, UCSC and other sources.

There are many different software tools which follow the concept of *FunctSNP*, *Postgwas and NCBI2R;* it is not my intention to discuss them all, but to highlight importance of such tools for livestock *sg*BLUP proposed earlier.

4.2. BLUP|GA (BLUP approach given the genetic architecture)

Recently, Zhang et al. (2014) proposed similar method as *sg*BLUP, called "BLUP|GA" ('BLUP approach given the Genetic Architecture'). They demonstrated that the performance of WGP can be improved by including the publicly available GWAS or QTL results for traits of interest in their BLUP|GA and that WGP accuracy can be improved especially in situations where the prediction accuracy is limited by a small sample size and/or when trait heritability is low. They illustrated the superiority of BLUP|GA over GBLUP and BayesB with a dairy cattle data (milk fat percentage, milk yield and somatic cell score) and a rice data set (11 different traits). This shows that, in general, GS methods that account for biologically important markers in WGP can indeed make a difference and be beneficial. The key component of BLUP|GA method is the separation of **M** into two subsets (one corresponding to markers located in known QTLs previously reported ($\mathbf{M_1}$) and the rest with small effects and not in QTL regions (**M**)). Here $\mathbf{M_1}$ is similar to $\mathbf{M}_B$ (but there are critical differences, as mentioned below). Then they specify a diagonal matrix, **D**, with marker weights for each locus on the diagonal to represent the relative size of variance explained by the corresponding loci in $\mathbf{M_1}$. Therefore it is a trait-specific genomic relationship matrix, like $\mathbf{M}_B$. Then they calculate **S** as: $\mathbf{M_1 D M_1}'/2\Sigma p_i(1-p_i)$. Here **S** corresponds to $\mathbf{G}_B$ in *sg*BLUP.

Finally, they use the relationship matrix **T** which contains two matrices, each weighted by ω as: $T=\omega\mathbf{S}+(1-\omega)\mathbf{G}$ where **S** is based on the set of markers being "important" for the considered trait and their relative weights are chosen based on mapping onto a significant QTL region during association studies previously carried out in the literature. The **G** corresponds to the standard genomic relationship matrix proposed by VanRaden (2008). In case $\omega=1$ then $\mathbf{M}=\mathbf{M_1}$, leaving entire WGP based on only important markers.

There are critical differences between *sg*BLUP and BLUP|GA. (1) *sg*BLUP has two random animal genetic effects (g_B and g_u) in the model whereas BLUP|GA has one random animal genetic effect. (2) *sg*BLUP utilizes not only markers that are in large effect QTL regions but also considers markers involved as expression QTL (eQTLs) and play a key functional role via biological or metabolic or signaling pathways underlying traits in question, *regardless of its effect size* (3) *sg*BLUP does not require estimated SNP effects neither does it calculate weights for each and every marker and for each trait evaluated. It simply assumes a different distributions and genetic variance, *apriori*, for all markers in a marker set $\mathbf{M_B}$ than $\mathbf{M_U}$. Such an approach is easily implemented in Bayesian methods such as Bayes R or Bayes Cpi. (4) In BLUP|GA, there is one GEBV while in sgBLUP, there are two GEBVs: *sg*BLUP assigns weights to two GEBVs but not to two relationship matrices.

5. Genome-wide gene expression (transcriptomics)

It is not only the genomics that have a prominent role in identifying key genes and variants useful for livestock production and health, but transcriptomics also play a critical role. In the last decade, microarray gene expression

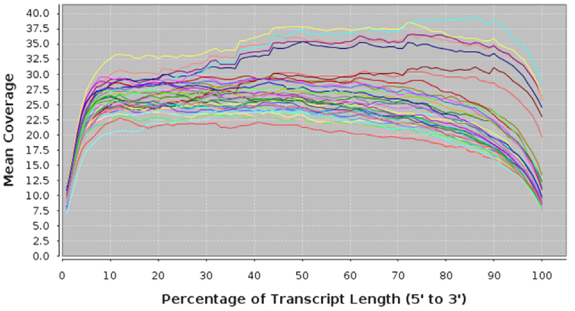
profiling (MGEP) has been and still is popular in livestock species. MGEP uses high-throughput transcriptomic arrays containing up to 30,000 transcripts to reveal underlying genetic (co) regulation in a set of biological conditions. These hybridization-based approaches typically involve incubating fluorescently labeled cDNA with custom-made microarrays or commercial high-density oligo microarrays. Most transcriptomics experiments are focused on detection and annotation of differentially expressed (DE) and co-expressed (CE) genes as well as construction of gene networks (for review of livestock transcriptomics, see (Kadarmideen and Reverter, 2007). We have used these approaches to unravel the biology and genomics underlying sheep resistance to gastrointestinal nematode (GIN) infections (Kadarmideen et al., 2011; Kadarmideen and Watson-Haigh, 2012), sheep muscle growth and development (Kogelman et al., 2011) and wool growth in Australian sheep (McDowall et al., 2013). We have also previously reported on new tools for gene co-expression network analyses using high-throughput transcriptomic datasets (Watson-Haigh et al., 2010; Kadarmideen and Watson-Haigh, 2012).

Hybridization methods have several limitations that include (over)reliance on existing knowledge of the genome sequence and high background levels due to cross-hybridization. A limited dynamic range of detection owing to both background and saturation of signals and comparing expression levels across different experiments is often difficult and can require complicated normalization methods. However, the limitations of microarray technologies were quickly overcome by next- generation sequencing (NGS) technologies in late 2000s. The main factors enabling the transition from Sanger sequencing (called first-generation sequencing) to next-generation sequencing (NGS) were the availability of robust sequencing technology platforms and dramatic reduction in costs associated with NGS. The fundamentals of NGS technologies are reviewed by Metzker (2009), Abecasis et al. (2010), Metzker (2010). When NGS approaches are applied to directly sequence mRNA, they are called *RNAseq*. RNA-seq is a quantitative approach in that it directly determines and counts the entire mRNA sequence, thereby estimating RNA expression levels in cells or tissues with higher accuracy than intensity-based microarrays. Consequently, results between RNAseq experiments can be compared directly without requiring complicated normalization methods. In addition to determination of gene expression levels, NGS also allows for the detection of cSNPs, novel and rare transcripts, novel protein isoforms, alternative splice sites, ncRNA and allele-specific expression in one single experiment. Traditionally, this would have required separate experiments, costing money and time. However, a major limitation associated with NGS data analyses compared to microarray data analyses is the requirement of large data storage and High Performance Computing (HPC) facilities. In general, RNAseq experimental design is no different than MGEP experimental design; both have to have solid experimental hypotheses to be tested and the corresponding statistical power to prove or disprove the set of hypotheses. In RNAseq, the sample/library preparations are different, obviously, due to the fact that it is sequencing the mRNA. Critical in the library preparation is the decision regarding read length (short or long) in base pairs and whether this is a single- or paired end read and read depth (in millions). The raw RNAseq data from the sequencing laboratory is supplied in the form of *fastq* files for each sample (each read in a sample comes with sequencing information, exact nucleotide sequence and the quality score in ASCII format). This fastq file could be well over 10 GB each depending on the read depth. They are then subjected to preliminary quality control, for example, using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and then sequence reads are assembled via mapping to the reference genome (using for example, Bowtie2, TopHat, GMAP, CLC-Bio software). There are several challenges in alignment to a reference genome/transcriptome or doing *de novo* assembly of transcriptome, as discussed in Vijay et al. (2013). The difference between genome mapping and transcriptome mapping comes in the form of alternate splicing, RNA editing, post-transcriptional modifications and variations from the reference such as substitutions, insertions and deletions, so any aligner or mapper has to take these issues into account. The ideal mapper algorithm would be the one that takes all RNAseq data types (single-end, paired-end, strand specific or non-strand specific) from all sequencing platforms (Illumina, SOLiD, Roche-454, PacBIO…). The mapped reads are available as.sam or.bam files for each sample, which needs to be quality controlled because some issues only appear after the mapping/alignment of reads are finished. Running the after-alignment/mapping quality control (e.g. using software such as *RNA-SeQC*, DeLuca et al., 2012) or *Qualimap* (http://qualimap.bioinfo.cipf.es/) results in, among others, an overview of the number of aligned reads, coverage and comparisons of samples by their expression. An example of this is the quality control results of paired-end RNAseq data from subcutaneous adipose tissue of a pig resource population (Kogelman et al., 2013) is illustrated in Fig. 4. After alignment, a normalization of the counted reads has to be performed to account for differences between samples (e.g. biological and library differences), but also to account for differences within samples (e.g. gene length and GC-bias). Several methods have been proposed and discussed (Dillies et al., 2013) and there are several expression data quantification tools including HTseq, Cufflinks and Qualimap. Moreover, the large size of the data files (several GB per raw sample) increases the need for large memory and storage requirements, but also for improved bioinformatics tools. A review of computational methods and bottlenecks involved in RNAseq data analyses is given in Garber et al. (2011) which underlines the importance of memory and storage requirements. After completing all the QC steps of RNAseq data, then high level, exploratory, bioinformatic and systems biology analyses can be performed, for instance, to detect differentially expressed (DE) genes, build co-expression (CE) networks, detect splice variants, allele-specific gene expression patterns, perform downstream enrichment and pathway analyses and so on. Many of the freely available tools that were originally made for MGEP in *Bioconductor* suite of R programs can now be used for RNAseq data (e.g. LIMMA) but there are

a

Sample	Note	Intragenic Rate	Exonic Rate	Intronic Rate	Intergenic Rate	Expression Profiling Efficiency	Transcripts Detected	Genes Detected
A		0.822	0.677	0.145	0.177	0.677	20,011	11,632
B		0.820	0.676	0.144	0.179	0.676	20,248	11,744
C		0.828	0.702	0.126	0.171	0.702	20,760	12,101
D		0.818	0.683	0.135	0.181	0.683	20,276	11,826
E		0.812	0.651	0.161	0.187	0.651	19,833	11,523
F		0.824	0.682	0.142	0.175	0.682	19,937	11,610
G		0.824	0.687	0.137	0.174	0.687	20,466	11,916

b



c

Rank	Transcript ID	Length	Total Cov.	Mean Cov.	Std Dev
1	ENSSSCT000	784	71408051	91081.70	60566.13
2	ENSSSCT000	681	51272218	75289.59	39019.04
3	ENSSSCT000	688	45588421	66262.24	35953.07
4	ENSSSCT000	204	18809646	92204.15	31122.86
5	ENSSSCT000	1140	48640816	42667.38	35537.76

Fig. 4. Quality control of RNAseq samples in one of our ongoing projects in pigs using RNA-SeQC. (a) showing an overview of the different samples and the results of the alignment against the pig reference genome, (b) visualization of the coverage of the median expressed transcripts in a particular sample, and (c) some transcript details including the length and coverage of a particular sample.

now new generation and dedicated RNAseq tools available. See examples of such analyses in livestock here: (Watson-Haigh et al., 2010; Kogelman et al., 2011; Kadarmideen et al., 2011; McDowall et al., 2013; Kogelman et al., 2014). Overall, just as it took several years to optimize the pipeline for Microarray Gene Expression studies, it will likely take a while before the entire RNAseq data analysis pipeline is optimized.

6. Systems genetics and systems biology

The term "systems genetics" in animal context was originally proposed by (Kadarmideen et al., 2006b) and expanded in Kadarmideen (2008) where perspectives were provided for how -omics scale measurements in livestock can be integrated to find important causal & regulatory genes and their variants and highly predictive biomarkers. This systems genetics method was then applied in livestock by Kadarmideen and Janss (2007), Kadarmideen et al. (2010), Kogelman et al. (2014) and recently also in human studies, as reviewed by Civelek and Lusis (2014). To elaborate on this, it is important to first introduce systems biology as a discipline. Systems biology approaches, by necessity, involve systematic data collected at all levels of the biological system and are aimed at studying interactions between these levels, but not at one level in isolation. With modern high-throughput technologies, hugely comprehensive data at all levels of the biological system are now available (genome-wide, transcriptome-wide or metabolome-wide, proteome-wide measurements). Systems biology collectively and iteratively models and analyzes these datasets using a combination of mathematical or statistical models, computational biology and bioinformatic principles and tools. Systems biology is not only about data-driven genome-scale measurements; it is also about a philosophy and a hypothesis-driven approach for experimental design and analysis (Kadarmideen, 2008)). The 'data-driven' modeling approach explaining the source of variation found in biological data is quite familiar to most statistical geneticists and biostatisticians. However, differences arise in modeling integrated multi-dimensional high-density omics data points; and this, in fact, is the formidable challenge now and for the future. The 'hypothesis- driven' modeling approach attempts to predict the outcome of new biological experiments iteratively using evidence from the 'experimental or wet data'. By "systems biology", I mean it is a discipline that iterates between wet and dry approaches to understand the whole biological system and provide a complete blueprint of functions of phenotype or a complex disease evolution. Therefore, it requires multi-disciplinary expertise in one team, from mathematical sciences through quantitative biology to molecular biology. Livestock systems biology is still an evolving field and only a very handful of true systems biology experiments are ongoing.

Systems genetics *is a branch of systems biology* that focuses *only* on integrating genetic factors (SNPs, CNVs, QTLs etc.) causing variation between individuals in intermediate -omics traits (whole genomic gene expression levels, metabolomic or proteomic levels, etc.). Fig. 5 illustrates the concept behind systems genetics. The systems genetics approach is seen as follows. The triangle has each corner representing a different type of data (genomic, other -omics and phenotypic) and it shows that the QTL or QTN affecting different biological measurements from genome through transcriptome to proteome or metabolome, all the

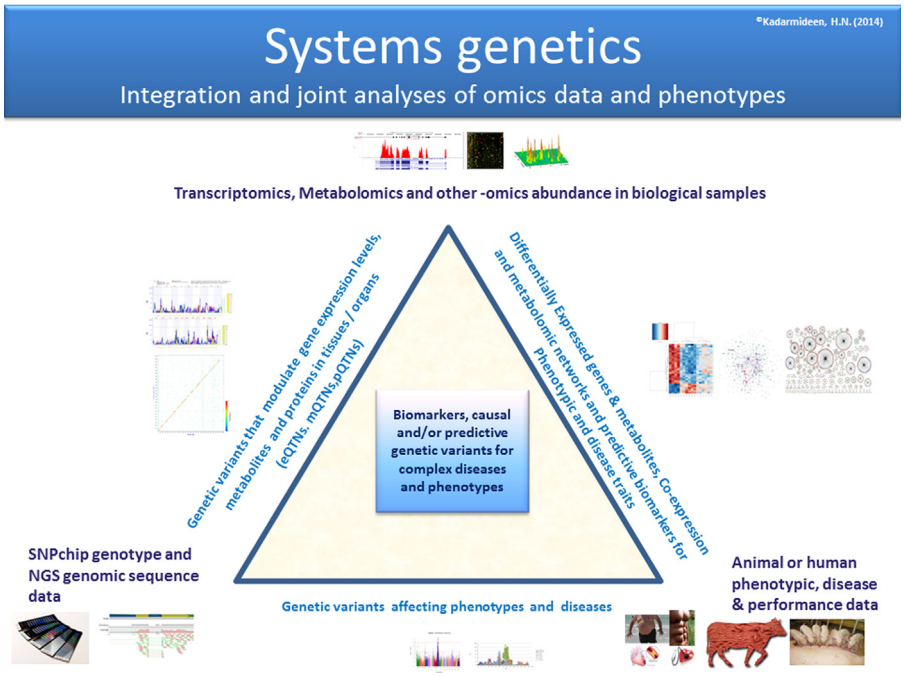


Fig. 5. Illustration of integrative systems genetics approaches that integrate genomic data and other -omics data types with diseases and phenotypic traits to detect highly predictive biomarkers, causal variants and master regulatory genes and variants in complex traits and diseases. Modified from (Kadarmideen, 2008).

way to an "exogenous" phenotype or a disease can all be integrated. This is still "genetics", because we are *systematically amassing* those genetic variants that exert their effects from DNA to phenotypic expression or disease manifestations, hence the term 'systems genetics'. In Fig. 5, the bottom line connecting genomic data with phenotype represents GWAS, QTL detection or whole genomic prediction performed directly on the end or exogenous phenotype or a disease. The detected SNPs or QTLs are causing variation in a complex polygenic trait or a disease measured on the animal itself. The left side of the triangle connecting genomic data with other -omics data represents genetic variants or QTLs causing variation in "endogenous phenotypes" or "endo-phenotypes" such as transcript abundance in tissues (*expression quantitative trait loci* or eQTLs) or equally protein levels or metabolite levels in biological samples measured in mass spectrometry or liquid chromatography (*protein* QTL or pQTL; *metabolite* QTL or mQTL, respectively). To be precise, with current SNP-based association tests, they can be called QTNs (eQTN, mQTN, pQTN, etc.). Finally, the right-hand side of the triangle connects -omics data (other than genomic) directly with a phenotype and hence represents gene expression, proteomic or metabolomic profiling experiments and analysis in clearly defined disease categories and phenotypic contrasts. This detects key gene transcripts or metabolites or proteins whose variations directly and significantly affect the disease outcome or phenotypic differences. Because SNPs or genetic variants (regardless of what they affect—be it a QTL or eQTL or mQTL or pQTL) can be assigned to genes or somewhere closer to them, and if these genes are also implicated in the right-hand side of the triangle, then we have co-locating genes or variants that affect or influence the entire trajectory from DNA through endo-phenotypes to exogenous phenotypes or diseases. It will be these genes or biomarkers that will be targets for future interventions, predictions and preventions, as they are highly and accurately predictive and/or causative from both biological and statistical viewpoints. Hence, systems genetics derives its name, as originally proposed in Kadarmideen et al. (2006b) and further expanded in Kadarmideen (2008), by being able to integrate analyses of all underlying genetic factors acting at different biological levels, namely, QTL, eQTL, mQTL, pQTL and so on. This leads us to provide a holistic view on complex trait heredity. Some of the recent examples of combining GWAS SNP chip with RNAseq for identification of key genetic and biomarkers for diseases based on systems genetics methods include those of Fu et al. (2012), Brown et al. (2013), Westra et al. (2013) and recently our own (Kogelman et al., 2014). It should also be noted that such systems genetics or eQTL approaches require much smaller sample size (around 40) to achieve a statistical power of 80% in detecting key genes and biomarkers (Kadarmideen, 2008). A nice and recent overview of systems genetics with applications in human genetics is given in Civelek and Lusis (2014).

One of the specific branches of systems genetics is *genetical genomics* that helps us to investigate the inheritance of regulatory loci, eQTL. The basic principle is that *transcript abundance is treated as a phenotype* and typical QTL or GWAS approaches are applied to this phenotype or expression trait (*e-trait*). This means that there could be several transcripts or e-traits (in livestock, in the range of potentially 20,000 to 35,000) and a correspondingly equal number of GWAS. The statistical methods of *genetical*

genomics are described in many key papers, starting with the landmark paper of (Jansen and Nap, 2001), followed by Kendziorski and Wang (2006), Kadarmideen (2008) and many others. The introduction of this concept to livestock sciences appeared first in Kadarmideen et al. (2006b) and was further developed and applied (Kadarmideen and Reverter, 2007; Kadarmideen, 2008; Kadarmideen et al., 2010). To date, there are many experimental eQTL studies that have successfully applied this approach in animal sciences (e.g. in chickens (de Koning et al., 2007), in mice (Kadarmideen et al., 2006b; Kadarmideen and Janss, 2007; Kadarmideen, 2008) and in pigs (Ponsuksili et al., 2011; Heidt et al., 2013, Steibel et al., 2011). A recent review of eQTL studies across many species can be found in Nica and Dermitzakis (2013). We have recently applied this eQTL approach to an F2 pig model for human obesity where we jointly modeled and analyzed paired-end RNAseq data (approx. 30 million reads) with Illumina Porcine 60k SNPchip data to reveal potential causative and master regulatory loci for obesity and related metabolic traits (Kogelman et al., 2014).

Briefly, with the use of dense SNP markers from genotyping SNPchip and gene expression data from microarrays or RNAseq experiments, one begins mapping eQTL with a general statistical model following Kadarmideen (2008). Similar to GWAS on regular phenotypes, one SNP is tested at a time for its association with the gene expression phenotype. Let y_{ij} be the logarithm-transformed expression phenotype for the jth individual with the ith marker genotype ($i=1$ to 3) corresponding to SNP genotypes AA, AB and BB, respectively. A general model is then

$$y_{ij} = \mu + a_m \times M_{ij} + \delta_m \times M_{ij} + e_{ij}$$

where μ is an overall mean for the expression of gene transcript, the coefficients or the values of M_{ij} for fitting additive effects, a_m, are 1, 0 and -1 and for fitting dominance effects, δ_m, are 0, 1 and 0, for AA, AB and BB genotypes, respectively and the e_{ij} are the (environmental) errors which are assumed to be independent of each other. Note that this is only true if marker allele transmission from parental lines to progeny is known; if not, then they would be probability values ranging from zero to one, as would be the case for outbred populations.

The phenotypic variance in transcript abundance includes all variances including genetic, environmental and technical variability of microarray or RNAseq experiments. A ratio of genetic variance to the total phenotypic variance gives the fundamental genetic parameter for gene expression traits, namely the heritability (h^2) of gene expression, h^2_{exp} and can be defined as $h^2_{\text{exp}} = (\sigma^2_A)/(\sigma^2_A + \sigma^2_E)$ where σ^2_A is the genetic variance and σ^2_E is the environmental variance. The eQTL effect, (α), can be defined as $\alpha = ((\mu_1 - \mu_2))/(2)$ where μ_1 and μ_2 are the means of gene expression levels of all individuals that respectively inherited AA genotype and BB genotype. The heritability (h^2) of eQTL, h^2_{eQTL}, following Kadarmideen (2008) is defined as

$$h^2_{eQTL} = \frac{\sigma^2_{eQTL}}{\sigma^2_{eQTL} + \sigma^{2*}_A + \sigma^2_E}$$

where, σ^2_{eQTL} is the variance in gene expression attributable to eQTL and $\sigma^{2*}_A = \sigma^2_A - \sigma^2_{eQTL}$.

The σ^2_{eQTL} can, for simplicity, be shown as $\sigma^2_{eQTL} = 2pq\,\alpha^2.\sigma^2_e$ where p and q are the frequencies of two different SNP alleles inherited from parents and σ^2_e is the residual transcript variance. The above formulation means that for 'g' number of genes, we would have 'g' number of h^2_{exp} estimates. Kadarmideen et al. (2006b), Kadarmideen et al. (2011) used this measure of heritability to predict 'estimated breeding values' of animals for gene expression traits (eEBVs) for eventual use in '*expression marker assisted selection* (*e*MAS)'. For instance, Kadarmideen et al. (2011) estimated heritability of 33 biomarker gene expression profiles from microarray data analyses on a sheep gastro-intestinal nematode experiment by random genetic effect mixed models in ASReml (Gilmour et al., 2009). Fig. 6 shows that biomarker gene expression profiles with heritabilities >0.15 (red dashed line) are shown in blue and those <0.15 are shown in pale blue. Of these 33 biomarkers, 33% (11 genes) had statistically significant ($p \le 0.05$) non-zero h^2 ranging from 0.92 for CAT (s.e. 0.52) to 0.02 for LOC51557 (s.e. 0.02). Moderate to

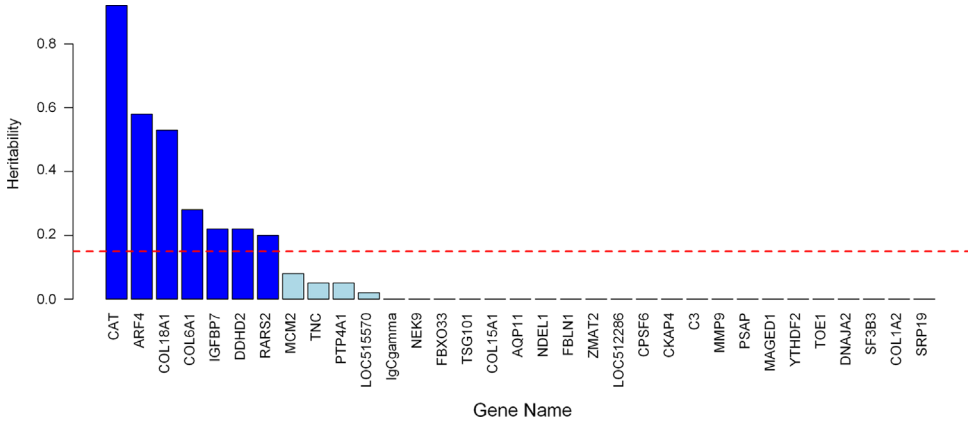


Fig. 6. Heritability of 33 biomarker genes derived from half-sib genetic variances in 10 sire families, estimated by random genetic effect mixed models in ASReml. Biomarkers with heritabilities >0.15 (red dashed line) are shown in blue and those <0.15 are shown in pale blue. From (Kadarmideen et al., 2011). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

high h^2 estimates were found for *CAT*, *ARF4*, *COL18A1*, *IGFBP7*, *DDHD2* and *RARS2*. For six biomarkers with heritabilities >0.15, eEBVs were also estimated. Overall, the above mentioned studies show that quantitative genetics principles and methods are not only applied to genomic, pedigree and phenotypic data but also to endo-phenotypes such as transcriptome or metabolome.

7. Conclusion

Current progress and lessons in livestock breeding and genomics with some applications in cattle, pigs and sheep were briefly discussed followed by new opportunities for transition to integrative systems genetics and systems biology. Particularly, this article focused on transition to next-generation methods to unravel the complete genetic architecture of complex traits and provide biologically meaningful and accurate genomic predictions of performance and disease risks. Methods such as Genome-Wide Epistasis Association (GWEA) and Weighted Interaction SNP Hub (WISH) network methods are recommended for future use in order to capture additional genetic variance arising from genome-wide epistasis and thus could explain part of the missing heritability or improve predictive power. Further, a new genomic prediction and selection method was developed and introduced here: *Systems genomic* BLUP (*sg*BLUP) prediction method that explicitly models SNPs with known biological role as a random effect in addition to conventional random SNP effects in SNP-BLUP or GBLUP methods. As we move towards next-generation GWAS, WGP and GS methods with a focus on biology, tools that link SNPs with biological functions such as *FunctSNP*, *Postgwas* and *NCBI2R* will become important. With a brief background on genome-wide gene expression (transcriptomics) analyses using RNAseq technology, integrative systems genetics and systems biology approaches for animal and veterinary sciences were introduced. Systems biology and systems genetics methods discussed here (and illustrated in Fig. 5) emphasizes that in order to fully understand and identify causal genes or variants and their networks for disease or phenotypic outcomes, it is important for omics studies to link with central theory in biology: from genes through transcription (to mRNA) and translation (proteins or metabolites) to eventual disease or phenotypic outcomes. These integrative approaches will not only become critical in understanding complete causal genetics and the biology of complex traits and diseases, but would also be useful in accurately predicting disease outcomes and phenotypic performance. Throughout this article, some applications of current and new methods in livestock species using data from our own experimental or field studies were provided. Finally, almost all of the next generations -omics tools and methods presented and discussed here are equally applicable to plant and human genetics and systems biology.

Conflict of interest statement

Author declares no conflict of interest.

Acknowledgments

The author is grateful for contributions from his PhD students Ali Abdirahman, Stephen Goodswen, Lisette Kogelman and Duy Ngoc Do. Dr. Sameer D. Pant is thanked for critical review of this article. The author is thankful for the EU-FP7 Marie Curie Actions—Career Integration Grant (CIG-293511) for funding this study.

References

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., Genomes Project, C., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

Ali, A.A., Khatkar, M.S., Kadarmideen, H.N., Thomson, P.C., 2013. Genome-wide epistasis effect on serum insulin-like growth factor-1: two-stage two-locus models. In: The 20th Conference of the Association for the Advancement of Animal Breeding and Genetics (AAABG). Association for the Advancement of Animal Breeding and Genetics 2013, Napier, New Zealand, p. 5.

Ali, A.A., Thomson, P.C., Khatkar, M., Raadsma, H., Kadarmideen, H.N., 2012. Epistasis association mapping for ultrasound carcass traits in tropical beef cattle. In: The Fourth International Conference of Quantitative Genetics: Understanding Variation in Complex Traits. The Genetics Society UK Edinburgh, UK, p. 1.

Bouquet, A., Juga, J., 2013. Integrating genomic selection into dairy cattle breeding programmes: a review. Animal 7, 705–713.

Brown, C.D., Mangravite, L.M., Engelhardt, B.E., 2013. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet. 9, e1003649.

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42, 2.

Civelek, M., Lusis, A.J., 2014. Systems genetics approaches to understand complex traits. Nat. Rev. Genet. 15, 34–48.

Clarke, A.J., Cooper, D.N., 2010. GWAS: heritability missing in action [quest]. Eur. J. Hum. Genet. 18, 859–861.

Cooper, G.M., Shendure, J., 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12, 628–640.

de Koning, D.J., Cabrera, C.P., Haley, C.S., 2007. Genetical genomics: combining gene expression with marker genotypes in poultry. Poult. Sci. 86, 1501–1509.

De Lobel, L., Geurts, P., Baele, G., Castro-Giner, F., Kogevinas, M., Van Steen, K., 2010. A screening methodology based on Random Forests to improve the detection of gene-gene interactions. Eur. J. Hum. Genet. 18, 1127–1132.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., Getz, G., 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28, 1530–1532.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinf. 14, 671–683.

Do, D.N., Ostersen, T., Strathe, A.B., Mark, T., Jensen, J., Kadarmideen, H.N., 2014. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genet. 15, 27.

Do, D.N., Strathe, A.B., Jensen, J., Mark, T., Kadarmideen, H.N., 2013a. Genetic parameters for different measures of feed efficiency and related traits in boars of three pig breeds. J. Anim. Sci. 91, 4069–4079.

Do, D.N., Strathe, A.B., Ostersen, T., Jensen, J., Mark, T., Kadarmideen, H.N., 2013b. Genome-wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. PLoS One 8, e71509.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J. H., 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450.

Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol., 43.

Fu, J., Wolfs, M.G., Deelen, P., Westra, H.J., Fehrmann, R.S., Te Meerman, G.J., Buurman, W.A., Rensen, S.S., Groen, H.J., Weersma, R.K., van den Berg,

L.H., Veldink, J., Ophoff, R.A., Snieder, H., van Heel, D., Jansen, R.C., Hofker, M.H., Wijmenga, C., Franke, L., 2012. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 8, e1002431.
Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8, 469–477.
Georges, M., 2007. Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. Annu. Rev. Genomics Hum. Genet., 131–162.
Gibson, G., 2010. Hints of hidden heritability in GWAS. Nat. Genet. 42, 558–560.
Gibson, G., 2012. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13, 135–145.
Gilmour, A.R., Gogel, B.J., Cullis, B.R., Thompson, R., 2009. ASReml User Guide. Release 3.
Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler, E.S., 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9, e90346.
Goodswen, S.J., Gondro, C., Watson-Haigh, N.S., Kadarmideen, H.N., 2010. FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. BMC Bioinf. 11, 311.
Goodswen, S.J., Kadarmideen, H.N., 2011. SNPpattern: A Genetic Tool to Derive Haplotype Blocks and Measure Genomic Diversity in Populations Using SNP Genotypes. InTech, 24.
Goodswen, S.J., Kadarmideen, H.N., Gondro, C., van der Werf, J.H.J., 2009. A framework to link whole genome SNP association studies to systems genetics. In: Proceedings of the 18th Conference of the Association for the Advancement of Animal Breeding and Genetics (AAABG) 18: 454–457. Sept 2009, Adelaide, Australia., Proceedings of the 18th Conference of the Association for the Advancement of Animal Breeding and Genetics (AAABG), pp. 454–457.
Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., Li, S., Larkin, D.M., Kim, H., Frantz, L.A.F., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C.W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S.-H., Chow, W., Clark, R.C., Clee, C., Crooijmans, R.P.M.A., Dawson, H.D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z.-Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G.J., Fowler, K.E., Fredholm, M., Fritz, E., Gilbert, J.G.R., Giuffra, E., Gorodkin, J., Griffin, D.K., Harrow, J.L., Hayward, A., Howe, K., Hu, Z.-L., Humphray, S.J., Hunt, T., Hornshoj, H., Jeon, J.-T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J., Kim, J.-H., Kim, K.-W., Kim, T.-H., Larson, G., Lee, K., Lee, K.-T., Leggett, R., Lewin, H.A., Li, Y., Liu, W., Loveland, J.E., Lu, Y., Lunney, J.K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M.P., Narayan, J., Truong Nguyen, D., Ni, P., Oh, S.-J., Onteru, S., Panitz, F., Park, E.-W., Park, H.-S., Pascal, G., Paudel, Y., Perez-Enciso, M., Ramirez-Gonzalez, R., Reecy, J.M., Rodriguez-Zas, S., Rohrer, G.A., Rund, L., Sang, Y., Schachtschneider, K., Schraiber, J.G., Schwartz, J., Scobie, L., Scott, C., Searle, S., Servin, B., Southey, B.R., Sperber, G., Stadler, P., Sweedler, J.V., Tafer, H., Thomsen, B., Wali, R., Wang, J., Wang, J., White, S., Xu, X., Yerle, M., Zhang, G., Zhang, J., Zhang, J., Zhao, S., Rogers, J., Churcher, C., Schook, L.B., 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393–398.
He, Y., Li, C., Amos, C.I., Xiong, M., Ling, H., Jin, L., 2011. Accelerating haplotype-based genome-wide association study using perfect phylogeny and phase-known reference data. PLoS One 6, e22097.
Heidt, H., Cinar, M.U., Uddin, M.J., Looft, C., Juengst, H., Tesfaye, D., Becker, A., Zimmer, A., Ponsuksili, S., Wimmers, K., Tholen, E., Schellander, K., Grosse-Brinkhaus, C., 2013. A genetical genomics approach reveals new candidates and confirms known candidate genes for drip loss in a porcine resource population. Mamm. Genome 24, 416–426.
Hiersche, M., Ruhle, F., Stoll, M., 2013. Postgwas: advanced GWAS interpretation in R. PLoS One 8, e71775.
Jansen, R.C., Nap, J.P., 2001. Genetical genomics: the added value from segregation. Trends Genet. 17, 388–391.
Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., de Bakker, P.I., 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24, 2938–2939.
Kadarmideen, H.N., 2008. Genetical systems biology in livestock: application to gonadotrophin releasing hormone and reproduction. IET Syst. Biol. 2, 423–441.
Kadarmideen, H.N., Janss, L.L., 2007. Population and systems genetics analyses of cortisol in pigs divergently selected for stress. Physiol. Genomics 29, 57–65.
Kadarmideen, H.N., Li, Y., Janss, L.L., 2006a. Gene-environment interactions in complex diseases: genetic models and methods for QTL mapping in multiple half-sib populations. Genet. Res. 88, 119–131.
Kadarmideen, H.N., Reverter, A., 2007. Combined Genetic, Genomic and Transcriptomic Methods in the Analysis of Animal Traits. CABI Review: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 2, 16.
Kadarmideen, H.N., von Rohr, P., Janss, L.L., 2006b. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. Mamm. Genome 17, 548–564.
Kadarmideen, H.N., Watson-Haigh, N.S., 2012. Building gene co-expression networks using transcriptomics data for systems biology investigations: comparison of methods using microarray data. Bioinformation 8, 855–861.
Kadarmideen, H.N., Watson-Haigh, N.S., Andronicos, N.M., 2011. Systems biology of ovine intestinal parasite resistance: disease gene modules and biomarkers. Mol. Biosyst. 7, 235–246.
Kadarmideen, H.N., Watson-Haigh, N.S., Kijas, J.W., Vuocolo, T., Byrne, K., Gondro, C., Oddy, V.H., Gardner, G.E., Tellam, R.L., 2010. Genetics of Global Gene Expression Patterns and Gene Networks Affecting Muscling in Sheep. The 9th World Congress on Genetics Applied to Livestock Production (WCGALP) World Congress on Genetics Applied to Livestock Production (WCGALP), 4. (Leipzig, Germany).
Kendziorski, C., Wang, P., 2006. A review of statistical methods for expression quantitative trait loci mapping. Mamm. Genome 17, 509–517.
Kogelman, L.J., Byrne, K., Vuocolo, T., Watson-Haigh, N.S., Kadarmideen, H.N., Kijas, J.W., Oddy, H.V., Gardner, G.E., Gondro, C., Tellam, R.L., 2011. Genetic architecture of gene expression in ovine skeletal muscle. BMC Genomics 12, 607.
Kogelman, L.J., Kadarmideen, H.N., Mark, T., Karlskov-Mortensen, P., Bruun, C.S., Cirera, S., Jacobsen, M.J., Jorgensen, C.B., Fredholm, M., 2013. An f2 pig resource population as a model for genetic studies of obesity and obesity-related diseases in humans: design and genetic parameters. Front. Genet. 4, 29.
Kogelman, L.J.A., Kadarmideen, H.N., 2014. Weighted Interaction SNP Hub (WISH) network method for building genetic networks for complex diseases using whole genome genotype data. BMC Syst. Biol. 8 (Suppl 2).
Kogelman, L.J.A., Zhernakova, D.V., Westra, H., Cirera, S., Fredholm, M., Franke, L., Kadarmideen, H.N., 2014. Systems Genetics Analysis of Obesity using RNA-Seq Data in an F2 Pig Resource Population. 10th World Congress of Genetics Applied to Livestock Production. (Vancouver, Canada).
Koivula, M., Stranden, I., Su, G., Mantysaari, E.A., 2012. Different methods to calculate genomic predictions-Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). J. Dairy Sci. 95, 4065–4073.
Legarra, A., Ducrocq, V., 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95, 4629–4645.
Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2013. Genomic selection for two traits in a maternal pig breeding scheme. J. Anim. Sci. 91, 3079–3087.
Madsen, P., Sørensen, P., Su, G., Damgaard, L.H., Thomsen, H., Labouriau, R., 2006. DMU-a Package for Analyzing Multivariate Mixed Models. 8th World Congress on Genetics Applied to Livestock Production. (Belo Horizonte, Brazil).
Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C. W., Allison, D.B., de los Campos, G., 2011. Beyond missing heritability: prediction of complex traits. PLoS Genet. 7, e1002051.
Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J. H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.
Marchini, J., Donnelly, P., Cardon, L.R., 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. 37, 413–417.
McDowall, M.L., Watson-Haigh, N., Edwards, N.M., Kadarmideen, H., Nattrass, G.S., McGrice, H.A., Hynd, P.I., 2013. Transient treatment of pregnant Merino ewes with modulators of cortisol biosynthesis coinciding with primary wool follicle initiation alters lifetime wool growth. Anim. Prod. Sci. 53, 1101–1111.
Metzker, M.L., 2009. Sequencing in real time. Nat. Biotechnol. 27, 150–151.
Metzker, M.L., 2010. Next generation technologies: basics and applications. Environ. Mol. Mutagen. 51, 691. (-691).

Meuwissen, T., Hayes, B., Goddard, M., 2013. Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci. Vol 1 (1), 221–237.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Misztal, I., Aggrey, S.E., Muir, W.M., 2013. Experiences with a single-step genome evaluation. Poultry Sci. 92, 2530–2534.

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92, 4648–4655.

Mogensen, M.S., Scheibye-Alsing, K., Karlskov-Mortensen, P., Proschowsky, H.F., Jensen, V.F., Bak, M., Tommerup, N., Kadarmideen, H.N., Fredholm, M., 2012. Validation of genome-wide intervertebral disk calcification associations in dachshund and further investigation of the chromosome 12 susceptibility locus. Front. Genet. 3, 225.

Moore, J.H., Ritchie, M.D., 2004. THe challenges of whole-genome approaches to common diseases. J. Am. Med. Assoc. 291, 1642–1643.

Nica, A.C., Dermitzakis, E.T., 2013. Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. London, Ser. B, 368.

Ponsuksili, S., Murani, E., Brand, B., Schwerin, M., Wimmers, K., 2011. Integrating expression profiling and whole-genome association for dissection of fat traits in a porcine model. J. Lipid Res. 52, 668–678.

Pryce, J.E., Daetwyler, H.D., 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. 52, 107–114.

Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., Ernest, S.R., O'Brien, W., Courtland, H.-W., Jepsen, K.J., Kirby, A., Kulbokas, E.J., Daly, M.J., Broman, K.W., Lander, E.S., Nadeau, J.H., 2008. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. Proc. Nat. Acad. Sci. 105, 19910–19914.

Steibel, J.P., Bates, R.O., Rosa, G.J.M., Tempelman, R.J., Rilington, V.D., Ragavendran, A., Raney, N.E., Ramos, A.M., Cardoso, F.F., Edwards, D.B., Ernst, C.W., 2011. Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. PLoS One 6.

The Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324, 522–528.

The International Sheep Genomics, C., Archibald, A.L., Cockett, N.E., Dalrymple, B.P., Faraut, T., Kijas, J.W., Maddox, J.F., McEwan, J.C., Hutton Oddy, V., Raadsma, H.W., Wade, C., Wang, J., Wang, W., Xun, X., 2010. The sheep genome reference sequence: a work in progress. Anim. Genet. 41, 449–453.

Tribout, T., 2014. Efficiency of genomic selection in a purebred pig male line. J. Anim. Sci. 92, 384. (-384).

Tribout, T., Larzul, C., Phocas, F., 2013. Economic aspects of implementing genomic evaluations in a pig sire line breeding scheme. Genet. Sel. Evol., 45.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.

Vijay, N., Poelstra, J.W., Kuenstner, A., Wolf, J.B.W., 2013. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol. Ecol. 22, 620–634.

Watson-Haigh, N.S., Kadarmideen, H.N., Reverter, A., 2010. PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. Bioinformatics 26, 411–413.

Wellmann, R., Preuss, S., Tholen, E., Heinkel, J., Wimmers, K., Bennewitz, J., 2013. Genomic selection using low density marker panels with application to a sire line in pigs. Genet. Sel. Evol., 45.

Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., t Hoen, P.A., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B., Franke, L., 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet., 45; 1238–1243.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., Visscher, P.M., 2013. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14, 507–515.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M., 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., Hill, W.G., Landi, M.T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R.A., Melbye, M., Pugh, E., Cornelis, M.C., Weir, B.S., Goddard, M.E., Visscher, P.M., 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43, 519–525.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., Simianer, H., 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One 9, e93017.