Incorporating Mouse Genome
Mammalian Genome
Genes and Phenotypes

From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding

Haja N. Kadarmideen,[1] Peter von Rohr,[1,2] Luc L.G. Janss[1]

[1]Statistical Animal Genetics Group, Institute of Animal Science, Swiss Federal Institute of Technology, ETH Zentrum (UNS D7), Universitaetstrasse 65, Zürich CH 8092, Switzerland
[2]Institute of Computational Sciences, Swiss Federal Institute of Technology, ETH Zentrum, Zürich CH 8092, Switzerland

Received: 29 November 2005 / Accepted: 21 February 2006

Abstract

This article reviews methods of integration of transcriptomics (and equally proteomics and metabolomics), genetics, and genomics in the form of systems genetics into existing genome analyses and their potential use in animal breeding and quantitative genomic modeling of complex traits. Genetical genomics or the expression quantitative trait loci (eQTL) mapping method and key findings in this research are reviewed. Various procedures and potential uses of eQTL mapping, global linkage clustering, and systems genetics are illustrated using actual analysis on recombinant inbred lines of mice with data on gene expression (for diabetes- and obesity-related genes), pathway, and single nucleotide polymorphism (SNP) linkage maps. Experimental and bioinformatics difficulties and possible solutions are discussed. The main uses of this systems genetics approach in quantitative genomics were shown to be in refinement of the identified QTL, candidate gene and SNP discovery, understanding gene-environment and gene-gene interactions, detection of candidate regulator genes/eQTL, discriminating multiple QTL/eQTL, and detection of pleiotropic QTL/eQTL, in addition to its use in reconstructing regulatory networks. The potential uses in animal breeding are direct selection on heritable gene expression measures, termed ''expression assisted selection,'' and genetical genomic selection of both QTL and eQTL based on breeding values of the respective genes, termed ''expression-assisted evaluation.''

Introduction

Animal breeding, just like its sister technology plant breeding, is concerned with steering the genetic makeup of agriculturally important species to make them better fit for use. Traditionally, this process of adapting species to human needs is known as domestication, and its potential is most vividly demonstrated by the domestication of the dog, which developed from the wolf into an impressive multitude of breeds suitable for herding, hunting, guarding, towing, racing, or just pets.

In the 20th century, the ''art'' of domestication developed into the more scientifically based ''animal breeding.'' In general, there are four ways of steering or changing the genetic makeup of a livestock species. The first well-known approach is classical breeding, in essence still ''domestication,'' in which parents of a next generation are selected based on their (or relatives') phenotypes; this approach has very much improved during the 20th century by the introduction of advanced statistical and computational genetics procedures to better assess heritability of traits and breeding values of animals. The second approach is marker-assisted selection (MAS), or directly changing frequencies in DNA by selecting parents that carry favorable polymorphisms at genetic markers and/or genes associated with economically important traits; this approach has been around since the 1980s. The third approach is the less known genetic modification (or single-gene addition or deletion). Finally, the fourth and latest approach is incorporation of –omics technologies into selection and breeding. As is argued in this article, this –omics approach could fundamentally change the practice of animal breeding, moving away from a basically ''black box'' approach toward an approach that considers the regulatory networks and pathways underlying the expression of important phenotypes.

Correspondence to: H.N. Kadarmideen; E-mail: haja.kadarmideen@inw.agrl.ethz.ch

DOI: 10.1007/s00335-005-0169-x • Volume 17, 548–564 (2006) • © Springer Science+Business Media, Inc. 2006

The –omics technologies will not replace the other approaches of quantitative genetic evaluations of populations or animal breeding but will complement and add to already existing approaches. Our review briefly describes existing joint transcriptome-genome analyses and proposes an integrated –omics approach or so-called systems biology or systems genetics approach (e.g., Cassman 2005; Kitano 2002) in which we consider ways to incorporate data on genetics, genomics, proteomics, and metabolomics into the already existing approaches of selection on phenotypes, selection on DNA polymorphisms, and combinations of both. Finally, possible uses of integrated or systems genomics approaches in quantitative genetic/genomic modeling and animal breeding are discussed.

We complement these discussions with data analysis of microarray gene expression (for diabetes- and obesity-related genes), pathway, and genetic marker linkage data from a BXD set of recombinant inbred (RI) mouse strains, showing in particular the promising combination of transcriptomics and QTL mapping known as ''genetical genomics'' (Jansen and Nap 2001). It is clear that genetical genomics and systems biology have strong multispecies and interdisciplinary components in research and hence we take examples from human, mouse, and yeast models and draw an analogous situation in livestock, if and whenever possible.

Expression QTL mapping by genetical genomics

The concept of *genetical genomics* (GG), or genome-wide genetic analyses of gene expression data, first proposed by Jansen and Nap (2001) and Jansen (2003), is also called *transcriptome mapping*. Genetical genomics has expanded the possibilities for identifying genomic (QTL) regions responsible for variation in gene expression patterns of individuals measured on microarrays, which is the main focus of this issue of *Mammalian Genome*. The basic principle of genetical genomi*cs* is to simultaneously use a segregating pedigreed or resource population for QTL mapping (e.g., F_2, recombinant inbred lines, backcross, half-sib, or full-sib families) and for expression profiling of the whole or a part of the genome. Such a population is studied to find out which genes are being expressed in different individuals and to what degree they vary in their expression patterns. The GG analyses proceed by treating the expression level of each gene on a microarray as a quantitative trait and use genetic markers on the linkage maps to identify genomic regions containing a QTL that affects or regulates gene expression phenotypes. Hence, individual differences in gene expression patterns are associated with sequence differences [e.g., single nucleotide polymorphism (SNP)] between such individuals. The procedure is called *expression QTL* (eQTL) mapping. For each gene (cDNA) or gene product analyzed (e.g., using proteomics and metabolomics) in the segregating population, eQTL analysis would pinpoint the regions of the genome influential for its expression. The suggestive biological meaning of these regions is that genes under eQTL peaks have a significant influence on transcriptional regulations of some of the genes that were probed on the microarray. These eQTL responsible for variation in gene expression could map within the gene itself (*cis-acting eQTL* or *cis-eQTL*) or map to some other location on the genome (in which case they are called *trans-acting eQTL* or *trans-eQTL*). In these locations, GG, combined with SNP mapping data on the same population, reveals not only cis- or trans-eQTL but also identification of SNP markers for expression differences causing changes in expression phenotype (called cis-SNPs or trans-SNPs). In addition to different scenarios, given by Jansen and Nap (2001) and Pomp et al. (2004) with respect to the relationship between eQTL vs. expression phenotypes (graphs with *X-Y* axes in Pomp et al. 2004), the expression profile of a single gene could be affected by many trans-eQTL. Such genes may be difficult to handle because some trans-eQTL may upregulate them while others downregulate them. It is also expected that there could be an abundance of trans-eQTL in one chromosomal region (so called *eQTL hotspots*). If the causal gene underlying a QTL that affects the expression profile of another gene is identified, then a direct link from the causal gene to the expression-profiled gene could be established to indicate a regulatory relationship by joining identified links. Later we perform eQTL mapping and genome-wide linkage analyses of clustered genes by using actual GG data sets on BXD RI mouse strains to identify cis-eQTL/SNP, trans-eQTL/SNP, and eQTL/SNP hotspots.

This combined genetic linkage analysis and expression profiling would be much more powerful than either approach alone, subject to noise reduction in data and control of false positives. The combined GG techniques could reveal a remarkable wealth of quantitative heritable variation in the transcriptome, as shown in human, mouse, and yeast (Bing and Hoeschele 2005; Brem et al. 2002; Schadt et al. 2003, 2005; Yamashita et al 2005). There are many studies (e.g., Brazhnik et al. 2002; de la Fuente et al. 2002; Friedman et al. 2000; Lee 2005) that describe application of transcription profiling by microarrays in constructing gene networks. Complex statistical methods have been proposed in (re)constructing such gene networks (e.g., Soinov et al. 2003; Yeung et al. 2002), but these studies did not capitalize on the use of eQTL on genetic linkage maps. The Jansen and Nap (2001) example of constructing such gene networks using GG was followed by experimental studies in human (Li et al. 2005) and mouse (Bystrykh et al. 2005).

Treating expression data as phenotypic observations leads to the situation that there are tens of thousands of observations (expression traits) per animal. This increases the dimensionality of the eQTL search problem tremendously. Lan et al. (2003) showed how to reduce the dimension of the mRNA abundance mapped as quantitative traits by using principal component analysis and hierarchical clustering to define new traits composed of a small collection of promising mRNAs that can be genetically mapped to identify eQTL. Kraft et al. (2003) demonstrated that the standard unstratified test based on Pearson's correlation coefficient can produce spurious results when applied to family data, and they presented a stratified family expression association test. Bing and Hoeschele (2005) proposed genome-wide QTL analysis of all expression profiles to identify eQTL confidence regions, followed by fine mapping of identified eQTL and then identify regulatory candidate genes in each eQTL region. Furthermore, they proposed a correlation analysis of the expression profiles of the candidate genes in an eQTL region and the gene affected by this eQTL; a substantial reduction in the number of causal genes then allows a finite set of candidate genetic networks to be identified immediately or through performing a small and feasible number of validation studies before network inference. Their methods infer networks by linking regulatory candidate genes to genes affected by the eQTL and joining such links to form networks. Statistical validation and refinement of the inferred network structure would be the final step. They used a segregating yeast population and retained 768 putative regulatory links, 331 of which are the strongest candidate links.

Carlborg et al. (2005) found that the GG approach helps to separate the significant QTL into high- and low-confidence QTL by using a false discovery rate (FDR) that incorporates prior information such as transcript repeatabilities and colocalization of gene transcripts and eQTL. Carlborg et al. (2005) also reported on the adapted QTL mapping methodology to perform automated mapping of QTL that affect gene expression. Li et al. (2005) developed a Bayesian approach that exploits the GG method to focus computational effort on the most plausible gene modulatory networks by exploiting a dense marker map for a genetic reference population that consists of 32 cross-bred strains of mice made by intercrossing two progenitor strains.

Key experiments in eQTL mapping. The GG technique has been quite successful, with its first application in yeast (Brem et al. 2002). A cross between a wild strain and a laboratory strain of yeast was used to identify over 1500 genes that showed differential expression, and eQTL mapping subsequently linked the expression levels of 308 of these genes to one or more genetic loci. Cheung and Spielman (2002) reviewed genetic analysis of expression phenotypes and indicated that this will contribute to our understanding of transcriptional regulation and will provide models for studying quantitative and complex traits. In yeast (Brem et al. 2002) and in the mouse (Schadt et al. 2003), only about 30%–40% of genes are cis-eQTL; the remaining are trans-eQTL. Other GG experiments were conducted in mouse (Schadt et al. 2003, 2005). They found many eQTL at a number of hotspots in the mouse genome, suggesting regulatory elements that may affect the expression levels of a number of obesity-related genes. In humans, Morley et al. (2004) reported genetic analysis of gene expression data from 14 Centre d'Etude du Polymorphisme Humain (CEPH) human families where approximately 1000 expression phenotypes were significantly linked to specific chromosomal regions containing mostly trans-eQTL. Correa and Cheung (2004) reported extensive genetic variation in transcriptional response to radiation exposure. Pomp et al. (2004) provided a number of examples of trans-eQTL that affect expression of genes related to obesity and immune response in humans and mice. Monks et al. (2004) measured the expression of 23,499 genes in lymphoblastoid cell lines for members of 15 CEPH human families. Of the total set of genes, 2340 were found to be expressed, of which 31% (762 genes) had significant heritability (from 0.1 to 1.0, with most genes having 0.3) when a FDR of 0.05 was used. They detected eQTL for 33 genes, of which 13 possessed a QTL within 5 Mb of their physical location, probably indicating closely linked trans-eQTL.

Yaguchi et al. (2005) reported candidate genes for type 2 diabetes modifier loci using expression profiling of segregating populations of diabetic F_2 progeny measured for susceptibility to diabetes and obesity. Palmer et al. (2005) identified behavioral genes involved in drug-abuse liability via GG methods using segregating populations of two mouse lines for high or low methamphetamine-induced activity. They detected expression differences for several genes, including casein kinase I epsilon (*Csnkle*). They then used the expression phenotypes to identify eQTL for *Csnkle* on Chromosome 15 (LOD = 3.8) that comapped with an eQTL for the methamphetamine stimulation phenotype (LOD = 4.5), suggesting that a single allele may cause both traits. Some earlier studies did conduct experiments similar to GG: Liu et al. (2001) identified 15 genes that showed differential expression between the resistant and susceptible lines of chickens to Marek's viral disease and subsequently mapped at least one of these genes to a known QTL that affects resistance to Marek's disease. Eaves et al. (2002) used GG to study genes underlying diabetes in a congenic mouse model and identified eight new candidate genes for one major gene that confers resistance to diabetes. Karp et al. (2000) identified 21 genes that were differentially expressed on exposure to an allergen between mouse strains with a high allergic response and strains that were low responders and mapped one of these differentially expressed genes to one of the two previously identified QTL. In all the above GG studies, it became clear that conclusions from data analyzed may indeed be different from one experiment to another experiment if different normalization and background adjustments were done and that most differentially expressed genes need to be confirmed by qRT-PCR methods wherever possible.

Toward systems genetics and systems genomics

Pathways. In the near future, livestock microarrays will not be very important in finding out regulatory systems because they are too expensive. An alternative scenario could be that, given all the pathway data that we have in humans or mice, we can infer from those results the structure of regulatory systems by homology to, e.g., pigs and cattle. A few confirmatory experiments can be well planned and therefore are more cost-efficient. We also expect that generalizability across species would increase, because exact gene effects may be less replicable, whereas effects of pathways are replicable.

To match gene expression patterns controlled at one genomic location with potential QTL in a different genomic location (trans-eQTL), pathway data could be very useful and sometimes needed for verification. Conversely, for pathway construction, gene expression analysis could be very useful in bringing additional evidence to point to one particular gene because this gene is also differentially expressed or matches a pathway that is differentially expressed. Therefore, pathway and gene expression data analyses are interdependant and could be mutually beneficial. The concept of using regulatory pathway, QTL map, and gene expression databases is that once a gene (e.g., gene *A*) is found to be differentially expressed but it does not reside within a QTL region (which contains other multiple candidate genes such as *B, C, D, E*,...), then a search can include pathway information. Pathway data can reveal a possible regulator (e.g., gene *B*) of a differentially expressed gene (gene *A*), which in turn resides within a QTL region. Hence, through the gene expression and pathway data, multiple candidates at the QTL region can be discriminated. With regard to clustering to use pathway information, common "unsupervised" clustering techniques (e.g., *k*-means, principal component approach) are generally not fully rewarding because the pure statistical association brings little biological significance to the clusters being made. More useful clustering can be obtained using "supervised" clustering techniques. For instance, this can be applied to combine gene expression data and various bioinformatics data sources into gene identification and eQTL mapping tools. In this approach, several sources of data (e.g., pathway databases, sequences, and literature) are used as "priors" in clustering gene expressions, thus adding information and cause-effect relationships that would otherwise not be available from pure statistical association.

Integrated genomics. The integrated genetics and –omics data could be helpful in studying the functions of causal genes underlying QTL regions. Most reported QTL in animals have large confidence intervals possibly harboring hundreds of genes. This is the biggest obstacle in finding genes or SNPs underlying identified QTL in livestock. Two steps to reduce this obstacle are, first, reduce the lengths of the initial QTL regions to, say, 1–2 cM by using existing fine-mapping techniques [e.g., RI line mapping, joint linkage disequilibrium (LD) and linkage mapping, interval-specific haplotype analyses, additional genotyping of loci and individuals in regions of interest] and, second, do positional cloning of QTL. DiPetrillo et al. (2005) proposed that one can investigate the fine mapped regions (say < 5 Mb) for the presence of a few strong candidate genes and evaluate whether the presence of polymorphisms in each one of those genes affects gene expression or function, using integrated bioinformatics approaches. They suggested querying the breed-specific transcriptomic, sequence, and proteomic databases for various types of tissues, if available. Mootha et al. (2003) have shown how such integrated –omics data sets can help identify and isolate a single gene called *LRPPRC* (among multiple candidates) that causes Leigh syndrome or COX deficiency in humans. Recently, Schadt et al. (2005) used the integrated genomics approach by means of cis-eQTL, trans-eQTL, and normal QTL data on obesity in mice to study intermediate gene expression phenotypes. They were able to identify and validate three new genes that cause susceptibility to obesity. In livestock species, efforts are underway to build such comprehensive databases; however, existing databases of closely related information-rich species such as mouse and human would be helpful, as illustrated below with the BXD RI mouse database of the *Ins1* gene.

Modular networks. One of the applications of the microarray-based expression profiles of entire genomes was to decipher the regulatory network of an entire organism. Friedman et al. (2000) published the first regulatory network spanning the whole genome of *Saccharomyces cerevisiae* (baker's yeast). This network was reconstructed from a series of microarray data using Bayesian networks. This approach suffers from the conceptual problem of trying to infer a complex network structure and, hence, a large number of unknown parameters from only a few observations. Besides the expression profiles of entire genomes, more data on genomic sequences became available in public databases. A good number of links can be seen at the NCBI website (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) and the EMBL website (http://www.ensembl.org/index.html).

These sequences together with functional characterizations of many protein families such as transcription factors constitute a large source of additional information. This information can be used besides the gene expression data to infer regulatory networks. Based on this idea of including prior information in the process of network inference, Segal et al. (2003) presented the so-called *module network algorithm*. The module network algorithm is based on probabilistic graphical models and uses gene expression data and an *a priori* known classification of genes into candidate regulators and nonregulators. These two sources of information are combined to infer *regulatory modules*. A regulatory module is a set of genes that are regulated by a shared regulation program. A regulation program describes the behavior of the nonregulator's gene expression in the module as a function of the expression levels of a small set of regulator genes. The algorithm takes as input a precompiled set of candidate regulator genes containing transcription factors and signaling molecules. Based on this input, the nonregulator genes are partitioned into modules according to their expression values and a regulatory program is searched for each of the newly created modules. The regulation program for each module specifies the set of regulator genes associated with the given module and describes the behavior of the expression values of the nonregulator genes in the module as a function of the expression values of the regulator genes. The two steps of establishing regulatory programs and reassigning genes to modules are iterated until convergence is reached.

The use of additional information about regulators in the module network algorithm reduces the dimensionality of reconstructing regulatory networks tremendously. Furthermore, by restricting only regulators to be parent nodes in the inferred networks, the biological significance of the resulting networks is increased. The module network algorithm assumes that regulators (transcription factors, signaling molecules) are themselves transcriptionally regulated. Any other regulation process such as post-transcriptional modifications cannot easily be quantified with this approach. The module network algorithm, in principle, can integrate genetic linkage data on fine-mapped trans-eQTL to support or form regulatory modules. However, this research needs further investigation. This integrated or systems genetics approach is attractive for animals because experimental designs (sample size) for animals tend to be smaller than those for plant breeding. However, opportunities to use comparative information from other well-studied mammals will be greater.

eQTL mapping using BXD mouse genome-transcriptome databases. In this section of the article we conduct a whole-genome-wide scan for eQTL for expression phenotypes from transcription of genes involved in type I and type II diabetes and obesity: insulin I (*Ins1*), insulin II (*Ins2*), and solute carrier family 2 (*Slc2a5*; a facilitated glucose transporter member 5). The objective of this investigation was to illustrate the various uses of integrated transcriptome-genome-pathway analyses (or equivalently systems biology) in quantitative genomics and animal breeding. The data set is based on a BXD set of RI mouse strains (obtained from WebQTL, http://www.genenetwork.org/) to show how to infer regulatory networks using interval and composite interval mapping of eQTL for *Ins1*, *Ins2*, and *Sclc2a5* in the mouse and by using pathway and proteomics databases. The interval eQTL mapping was done using software available at the WebQTL website that is suitable for crosses between inbred lines (e.g., Haley and Knott 1992) or for outbred lines on a within-family basis (e.g., Kadarmideen et al. 2000). The genome-wide significance testing was done using the methods of Churchill and Doerge (1994), and bootstrap samples for eQTL location were drawn

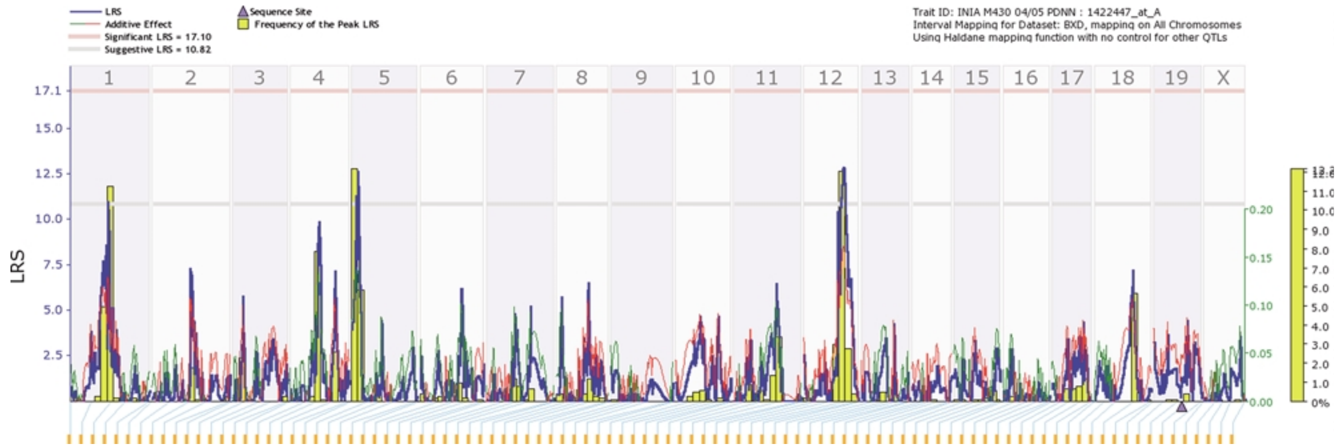

Fig. 1. Whole-genome scan for eQTL that influence gene expression of insulin (*Ins1*) on Chr 19 (at 51.83 Mb) using mouse BXD recombinant inbred line data from INIA Brain mRNA M430 PDNN database of WebQTL (trait ID: 1422447_at_A,). Effects of QTL alleles on gene expression and likelihood ratio statistics (LRS; based on 1000 permutation tests) are plotted. (The additive effect is half the difference in the mean phenotype of all cases that are homozygous for one parental allele at this marker minus the mean of all cases that are homozygous for the other parental allele at this marker. In the case of BXD strains, e.g., a positive additive effect indicates that DBA/2J alleles increase trait values. Negative additive effect indicates that C57BL/6J alleles increase trait values.)

based on the work of Visscher et al. (1996) . Further details on software and statistical methods used are described in Wang et al. (2003). In addition to eQTL mapping, we show how to conduct homology mapping by querying databases in Ensembl using mouse eQTL regions to find candidate regions in livestock and humans. The same database is used in a section below to illustrate the use of eQTL technology in quantitative genetic modeling and animal breeding.

The source data set. The BXD RI strains were derived by crossing C57BL/6J (B) and DBA/2J (D) and then inbreeding progeny for over 21 generations. This set of RI strains is a remarkable resource because many of these strains have been phenotyped extensively for hundreds of interesting traits over a 25-year period. A significant advantage of this RI set is that the two parental strains (B6 and D2) have both been extensively sequenced and are known to differ at approximately 1.8 × 10^6 SNPs. Coding variants (mostly SNP and insertion deletions) that may produce interesting phenotypes can be rapidly identified in this particular RI set.

INIA Brain mRNA M430 (April05) PDNN database. This data set, publicly available at http://www.genenetwork.org/, provides estimates of mRNA expression in adult forebrain and midbrain from 45 lines of mice including C57BL/6J and DBA/2J, their F_1 hybrids, and 42 BXD RI strains. Samples were hybridized in small pools ($n = 3$) to a total of 105 Affymetrix M430 A and B array pairs. Physical maps in WebQTL incorporate approximately 2 × 10^6 B vs. D SNPs. These strains and advanced intercross progeny from this RI and many of the 50 new BXD strains are available from The Jackson Laboratory (http://www.jax.org/). Most BXD animals were born and housed at the University of Tennessee Health Science Center (http://www.utmem.edu/). Detailed description of the above and all other details about the tissue used to generate this data set and background information is at http://www.genenetwork.org.

Figure 1 shows results from a whole-genome scan for eQTL influencing gene expression of insulin (*Ins1*) on Chr 19 (at 51.83 Mb) using mouse BXD RI data. The sequence site of *Ins1* is shown by a triangle on Chr 19. Peaks in the heavy blue line (LRS) show locations of a putative QTL, and the yellow histogram beneath it shows frequent peak location for bootstrap samples. Permutation-based significance thresholds are given by dashed lines; the upper line corresponds to a genome-wide 5% significance threshold. A positive additive coefficient (green line) indicates that DBA/2J alleles increase trait values. In contrast, a negative additive coefficient (red line) indicates that C57BL/6J alleles increase trait values, in this case, expression of the *Ins1* gene. It is clear from the whole-genome eQTL mapping results that there are three locations (Chr 1, 5, and 12) that are above the suggestive LRS of 10.82 but none over the genome-wide significance threshold (LRS) value of 17.10. Each point on the *x* axis of this LRS profile is a marker, the exact identity of which is readily seen in WebQTL. For our results, marker loci flanking these eQTL peaks are identified as

- Chr 1 between marker loci *D1Mit216* at 80.14 Mb and CEL-1_98681809 at 98.57 Mb
- Chr 5 between marker loci *mCV23582150* at13.27 Mb and *rs13478123* at 17.66 Mb
- Chr 12 between marker loci *D12Mit259* at 87.26 Mb and *rs13481611* at 95.01Mb

Hence, we see here a common picture from eQTL mapping, where expression of *Ins1*, which itself is a locus on Chr 19 (at 51.83 Mb), is associated with polymorphisms in several genomic locations different than its own location on Chr 19. To make regulatory inference on this eQTL map, these three trans-eQTL regions can be viewed as candidate regions that contain regulatory genes. The eQTL mapping to detect these trans-eQTL was performed without fitting other markers as cofactors in the analysis, in other words, without composite interval mapping to suppress the effects of "ghost eQTL." With composite mapping for *Ins1*, we found that none of the eQTL peaks reached suggestive LRS, but a peak in Chr 6 appeared (results not shown).

To investigate the mapped eQTL in detail, chromosome-wide genetic linkage and physical maps (e.g., for Chr. 12) are better visualized as in Fig. 2. It is clear from the Chr 12 map that the eQTL allele from C57BL/6J increases expression of the *Ins1* gene (red line) with almost no allelic effect from the DBA/2J line. It is also seen that the eQTL region mapped could indeed contain more than one eQTL judging from bimodal distribution of LRS, indicating a need for further fine mapping of eQTL.

Comparative eQTL mapping and candidate gene search. We considered specific eQTL peaks and investigated physical maps for further inference on regulatory network and comparative eQTL mapping, including investigation of the possible sequence variations and candidate gene identification. Figure 2 gives a physical map of eQTL regions in Mb. It is seen that the first eQTL peak is between 80 and 90 Mb and the second peak is between 95 and 110 Mb. There were 15 known candidate genes present between approximately 90 and 95 Mb on Chr 12 (names not shown). Similar searches yielded 31 genes on Chr 5 (approximate range = 90–95 Mb) and 15 genes on Chr 1 (approximate range = 9–15 Mb). Considering, for instance, 80–110 Mb on Chr 12 as one region, a synteny mapping can be initiated by searching: This is not restricted to homology mapping only but, in general, we have to find the location in the livestock genome (pig, chicken, or cattle) that is somewhat related to the candidate regulator regions that we mapped in mouse between 80 and 110 Mb on Chr 12. By such synteny mapping, results would indicate candidate regulator regions in the livestock species of interest. By comparative mapping, the first region on mouse Chr 12 maps by synteny to chicken Chr 3 and human Chr 2 (not shown).

Pathway analysis. It is important to link the gene expression data with proteomics and pathway databases to be able verify and ascertain candidate gene or pathway networks. Querying STRING (Search Tool for the Retrieval of Interacting Genes/Proteins; http://string.embl.de/) for *Ins1*, retrieved gene networks involved for the protein encoded by the *Ins1* gene were glucagon precursor (Gcg), insulin 2 precursor (Ins2), lipoprotein lipase precursor (Lpl), gastric inhibitory polypeptide precursor (Gip), glucagon-like peptide 1 receptor precursor (G1p1r), 85-kDa calcium-independent phospholipase a2 (Ipla2), hexokinase d (Gck), insulin promoter factor 1 (Ipf-1), and serine/threonine-protein kinase (Sgk1). From this gene network, one can validate detected trans-eQTL peaks on Chr 1, 5, and 12 and ascertain whether those eQTL indeed map into physical regions of these regulators (i.e., checking whether peaks in Fig. 1 correspond to any one of the regulatory gene network nodes from pathway analyses).

Whole-genome linkage cluster analysis. Genome-wide cluster maps are sets of eQTL heat maps for a cluster of expression traits that are analyzed side by side to enable easy detection of possible common eQTL. Traits (expression values) are clustered along one (x) axis of the heat map by similarity of expression phenotypic values based on hierarchical clustering. More tightly correlated expression traits are close together. The longer (y) axis plots genome location, SNP by SNP, from Chr 1 to Chr X. Colors are used to encode the probability of linkage and the additive effect of polarity of alleles at each marker.

Figure 3 shows the results of such a linkage cluster analysis in the form of a heat map. The vertical bar right below each gene on the x axis contains an orange triangle that shows the chromosomal location of that gene itself. Based on the correlation and distance analyses, these genes are clustered together side by side. Results are shown for the top 100 most correlated genes based on their expression phenotypes in a hierarchical cluster tree. This tree was built based on distances computed using $1 - r$, where r is the Pearson correlation values. On the y axis all the chromosomes from 1 to 19 and X are aligned. The regions in the map with high color intensities mark chromosomal regions with high linkage statistics. Horizontal colored bars in the heat

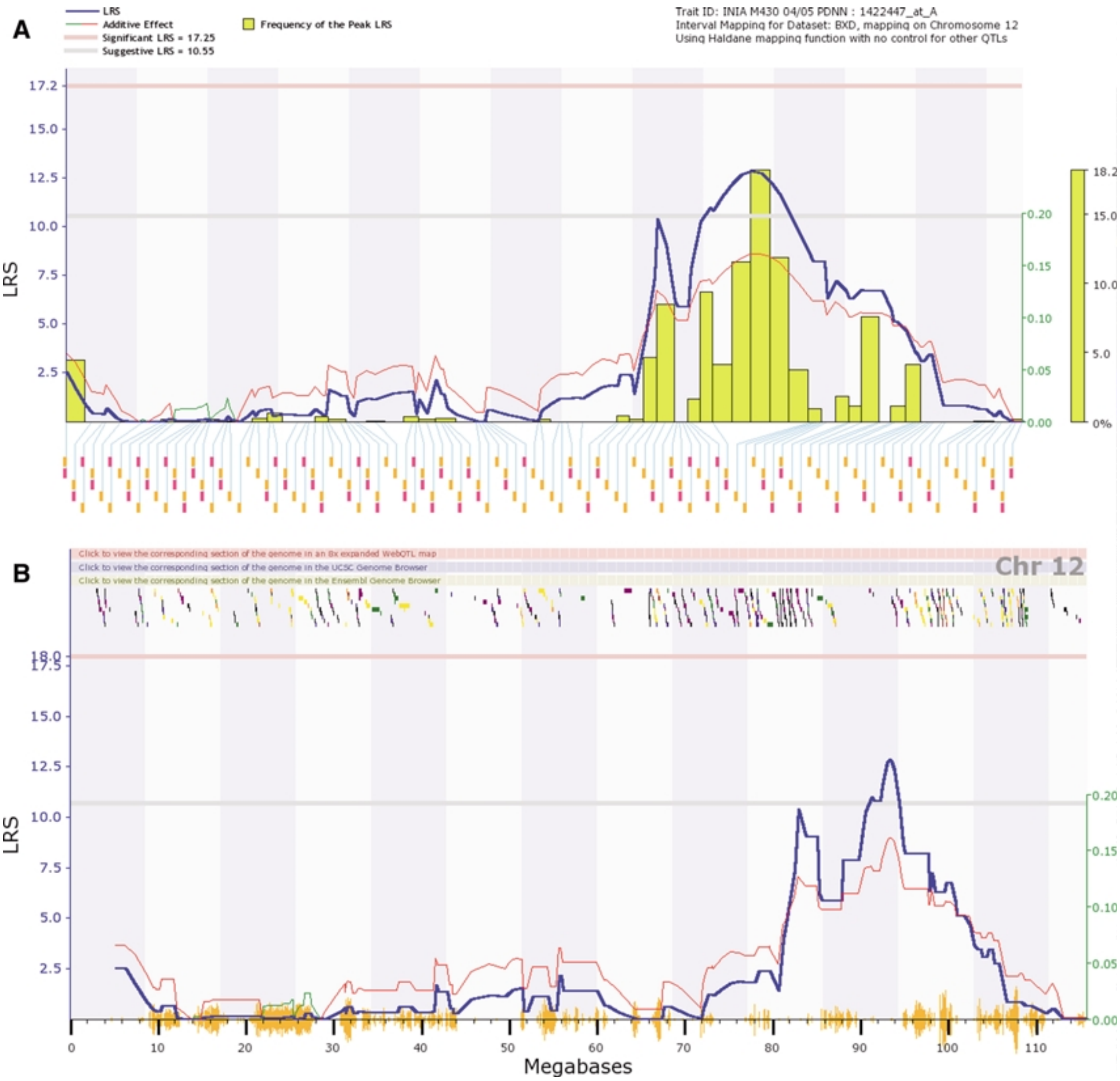

Fig. 2. Detailed view of eQTL that affect expression of the *Ins1* gene that maps to Chromosome 12 showing genetic linkage map (**A**) and corresponding physical map (**B**).

map denote expression trait values of a group of genes all of which map into the same chromosomal region. Based on the data shown in Fig. 3, we could identify six chromosomal eQTL regions (Chr 1, 2, 4, 5, 12, and 18) into which the majority of the shown expression values map. More regions important for the expression trait values of smaller groups of genes can be identified easily.

A given chromosomal region identified by the horizontal colored bar is important for the regulation of the expression values of the genes across which the bar extends. Hence, one could speculate that this chromosomal region contains genes that are important in the regulation of the expression of other genes contained in the horizontal colored bar. Based on this observation, the identified chromosomal regions are good candidates for searching regulatory genes. The identified chromosomal regions together with the genes contained in the horizontal colored bars can also be used as input for more advanced approaches of reconstructing regulatory networks such as the module network algorithm.

Main challenges

Experimental issues. At this time, the design and the analysis of microarrays to produce the expression measurement are far from routine. The price for good-quality microarrays or GeneChips is still in the area of $400 US per array for an Affymetrix array (http://www.affymetrix.com/products/arrays/index.affx),

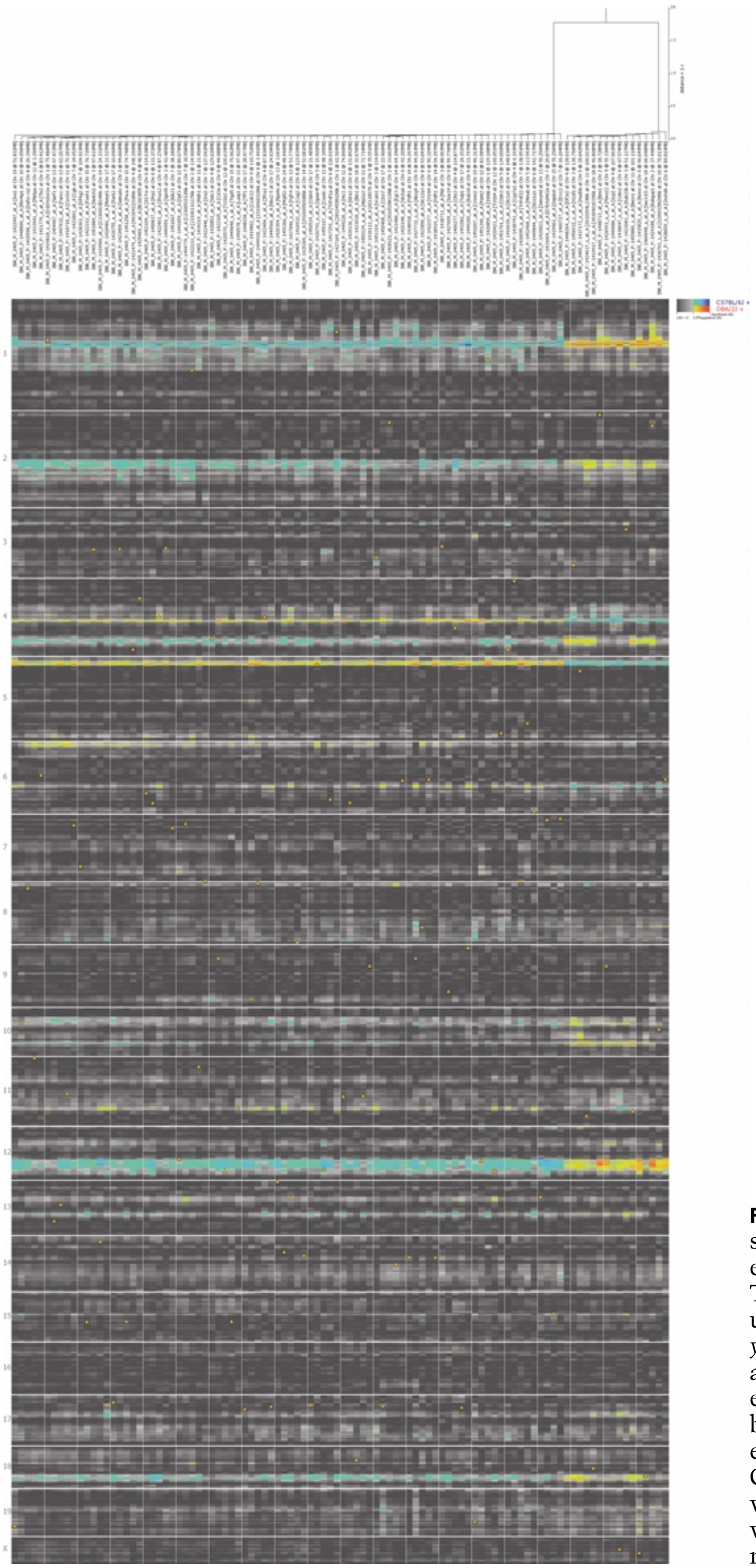

Fig. 3. Global display of linkage clusters. The *x* axis shows the 100 most correlated genes based on their expression phenotypes in a hierarchical cluster tree. This tree was built based on distances computed using 1 – *r*, where *r* is the Pearson correlation. On the *y* axis all the chromosomes from 1 to 19 and X are aligned. The colors in the heat map symbolize different values of the linkage statistics. For example, blue-green regions are those with higher phenotypic expression values associated with the parental allele C57BL/6J and the red-yellow regions are those in which the other parental allele DBA/2J is associated with high trait values. Gray and black areas in the map have insignificant linkage statistics.

or about $100 US for a spotted array (but with less features). This price is certainly too expensive for using arrays in routine evaluations. Using this technology also requires that tissue samples be taken in sufficient quantities. From those samples a large amount of mRNA has to be extracted. These technologic requirements impose some limitations on the potential practical implementation of the eQTL technology. Besides, the main implementation problem would be the cost of *microarraying* animals, especially with livestock populations where a large number of animals are recorded for quantitative traits.

Bioinformatics issues. There are notable review articles and books on the design and analysis of gene expression microarray data (Parmigiani et al. 2003; Sebastiani et al. 2003; Yang and Speed 2002), including normalization, differential gene identification, unsupervised clustering, and supervised classification methods. There is an ongoing list of new approaches also investigated by many others (e.g., Bing et al. 2005; Moser et al. 2004; Qin and Kerr 2004; Tempelman 2005). With respect to including expression phenotypes in traditional best linear unbiased prediction (BLUP) genetic evaluation, we will have substantial dimensionality problems that even the most recent advances in computing power would not alleviate. Given the multiple-testing problems (and correlated tests) in gene expression data analysis, as well as multidimensionality problems, setting significance threshold for identified genes is a complex but crucial point. A ''paradigm shift'' that may be needed here is to change from the use of *p* values to the use of FDR (Fernando et al. 2004). The application of FDR is now common in the analysis of gene expressions and could also be applied in QTL mapping and other multiple-testing problems where the adherence to *p* values now results in large false-negative rates.

Common approaches for analysis of microarray data imply the sequential application of a large number of statistical estimation and correction procedures, which can accumulate errors in the pipeline. Also, bioinformatics data that may be used to ultimately annotate genes have many sources of potential error: The largest databases are the least curated (so poor data are overwhelming good data), and making comparative links between species (on which animal breeding will largely rely) can cause errors because of inaccuracies in comparative maps and mistakes caused by gene duplications. Also, the matching of traits between species can be risky; to do so properly, good trait ontologies should be developed. Two solutions to the problem of accumulating errors in pipelines are to perform more integrative analyses (which may be feasible in some areas but not in all) and to assign levels of confidence to information, which could be used in further steps. Meta-analysis tools or Bayesian modeling could be used to sequentially update knowledge and uncertainty. To perform the last steps to ultimately identify genes underlying eQTL, a large array of tools and (genomics) data would have to be combined and streamlined. In general, more relaxed (in using FDR) and better (in pipelines) approaches to handle errors should come into use. The ultimate identification of genes, but also the systems genetics and bioinformatics information on pathways, would ultimately help breeders better understand gene effects, make eQTL results more generalizable across populations, and so devise more robust selection programs based on molecular data. While the GG/eQTL technology offers an insight into the putative location of genes that are important in running an organisms' regulatory program, there are also some limitations. Most studies on eQTL analyses published to date (e.g., Bing and Hoeschele 2005; Yamashita et al. 2005) have assumed the expression data to be independent phenotypic observations. However, it is widely known that expression data are highly correlated. Ignoring the correlation structure in these data might be one reason for the high number of significant eQTL.

Animal breeding issues. During the last ten years, a plethora of genetic data (genetic markers, QTL, and candidate genes) on economically important traits in animals became available. The main problem is to integrate this information into traditional genetic evaluation (BLUP) programs, which already incorporate large amounts of phenotypic and pedigree information. There is no common opinion or method on how one should optimally integrate information from marker-QTL data into a genetic evaluation and breeding program but several options have been proposed (Dekkers 2004). Application of GG and systems genetics methods in animal populations would potentially lead to identification of a set of differentially expressed genes (e.g., for a given disease or an economically important trait), normal QTL/SNPs, trans-eQTL/SNPs, and candidate master regulators. To use this large amount of information from the systems genetics, we need to integrate these data into routine genetic evaluations, much like marker-QTL data, possibly in the BLUP framework. This will be a major challenge for breeders in the near future. There is some speculation on some potential use of microarray data by breeders by Walsh and Henderson (2004) but not so much at the level of GG and systems biology. Given that there is a way to incorporate all data in genetic evaluations, we envision some form of a genetical genomic selection index in the future that could be used in selecting animals based on traditional QTL effect estimates from genome scans with their (favorable) ''expression value'' (in a given condition and time). This could be important in making time-specific selection decisions in animal breeding, e.g., if we are interested in studying mastitis resistance in cows, we might put more emphasis on eQTL related to the expression of resistance genes in udder tissues. The open question or challenge to using the expression data is to what extent are these data *reliable and repeatable* because the data on gene expression are very specific (down to a cell type, time, etc.) and may have very high sampling variance and/or not be replicable.

Potential uses of eQTL mapping in quantitative genomics

Candidate gene and SNP discovery. Whole-genome genotyping tools based on SNP markers are now available as Microarray-based genotyping arrays (from http://www.affymetrix.com/ and http://www.perlegen.com). Such arrays allow genotyping individuals for the entire genome with tens of thousands of SNP markers in a single hybridization step, thus significantly increasing the throughput and decreasing the cost of current gel-based techniques for molecular mapping. Statistical genetic evaluation of animals with genotype information at tens of thousands of SNPs and phenotypes allows an association between animals' phenotypes and SNP markers to be estimated, virtually resulting in a genome-wide genetic evaluation of animals. When combined with linkage studies (genetical genomics), it helps to disentangle the fine-mapped QTL linkage blocks (with potentially 100 genes) into a few candidate genes and SNPs and study their effects on phenotypes. This dissection is not possible with typical association studies of polymorphisms at these loci with phenotypes because such methods would not be able to distinguish between the large number of SNPs and candidate genes within such a linkage block. The other advantage of eQTL mapping is that it would also show trans-eQTL regions and trans-SNPs that affect expression of a QTL or a candidate gene under study. This approach naturally offers identification, isolation, and characterization of economically important causal cis-mutations within QTL as well as trans-acting genes and trans-acting SNPs within eQTL regions. This is advantageous for breeding applications because it allows simple selection tools to enhance milk or meat production with quality or disease resistance for which some evidence of segregating a major gene or a QTL already exists (e.g., Ilahi and Kadarmideen 2004; Kadarmideen and Janss 2005). Furthermore, such identified genes and markers can serve as direct markers and provide excellent candidates for future translational genetic studies and for biomedical, pharmacogenetic, and agribiotechnological interventions.

Understanding gene (QTL) × environment interactions. Genotype × environment interaction (G × E) may be present if a difference in trait performance (e.g., heat or disease resistance) between two genotypes (e.g., breeds) in one environment (e.g., temperate climate) is not the same in another environment (e.g., tropical climate). Livestock species such as dairy cattle are often imported or exported by importing or exporting semen, embryos, and animals. Recently, the detection and mapping of QTL with GxE in complex diseases was described and is an emerging field in quantitative genomics and animal breeding (Kadarmideen et al. 2006). In the case of genome scans that identified significant interaction of QTL with environment, one can further understand the biology of GxE and validate such interacting QTL by GG approaches. This can be performed by testing the expression patterns of such QTL in microarrays specific for different environments (e.g., cDNA arrays made up of udder tissue mRNA and testing the expression of mastitis QTL in a high- versus a low-hygiene environment). The GG approach can help in identifying eQTL which up- or downregulates expression of normal QTL or candidate genes depending on the environment. The ultimate use of understanding GxE at QTL and identifying trans-eQTL involved in these phenomena would be in weighing the relative importance of genes and making more appropriate genetic evaluations and decisions.

Genetical genomics can also help one understand behavioral or stress responses of animals that are often difficult to measure and evaluate. Such responses can be invoked by environment, which, in principle, includes all nongenetic factors such as low or high nutrition, housing comfort, altitude, temperature, humidity, and exposure to pathogens. In yeast, for instance, Gasch et al. (2000) and Causton et al. (2001) used microarray technology to evaluate gene expression patterns in cells exposed to different environmental conditions and concluded that yeast responds to a range of environmental challenges with differential expression of a common set of genes described as the environmental stress response, while other genes were specialized for specific conditions or stressors. Expression data on environmental stress responses of animals exposed to environmental challenges would indicate how animals adapt to different conditions. This information can be used in selection based on stress-

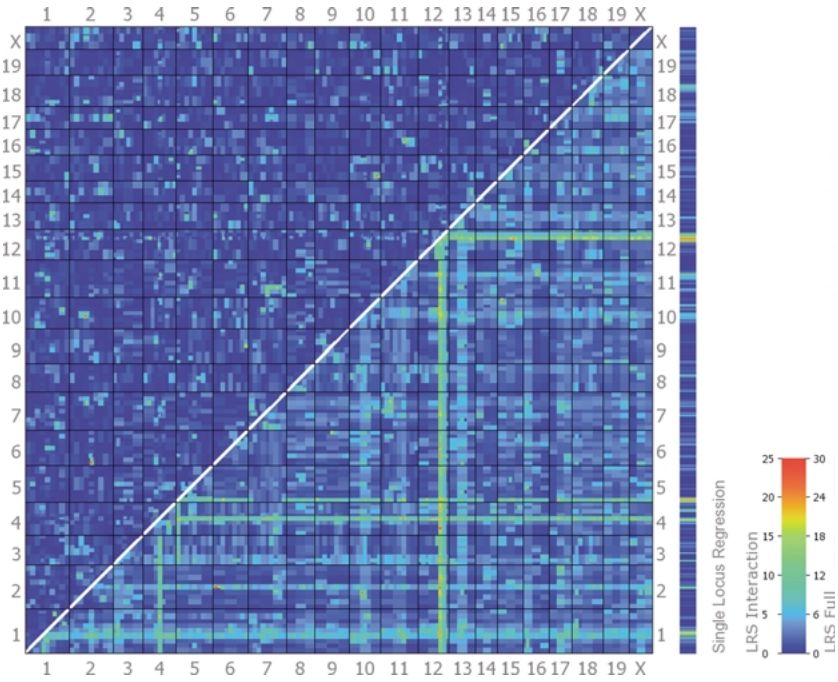

Fig. 4. Whole-genome-wide pairwise scanning of epistatic genes influencing gene expression of insulin 1 (*Ins1*) on Chr 19 (at 51.83 Mb) using mouse BXD recombinant inbred line data from INIA Brain mRNA M430 PDNN database of WebQTL (trait ID: 1422447_at_A). The upper left half of the plot highlights any epistatic interactions (corresponding to the column labeled ''LRS Interact''). In contrast, the lower right half provides a summary of LRS of the full model, representing cumulative effects of linear and nonlinear terms (column labeled ''LRS Full'') based on WebQTL.

specific gene expression patterns and perhaps combined with identified trans- and cis-eQTL on linkage map as ''candidate genes.''

Understanding and detecting epistatic interactions. The same principle of understanding GxE interaction at QTL also applies to elucidating patterns of QTL-QTL (epistatic) interactions at the population level and down to gene expression levels. In this case, all epistatic QTL, by definition, would be detected as a trans-eQTL from a microarray study. The major drawback of QTL mapping is that a multidimensional grid search is needed to estimate and detect interacting QTL; some improvements were proposed (e.g., Carlborg and Andersson 2002; Carlborg et al. 2004) but the computational and modeling difficulties still exist. Additional small but interacting QTL would be missed (false negatives) unless a model explicitly accounting for such interactions is used (shown by Carlborg et al. 2004; Kadarmideen et al. 2006). The GG approach can help in the following way: First, it helps to identify with which genes in the entire genome the identified gene is interacting, and, second, how the expression of the identified gene is affected by gene-gene interactions in the rest of the whole genome. For this we need to identify expressed genes within the QTL region by querying expressed sequence tag (EST) databases and synteny mapping. Then one can study how expression of these genes is affected by other pairs of genes. Thus, it is a three-way interaction: The first is an expression measurement on the genes and the second and third genes are interaction pairs. Microarray-based eQTL mapping could be very helpful here in showing epistatic hot spots. With mouse data we show that the GG approach reveals epistatic eQTL using data on 1.8 × 10^6 SNPs in Fig. 4. In Fig. 4, the interaction of eQTL with itself is represented by white diagonal lines. Off-diagonal elements represent pairwise epistatic interactions of any genes throughout the genome that affect the expression phenotype of *Ins1*. The full epistatic model fitted a pair of genes as a main eQTL effect (linear regression) and interaction between genes in this pair as a nonlinear regression in the model (LRS full). The result from LRS full is below the diagonal, which represents both the main and epistatic interactions, but it is difficult to distinguish whether the significance found is due to main effects or interactions effects or both. Therefore, the area above the white diagonal line represents ''pure'' epistatic interactions (LRS interact). Whether above or below the diagonal, the significance of epistatic interactions is seen from the intensity of the color (red being the strongest and blue being the weakest epistasis). As expected, based on Figs. 1 and 2, Chr 12 has the most interacting pairs of genes that affect the expression of *Ins1* (light green and red spots) followed by Chr 5 and Chr 1. All light green and red areas above the diagonal represent significant pairwise interactions of genes throughout the genome that affect the expression of the trait (*Ins1*).

Detection of candidate regulator genes and eQTL. The global linkage cluster analyses of gene expression data (Fig. 3) combined with the results of GG analyses (Figs. 1 and 2) can reveal chromosomal locations with candidate regulator genes and transcription factors binding sites for clusters of coexpressed genes. These results can be used as input to future studies aimed at inferring regulatory networks, using all available genetic, genomic, proteomic, and pathway information. The most straightforward application of such trans-eQTL would be in the module network algorithm of Segal et al. (2003), which so far has not explored this trans-eQTL information to infer regulatory networks. Alternatively, known regulatory genes close to putative eQTL can be used together with groups of coexpressed genes to discover novel transcription factor binding sites. The above strategies are a part of the systems genetics approach and can be seen as a good tool for selecting regions in the genome where interesting candidate (regulator) genes can be found. These candidates have to be further characterized in more detailed analyses and in confirmatory experiments. Trans-eQTL identified as a major regulator of gene expression would be important, for example, in disease risk or susceptibility in that if such trans-eQTL are segregating, they can be selected for or against by conventional animal breeding strategies. The module network algorithm analyses and linkage cluster analyses can reveal major candidate regulator genes that control a cluster of genes; such major genes could be used in gene- or marker-assisted selection (GAS/MAS). Once we understand pathway, genome sequence, and expression background, such data have to be transformed into information for selection strategies.

Discriminating multiple QTL and eQTL. The GG analysis would also pinpoint multiple eQTL regions that affect the same (expression) trait. These trans-eQTL mapped by GG may help us validate or add a completely new set of QTL to the trait under investigation (e.g., Fig. 1 showing multiple eQTL at Chr 1, 5, and 12; all or some of these positions may have shown up as multiple QTL regions by conventional genome scan for end phenotype controlled by the *Ins1* gene).

Multitrait QTL mapping and detection of pleiotropic eQTL. Genetic analysis of genome-wide expression analysis leading to identification of cis-eQTL/SNPs or trans-eQTL/SNPs would help in unraveling the genomic basis of genetic correlations between different phenotypes [e.g., for animals, body fat reserves versus fertility or health problems, weigh gain or growth versus meat quality and diseases, as reported by Kadarmideen (2004) and Kadarmideen et al. (2004)]. It helps in understanding the molecular basis of population-based pleiotropic effects that are favorable or unfavorable. To be able to investigate the molecular basis of pleiotropy, microarray experiments can be set up to address two or more different phenotypic traits. The claim made above is illustrated in Fig. 5, where we chose three most likely related traits, insulin 1 (*Ins 1*), insulin 2 (*Ins 2*), and glucose transporter gene (*slc2g5*) in type I and type II diabetes and obesity in humans or mice and conducted a multitrait eQTL mapping. Results (Fig. 5) show that Chr 1, 5, and 12 contain pleiotropic eQTL affecting all three traits.

Potential uses of eQTL in animal breeding

Prediction of eQTL effects. Microarray analysis offers a new way to study genotype-''end'' phenotype relationships by identifying gene and metabolic networks through ''intermediate'' phenotype (e.g., expression data, protein-interaction data) and SNP markers within genes, and it lends itself to *Expression-Assisted Selection* (*EAS*). Although establishing such a link is often difficult, the prospects of translating ''intermediate'' phenotype to its actual effect on ''end'' phenotype at the animal level exist, as shown for mouse data by Schadt et al. (2005). In this section we discuss the estimation of heritability on gene expression patterns observed on a microarray. At the outset, it must be noted that mRNA levels themselves are not transmitted from parent to progeny but can be correlated to end phenotypes of interest that are inheritable. Thus, mRNA levels can serve as intermediate traits. The major constraint of using expression data is that mRNA levels are highly variable across time and cellular levels and thus repeatability of data is very poor. Therefore, several replicates and large sample sizes would be needed. The estimation of heritability may be impractical in many situations but could be feasible, for instance, in the improvement of meat quality where the necessary material for gene expression measurements can be obtained from a muscle biopsy. To estimate heritability of gene expression phenotype at gene (spot) *j* on the microarray, we need observations on the same gene *j* from different individuals for example in the hundreds or even thousands (1 array per individual). To have such a data set, we need a large number of individually microarrayed animals. This may become possible as costs of microarraying go down or different strategies are used (described below). Assuming that all the ''low-level'' microarray data processing has been accomplished, fitting a statistical model to estimate additive genetic var-

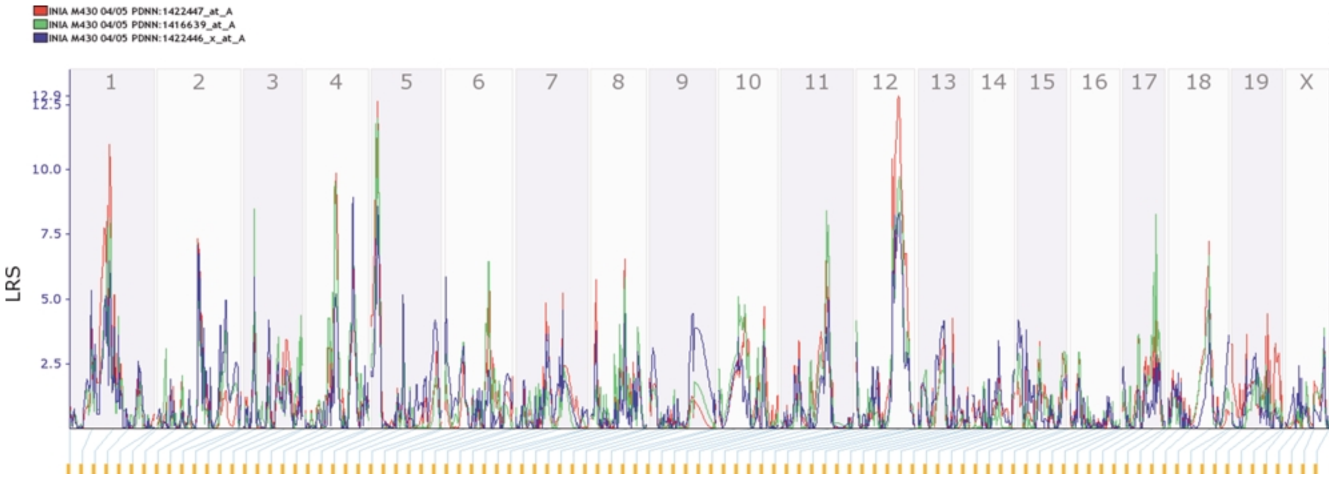

Fig. 5. Multiple (pleiotropic) mapping of eQTL that influence gene expression traits insulin 1, insulin 2, and glucose transporter gene (*slc2g5*). Example trans-eQTL on Chr 12 is significant and affects all three transcripts.

iance of gene expression data would ideally result in estimating h^2 for each spot or gene j. For G number of genes on a microarray, we would have G number of h^2 estimates, from $h_1^2, h_2^2, h_3^2, \ldots, h_G^2$ and predicting Estimated Breeding Values (EBVs) for each animal i for each gene j such that EBV_{ij} indicates $i = 1,2,\ldots N$ and $j = 1,2,\ldots,G$, with a total of $N \times G$ EBVs (i.e., 1000 animals with 30,000 genes on a array would lead to 30,000,000 EBVs and 30,000 h^2 estimates). The method of estimating heritability could be based on the methods used by, for example, Monks et al. (2004) and can be extended to include typical BLUP mixed models and commonly used variance component methods such as REML or Bayesian-MCMC methods. The main question, in the context of an economically important trait, is what exactly is the phenotype and how is that phenotype related to quantitative traits such as milk, meat yield or quality, or health status. If we consider gene expression as an intermediate phenotype, it is implicit that this phenotype relates very specifically to the condition or treatment in which mRNA samples were collected from, say, diseased versus healthy kidney or brain samples. Then, the heritability or breeding value for gene expression relates to this particular condition, which in turn results in observable end phenotypes at the animal level by showing kidney-oriented diseases or neurologic problems. The dependency of gene expression patterns (and their heritabilities) on particular environmental conditions essentially is not different from other phenotypes, which also depend on environmental conditions, but these intermediate phenotypes may be more strongly affected by environmental conditions than aggregated ''end'' phenotypes. The breeding values obtained for gene expression of animals could then be used in direct EAS. Immediate concerns are the dimensionality and volume of EBVs available per animal and h^2 estimates. We propose one obvious solution: Instead of estimating h^2 for all G genes, one could restrict the analysis to those that are significantly differentially expressed (say 30% of all G genes) by applying the t-test. The two-sample t-test can detect gene expression differences between two conditions, e.g., healthy versus diseased animals, by calculating a gene-specific significance for each gene j. Because of multiple testing with tens of thousands of t-tests, permutation techniques are used to derive significance threshold values from null hypotheses of no differential gene expression. Then the gene j is said to be differentially expressed at a given FDR of, say, 5%. These genes are then ranked based on t value to select the high versus the low differentially expressed genes. Many other proposals have also been made for dimensionality reductions in GG methods such that it should be possible to keep the genomic-transcriptomic evaluations to a manageable size.

The use of EBV for gene expression will depend on the goals of the breeding program. It is obvious that the EAS has a long way to go before it is used in animal breeding. However, there are several other uses for such EBVs. EAS can help improve genetics for economically important traits such as milk and meat yield by selecting favorable gene expression profiles of certain candidate loci. As opposed to direct selection of gene expression patterns, we may be interested mainly in eQTL that affect gene expression and incorporate such eQTL into a breeding program much like GAS/MAS programs that try to incorporate normal QTL or markers that bracket a normal QTL. Hence, no new principle needs to be introduced here in addition to what is already available in the literature for incorporating QTL information in MAS and BLUP evaluations (e.g., Fernando and Grossman 1989; Meuwissen et al. 2001). The main difference between normal MAS versus eQTL-based MAS is that MAS is a very coarse or broad genomic-based selection, whereas eQTL-based MAS is a very specific transcriptomic-based selection, specific to a particular treatment of tissue and time from which the mRNA population originated. It is clear that the animal breeding community is still struggling to include QTL information in the BLUP model mainly because of the costs involved in genotyping all individuals in the population for dense marker maps and computational difficulties in deriving and inverting the identity-by-descent variance-covariance matrix for a large pedigree, with and without marker-pedigree data. Assuming that these issues are solved in the near future, it is easy to include eQTL in BLUP animal models because it is still a QTL that affects an intermediate phenotype (expression of some genes), similar to normal QTL that affect an end phenotype. Hence *Expression-Assisted Evaluation* (*EAE*) is still possible. Recently, there have been a few attempts (Bing et al. 2005; Tempelman 2005) to extend ANOVA-based methods (e.g., Kerr 2003; Kerr et al. 2000) to BLUP mixed models commonly used in animal breeding. Such gene expression prediction models may have to be integrated into models of Meuwissen et al. (2001) to exploit the full potential of all available genetical genomics tools.

Expected benefits for genetic improvement. In animal breeding, the rate of genetic gain, ΔG, that can be obtained by selection of high-genetic-merit animals and breeding is given by

$$\Delta G = \frac{i.r_{\mathrm{EBV}}\sigma_a}{L_{\mathrm{m}} + L_{\mathrm{f}}}$$

where i is the selection intensity, r_{EBV} is accuracy of EBVs, σ_a is a genetic standard deviation, and L_{m} and L_{f} are generation intervals in male and female animal selection paths. Without showing any proof, it is intuitive that EAS would affect all components of this genetic gain, normally achieved by using phenotypic pedigree data and sometimes with markers (MAS). For example, eQTL data allow decisions to be made using gene expression profiles at juvenile stages as a predictor of adulthood performance (e.g. before sexual maturity or before observations on carcass quality are made) which would shorten the generation interval and thus improve genetic gain. Another example of reducing generation interval is our ability to predict the development of late-onset diseases and disorders (e.g., prion diseases and impaired metabolism) using gene expression *signatures*. The intensity of selection, *i*, is a standardized selection differential (the difference between the mean of selected animals versus the mean of the whole population) and depends on the proportion of animals selected as parents and distribution of phenotype. With eQTL, it would be possible to further distinguish animals based on similar EBVs (e.g., full-sibs or close relatives); thus reducing the proportion selected (high intense selection). For some traits it would also mean differentiating animals genetically much better using gene expression than one would do with the conventional phenotypic or EBV or marker data; this should result in high accuracy (i.e., r_{EBV}) which in turn would contribute to genetic progress.

As for MAS, the full potential of EAS for animal breeding may be realized only under certain circumstances, e.g., where the heritability of a polygenic trait is low (in which case there may be only a few QTL, each with small additive effects) or where the value of information in individual QTL expression would be higher than when the heritability of a polygenic trait is high. Traits with very low heritability in livestock include health or disease resistance, fertility, and longevity (e.g., Kadarmideen 2004; Kadarmideen et al. 2001, 2004). In case of high heritability (at QTL), it can be hypothesized that not only the additive value of a QTL variant but also the additive value of its own expression (or that of trans-eQTL) is inherited by progeny. For traits that are difficult to measure on the farm (e.g., disease susceptibility and health traits, reproductive traits, carcass quality, feed intake), eQTL or SNPs within the eQTL regions can be used as predictors. Such fast-track decisions have clear commercial benefits to animal breeders and producers. Sex-limited traits (e.g., milk production, reproduction) can also be predicted in animals using eQTL/SNP information.

Acknowledgments

The authors thank the WebQTL providers Drs. Jintao Wang, Elissa Chesler, Robert Williams, and Kenneth Manly for allowing us to access the BXD databases and software for this study.

References

1. Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170, 533–542
2. Bing N, Hoeschele I, Ye KY, Eilertsen KJ (2005) Finite mixture model analysis of microarray expression data on samples of uncertain biological type with applica-

tion to reproductive efficiency. Vet Immunol Immunopathol 105, 187–196

3. Brazhnik P, de la Fuente A, Mendes P (2002) Gene networks: how to put the function in genomics. Trends Biotechnol 20, 467–472
4. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752–755
5. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, et al. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using ''genetical genomics.'' Nat Genet 37, 225–232
6. Carlborg O, Andersson L (2002) Use of randomization testing to detect multiple epistatic QTL. Genet Res 79, 175–184
7. Carlborg O, Hocking PM, Burt DW, Haley CS (2004) Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. Genet Res 83, 197–209
8. Carlborg O, De Koning DJ, Manly KF, Chesler E, Williams RW, et al. (2005) Methodological aspects of the genetic dissection of gene expression. Bioinformatics 21, 2383–2393
9. Cassman M (2005) Barriers to progress in systems biology. Nature 438, 1079
10. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, et al. (2001) Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 12, 323–337
11. Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. Nat Genet 32(Suppl), 522–525
12. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971
13. Correa CR, Cheung VG (2004) Genetic variation in radiation-induced expression phenotypes. Am J Hum Genet 75, 885–890
14. Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J Anim Sci 82, E313–328E
15. de la Fuente A, Brazhnik P, Mendes P (2002) Linking the genes: inferring quantitative gene networks from microarray data. Trends Genet 18, 395–398
16. DiPetrillo K, Wang X, Stylianou IM, Paigen B (2005) Bioinformatics toolbox for narrowing rodent quantitative trait loci. Trends Genet 21, 683–692
17. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, et al. (2002) Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of Type 1 diabetes. Genome Res 12, 232–243
18. Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. Genet Sel Evol 21, 467–477
19. Fernando RL, Nettleton D, Southey BR, Dekkers JCM, Rothschild MF, et al. (2004) Controlling the proportion of false positives in multiple dependent tests. Genetics 166, 611–619
20. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7, 601–620
21. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241–4257
22. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69, 315–324
23. Ilahi H, Kadarmideen HN (2004) Bayesian segregation analysis of milk flow in Swiss dairy cattle using Gibbs sampling. Genet Sel Evol 36, 563–576
24. Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4, 145–151
25. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17, 388–391
26. Kadarmideen HN (2004) Genetic correlations among body condition score, somatic cell count, production, reproduction and conformation traits in Swiss Holsteins. Anim Sci 79, 191–201
27. Kadarmideen HN, Janss LLG (2005) Evidence of a major gene from Bayesian segregation analyses of liability to osteochondral diseases in pigs. Genetics 171, 1195–1206
28. Kadarmideen HN, Janss LLG, Dekkers JCM (2000) Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. Genet Res 76, 305–317
29. Kadarmideen HN, Rekaya R, Gianola D (2001) Genetic parameters for clinical mastitis in Holstein–Friesians: a Bayesian analysis. Anim Sci 73, 229–240
30. Kadarmideen HN, Schwörer D, Ilahi H, Malek M, Hofer A (2004) Genetics of osteochondral disease and its relationship with meat quality and quantity, growth and feed conversion traits in pigs. J Anim Sci 82, 3118–3127
31. Kadarmideen HN, Li Y, Janss LLG (2006) Gene–environment interactions in complex diseases: genetic models and methods for QTL mapping. Genet Res [in press]
32. Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, et al. (2000) Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat Immunol 1, 221–226
33. Kerr MK (2003) Design considerations for efficient and effective microarray studies. Biometrics 59, 822–828
34. Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7, 819–837
35. Kitano H (2002) Systems biology: a brief overview. Science 295, 1662–1664
36. Kraft P, Schadt E, Aten J, Horvath S (2003) A family-based test for correlation between gene expression and trait values. Am J Hum Genet 72, 1323–1330
37. Lan H, Stoehr JP, Nadler ST, Schueler KL, Yandell BS, et al. (2003) Dimension reduction for mapping mRNA

abundance as quantitative traits. Genetics 164, 1607–1614

38. Lee NH (2005) Genomic approaches for reconstructing gene networks. Pharmacogenomics 6, 245–258
39. Li HQ, Lu L, Manly KF, Chesler EJ, Bao L, et al. (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14, 1119–1125
40. Liu H, Cheng HH, Tirunagaru V, Sofer L, Burnside J (2001) A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. Anim Genet 32, 351–359
41. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829
42. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, et al. (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75, 1094–1105
43. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, et al. (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci USA 100, 605–610
44. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743–747
45. Moser RJ, Reverter A, Kerr CA, Beh KJ, Lehnert SA (2004) A mixed-model approach for the analysis of cDNA microarray gene expression data from extreme-performing pigs after infection with *Actinobacillus pleuropneumoniae*. J Anim Sci 82, 1261–1271
46. Palmer AA, Verbitsky M, Suresh R, Kamens HM, Reed CL, et al. (2005) Gene expression differences in mice divergently selected for methamphetamine sensitivity. Mamm Genome 16, 291–305
47. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (eds.) (2003) *The Analysis of Gene Expression Data, Methods and Software* (New York: Springer)
48. Pomp D, Allan MF, Wesolowski SR (2004) Quantitative genomics: Exploring the genetic architecture of complex trait predisposition. J Anim Sci 82, E300–E312
49. Qin LX, Kerr KF (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res 32, 5471–5479
50. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302
51. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37, 710–717
52. Sebastiani P, Gussoni E, Kohane IS, Ramoni MF (2003) Statistical challenges in functional genomics. Stat Sci 18, 33–60
53. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34, 166–176
54. Soinov LA, Krestyaninova MA, Brazma A (2003) Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol 4, R6
55. Tempelman RJ (2005) Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol 105, 175–186
56. Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. Genetics 143, 1013–1020
57. Walsh B, Henderson DA (2004) Microarrays and beyond: What potential do current and future genomics tools have for breeders? J Anim Sci 82, E292–E299
58. Wang J, Williams RW, Manly KF (2003) WebQTL: Web-based complex trait analysis. Neuroinformatics 1, 299–308
59. Yaguchi H, Togawa K, Moritani M, Itakura M (2005) Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL. Genomics 85, 591–599
60. Yamashita S, Wakazono K, Nomoto T, Tsujino Y, Kuramoto T, et al. (2005) Expression quantitative trait loci analysis of 13 genes in the rat prostate. Genetics 171, 1231–1238
61. Yang YH, Speed TP (2002) Design issues for cDNA microarray experiments. Nat Rev Genet 3, 579–558
62. Yeung MKS, Tegner J, Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A 99, 6163–6168