The integration of 'omic' disciplines and systems biology in cattle breeding

Berry, D P; Meade, K G; Mullen, M P; Butler, S; Diskin, M G; Morris, D; Creevey, C J

Published in:
Animal

DOI:
10.1017/S1751731110002120

Publication date:
2011

Citation for published version (APA):
Berry, D. P., Meade, K. G., Mullen, M. P., Butler, S., Diskin, M. G., Morris, D., & Creevey, C. J. (2011). The integration of 'omic' disciplines and systems biology in cattle breeding. *Animal*, *5*(4), 493-505. https://doi.org/10.1017/S1751731110002120

Animal, page 1 of 13 © The Animal Consortium 2010
doi:10.1017/S1751731110002120

animal

The integration of 'omic' disciplines and systems biology in cattle breeding

D. P. Berry[1†], K. G. Meade[2], M. P. Mullen[3], S. Butler[1], M. G. Diskin[3], D. Morris[3] and C. J. Creevey[2]

[1]*Animal and Bioscience Research Department, Teagasc, Moorepark, Co. Cork, Ireland;* [2]*Animal and Bioscience Research Department, Teagasc, Grange, Co. Meath, Ireland;* [3]*Animal and Bioscience Research Department, Teagasc, Athenry, Co. Galway, Ireland*

(Received 27 November 2009; Accepted 28 July 2010)

Enormous progress has been made in the selection of animals, including cattle, for specific traits using traditional quantitative genetics approaches. Nevertheless, considerable variation in phenotypes remains unexplained, and therefore represents potential additional gain for animal production. In addition, the paradigm shift in new disciplines now being applied to animal breeding represents a powerful opportunity to prise open the 'black box' underlying the response to selection and fully understand the genetic architecture controlling the traits of interest. A move away from traditional approaches of animal breeding toward systems approaches using integrative analysis of data from the 'omic' disciplines represents a multitude of exciting opportunities for animal breeding going forward as well as providing alternatives for overcoming some of the limitations of traditional approaches such as the expressed phenotype being an imperfect predictor of the individual's true genetic merit, or the phenotype being only expressed in one gender or late in the lifetime of an animal. This review aims to discuss these opportunities from the perspective of their potential application and contribution to cattle breeding. Harnessing the potential of this paradigm shift also poses some new challenges for animal scientists – and they will also be discussed.

Keywords: 'omic', systems biology, breeding, application

Implications

Traditional methods of animal breeding, based on sophisticated statistical methodology, have resulted in considerable genetic gains in traits that were selected on. However, tools of systems biology such as genomics, transcriptomics, proteomics, metabolomics and bioinformatics can further increase the genetic gain achievable over and above traditional methods. This review provides an overview of the current state-of-the-art as well as providing ideas on how new 'omic' disciplines, including systems biology, may be used in cattle breeding. Challenges and future research are also discussed.

Introduction

The observed performance, or phenotype of an individual, is the outcome of the interacting development between the genotype of individual and its specific environment throughout life since fertilisation (Bowman, 1974). For many decades quantitative geneticists, through the development and refinement of increasingly sophisticated statistical methodology, have tried to disentangle from the phenotype, the additive genetic, non-additive genetic and various environmental components and their interactions. The end goal was to accurately predict the additive genetic merit of an animal.

More recently, the paradigm shift in technology has facilitated a move toward a systems biology approach and its component disciplines (Figure 1), to prise open the 'black box' approach adopted by quantitative geneticists where it is assumed that each quantitative trait is a function of an infinite number of genes each with an infinitesimally small effect (i.e. Fisher's infinitesimal genetic model; Bulmer, 1980). This paradigm shift involves the study of actual genomic regions and their effects or associations with performance, in contrast to current methods that rely on the statistical analysis of large quantities of phenotypic data. Systems biology is an inter-disciplinary study of genomics, transcriptomics, proteomics, metabolomics and bioinformatics (van Ommen and Stierum, 2002). However, often omitted from this description is arguably the most important discipline, the definition of the phenotype (i.e. phenomics). A clear definition of the phenotype under investigation, and how it mimics the actual phenotype of interest is vital to the success of the whole process.

† E-mail: donagh.berry@teagasc.ie

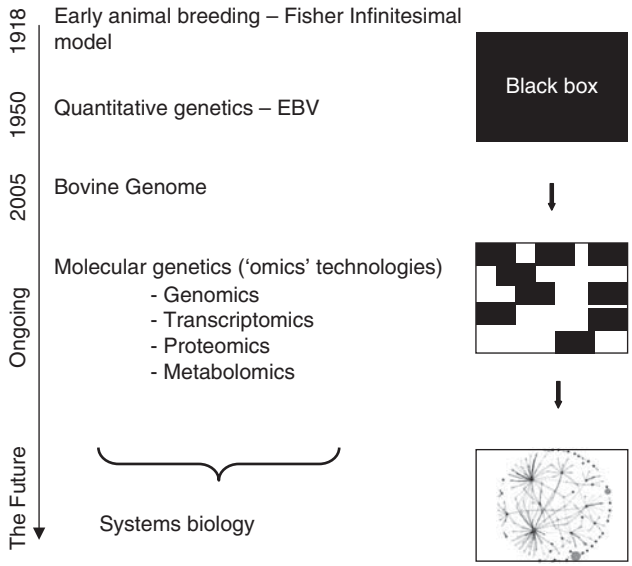

Figure 1 The paradigm shift in approach to animal breeding.

The objective of this review is to summarise the current state-of-the-art in quantitative genetics and in selected molecular science disciplines including systems biology, and explore how they can be used to accelerate genetic gain. The review concludes with challenges to animal breeding as well as identifying topics for future research.

Animal breeding using quantitative genetics

Quantitative genetics is the study of mainly, although not always, traits expressed on a continuous scale. Animal breeders were traditionally interested in estimating the additive genetic merit of an animal, commonly referred to as the animal's true breeding value. However, the true breeding value is never known using traditional quantitative methods and therefore one of the main objectives of quantitative geneticists is to predict, as accurately as possible, the true breeding value of the animal; the predicted value is termed the estimated breeding value (EBV).

Heritability

A common statistic used in quantitative genetics is the heritability that is defined as the ratio of genetic variance to phenotypic variance after excluding the variance attributable to systematic environmental effects. Usually the narrow sense heritability (h^2) statistic is of most interest where the numerator is the additive genetic variance (i.e. variance of true breeding values). Heritability is useful in quantifying the expected response to selection as well as being a useful indicator of the potential efficiency of gene-mapping experiments that use pedigree information (Visscher *et al.*, 2008).

Limitations of quantitative genetics

Genetic gain achieved using traditional quantitative genetic methodology has been immense in recent decades. However, traditional methods are not without their limitations.

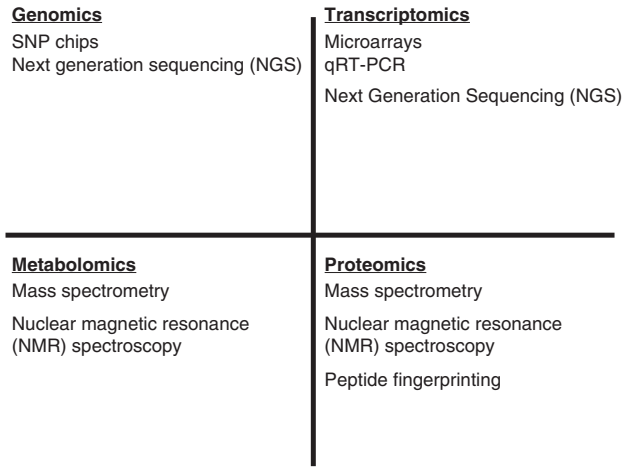

Figure 2 Core technologies enabling developments in animal breeding.

Examples of the weaknesses of quantitative genetics and the infinitesimal model are (i) the phenotype measured contains error (i.e. low heritability trait), (ii) the phenotype may not be measurable in both genders (e.g. milk yield in dairy cattle), (iii) adult performance cannot be measured in juveniles although it can be predicted, and some traits like longevity require a long time horizon to measure, (iv) the animal may need to be sacrificed to obtain the phenotype, (v) antagonistic genetic correlations between traits of interest cannot be easily resolved and (vi) genotype by environment interactions may exist, which may complicate the statistical analysis. Furthermore, the estimation of accurate EBVs using quantitative methods requires large and expensive breeding schemes such as progeny testing.

The different levels of systems biology

Systems biology is an inter-disciplinary study of genomics, transcriptomics, proteomics, metabolomics and bioinformatics (van Ommen and Stierum, 2002). Systems biology is concerned with understanding the dynamic outcome of molecular interactions among biomolecules at the pathway, cellular network, cell, tissue and organismal levels. Instead of analysing individual components or aspects of the organism, such as the response of a single cell type to a specific disease, systems biologists focus on all the components and the interactions among them, all as part of one system. The new technologies used in the different disciplines that contribute to a systems biology approach are detailed in Figure 2 and are now described.

Genomics

Genomics is the study of the structure, function and intragenomic interactions within the genome. Genomics emerged with the sequencing of the first complete genome, bacteriophage φX174, by Fredrick Sanger in 1977, spanning approximately 5 kb encompassing just 10 genes (Sanger *et al.*, 1977). Since then genomic sequencing has rapidly progressed and complete genomes are now known for many

viruses, bacteria, fungi, plants and animals. A draft human genome was released in 2001 by the human genome project with the finalised version completed in 2003 spanning approximately 3 Gb (IHGSC 2004). The recent completion of the draft bovine genome, also encompassing approximately 3 Gb (Liu *et al*., 2009), has provided an invaluable resource for genomic studies related to cattle breeding. It enables the localisation and direct examination of any gene or groups of genes for mutations within the cattle genome, using the widespread and conventional techniques of PCR and capillary sequencing. Global sequence analysis of the bovine genome has huge potential to transform our understanding of the genetics underpinning complex traits.

The principle behind searching for mutations within the genome has, until recently, largely remained unchanged since its conception by Sanger in the 1970s, and has been the gold standard for DNA sequencing for the last 30 years (Hutchison, 2007). The recent development of the next generation sequencers (NGS) utilises a novel approach and is a paradigm shift in both sequencing methodology and in the quantity of sequence data generated. NGS has the ability to produce millions of DNA sequence reads in a single run and is rapidly changing the landscape of genetics (Mardis, 2008). A number of different platforms are available for NGS, and although they differ in their engineering configurations and sequencing chemistries, they share a technical paradigm in that sequencing of spatially separated, clonally amplified DNA templates or single DNA molecules is performed in a flow cell in a massively parallel manner (Voelkerding, *et al*., 2009). This massively parallel approach has reduced the base-pair cost of sequencing by several orders of magnitude (Shendure and Ji, 2008). Using this technology complete genome sequencing could in theory, be carried out in a single run, although multiple runs are currently required for sufficient coverage due to the small fragment sizes sequenced (Mardis, 2008). The use of NGS in animal genomics is still at a very early stage, and due to the costs and challenges of handling such large amounts of data, currently chain termination capillary-based sequencing remains the most widespread method of use for mutation discovery. This is likely to change however, as protocols for single-step polymorphism discovery using NGS become more widely available (Van Tassell *et al*., 2008).

Transcriptomics

The recent completion of the draft cattle genome has also provided a rich resource for the application of technologies that allow large-scale high throughput analysis of the gene expression changes underlying various bovine phenotypes and biological responses. Transcriptomics, the study of messenger RNA (mRNA) dynamics in a given sample, facilitates a global understanding of the molecular changes in gene activation (or suppression) levels that controls the synthesis of proteins within the cell, which ultimately affect function and the phenotype of the animal.

At the most basic level, the coordination of biological functions requires activation of gene expression, and probing the transcriptome can yield insights on the genes and pathways involved in the molecular regulation of animal performance. This in turn aids in the understanding of the genetic architecture of complex traits as well as identifying functional candidate genes. During the last decade, DNA microarrays have emerged as the tools of choice for interrogating the transcriptome. A wide range of large-scale gene expression studies in cattle have been published that describe the application of this technology to the understanding of reproduction (Evans *et al*., 2008), lactation (Maningat *et al*., 2009) and immunobiology (Meade *et al*., 2006).

It is generally regarded that there are several limitations to the use of microarray technology, including sensitivity and reproducibility, many of which have been circumvented by the advent of NGS. There are a number of published methods for NGS, including sequencing of selected mRNA regions (Tag-seq) and genomic DNA sequencing. However, the predominant use is the RNA-Seq method which involves the sequencing of the full-length mRNA transcript library in an experimental sample (Mortazavi *et al*., 2008). RNA-seq facilitates deep-sequencing of the transcriptome providing digital defined counts of transcribed sequences in an unbiased manner. The applications of this technology extend beyond that of traditional microarrays to include detection of alternative splicing, epigenetic effects, microRNAs and single nucleotide polymorphism (SNP) discovery (Morozova and Marra, 2008).

The lack of a requirement for previous sequence information with NGS technology has facilitated a deeper and more comprehensive understanding of the complexity of the mammalian transcriptome, including copy number variation, microRNAs and epigenetic effects (Hurd and Nelson, 2009). Recent studies have estimated that copy number variation (CNV) account for almost 18% of the total genetic variation in gene expression, thereby significantly contributing to the variation underpinning complex phenotypes (Stranger *et al*., 2007), and thus accuracy of selection in genetic evaluations. Furthermore, considering that almost all multi-exon genes are alternatively transcribed in humans, and likely to also be the case in cattle, it is imperative that animal scientists use technology that can account for the effects of alternative isoforms on the resulting animal phenotype. The application, however, of this new technology in cattle research is in its infancy.

Proteomics

The proteome refers to the entire complement of proteins expressed by a genome, cell, tissue or organism at any given time. Proteomics is the large-scale study of the structure and function of these proteins and how they interact with each other. The human and bovine genomes are now thought to contain 20 000 to 25 000 genes, however, the proteome is potentially more complex. Whereas genes are comprised of a relatively unvarying linear sequence of nucleotides, the proteins resulting from such genes are much more varied in structure, function and dynamic range (at least in our current understanding of gene regulation). This increased complexity is a result of alternative gene splicing (Matlin *et al*., 2005), post-translational modification (Walsh, 2006) and

protein/protein interactions (Royer, 1999). Many hundreds of post-translational modifications exist including proteolytic cleavage, acylation, glycosylation, methylation, phosphorylation, sulfation, and di-sulfide bond formation (Krishna and Wold, 1998). Indeed up to 2000 genes, or more than 5% of the genome, may be involved in these processes with tens or even many hundreds of enzymes associated with each class of post-translational modification (Walsh, 2006). The proteomic complement of any one cell may therefore be two or more orders of magnitude greater than the genome and may consist of up to a million functionally different proteins. To further add to the complexity, some genes may not be translated at all or to a lesser degree than that indicated by the abundance of transcript. Indeed, several recent investigations have revealed a rather poor correlation between mRNA and protein profiles (Tian *et al*., 2004; Waters *et al*., 2006), suggesting that different control mechanisms are present at the transcriptomic and proteomic level (Rogers *et al*., 2008) and that this complexity can only be unveiled through integrated analyses of both protein and mRNAs (Tian *et al*., 2004; Waters *et al*., 2006).

The measurement of the protein complement of a tissue or cell type is, like genomics, hampered by the wide dynamic range in the copy numbers of each individual protein within a cell and ultimately on the low absolute abundance of a high proportion of many biologically active proteins. Unlike mRNA, however, where low copy numbers can be amplified in order to facilitate detection, no such technologies currently exist for proteomics. The concentration range of cellular and plasma proteins are estimated to span 6 to 10 orders of magnitude, respectively, whereas the best high throughput protein detection technologies incorporating mass spectrometry are limited to 2 to 4 orders of magnitude (Anderson and Hunter, 2006). Furthermore, the variety of methods required to extract, purify, isolate, analyse and identify the many different types of proteins from a complex mixture are significantly more varied and complex than those typically associated with DNA technologies. Like genomics, proteomics is finding a place in animal science (Lippolis and Reinhardt, 2008) and is spawning an ever increasing range of complex separation technologies and protocols as it moves towards a high throughput mode generating vast amounts of data (Graham *et al*., 2005; Falk *et al*., 2007; Yates *et al*., 2009).

Metabolomics

Metabolomics is the large-scale study of the metabolome or small-molecule metabolic profiles. The metabolome is the entire complement of metabolites in an organism, which are the end-product or by-product of gene and protein expression; ultimately it is a snapshot or fingerprint of the cellular processes operating in a cell at a particular point in time. Therefore, metabolic profiling can give an instantaneous description of the physiology of a cell. Metabolomics may also be used to describe the changes that occur in the metabolome due to various physiological or developmental conditions, or as the result of genetic differences. One of the difficulties with metabolomics is that no one analytical technology can encompass the variety of metabolites in existence. Greater than 50 000 metabolites have been identified and catalogued for plants to date, but only 6500 have so far been catalogued for humans (Wishart *et al*., 2007). Targeted analysis of metabolomics requires that the metabolite of interest be identified *a priori* and available in a pure form. Currently, a large number of metabolites cannot be positively identified in samples using existing analytical techniques, and for many metabolites that can be identified, purified standards are not available (Shulaev, 2006). The use of hyphenated mass spectrometry methods including HPLC-MS, GC-MS and CE-MS as well as NMR and microfluidic-based methods (Kraly *et al*., 2009), all of which lend themselves to high throughput analysis, will ensure that the number of identified metabolites will dramatically increase in the future.

Bioinformatics

Bioinformatics research can be loosely described as the application of computational techniques to address biological problems and may be considered as the over-arching discipline within systems biology research. Bioinformatics covers a wide range of disciplines from software development and database design, to the functional annotation of sequenced genomes and unravelling the complex interactions of its constituent parts (Chicurel, 2002). Perhaps the most beneficial aspect of including bioinformatics in a research program and animal breeding, is the ability to handle enormous amounts of information in a high-throughput manner for the purposes of hypothesis testing or knowledge discovery. It is in this context that bioinformatics finds its niche in systems biology (Kitano, 2002).

Biologists are using increasingly larger-scale analyses to tackle biological problems in preference to the more traditional reductionist approaches, and animal breeding is no exception. Genomic, transcriptomic, proteomic and metabolomic experiments all produce vast amounts of data that must be efficiently handled, analysed, and crucially, compared prior to the inclusion in a breeding programme. While whole areas of bioinformatics research are devoted to the efficient analysis and storage of data, the area which is most relevant to systems biology is the ability to compare and contrast data from disparate sources. This is the 'glue' which makes a whole systems biology approach possible and is the approach which is likely to lead to rapid gains in animal breeding in the future.

Defining the parts list

The first step in any bioinformatic analysis of a system is to obtain a 'parts list' of the genes, proteins, metabolic and signalling pathways involved as well as any gene regulatory networks present along with their biochemical interactions (Raes and Bork, 2008). From this list, a higher-level picture can be constructed of the system as a whole (Kitano, 2002).

At the simplest level, we generally have a list of genes, proteins and metabolites and we wish to know more about their functional properties. This information is available from various publicly available online databases or may have been generated from animal experiments to unravel the genetic architecture of a performance trait. One of the main repositories of genetic sequence information is the USA National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/; Baxevanis, 2008). The NCBI contains many databases including genbank, which holds all publicly available nucleotide and amino acid sequences. While this repository is significant in itself, it is the tools that are available to analyse the data, which make the NCBI one of the major resources for computational biologists around the world. The most widely used bioinformatics tool provided by the NCBI is BLAST (Basic local alignment search tool; Altschul *et al.*, 1990). This tool searches for regions of similarity between pairs of sequences. High-sequence similarities suggest possible functional if not evolutionary relationships between the sequences compared. This information is extremely useful when investigating possible functional roles for unknown sequences or as a first step in the identification of homologs in other species.

If sequence similarity is found between multiple overlapping pairs of sequences, this also allows the definition of a higher level of organisation: gene families. Gene families are sets of genes with known homology (shared ancestry), which are usually biochemically similar. The NCBI has a database of precomputed gene families to which novel sequence can be compared. These families are called clusters of orthologous groups (COGs) and are defined by the comparison of protein sequences from completely annotated genomes, representing major phylogenetic lineages. Each COG consists of proteins from at least three distinct lineages and represents an ancient conserved domain (Tatusov *et al.*, 1997; Tatusov *et al.*, 2003). This idea has been further refined by researchers from the European molecular biology laboratory with the eggNOG (evolutionary genealogy of genes: non-supervised orthologous groups) database (http://eggnog.embl.de; Jensen *et al.*, 2008) which defines orthologous groups for multiple taxonomic levels in the eukaryotes including mammalian, vertebrate, and metazoan specific gene families.

Identifying connections in the system

With the knowledge of the gene families involved in a biological system, the next step is to build up a picture of the interactions between them. When a gene family codes for an enzyme that converts a substrate into a product, it can form part of a biochemical pathway where the product of one enzyme becomes the substrate for the next. These pathways can be as simple as two-enzyme interactions or as complicated as the multiple interactions found in metabolic pathways. The Kyoto encyclopedia of genes and genomes (KEGG) (http://www.genome.jp/kegg/; Kanehisa and Goto, 2000) is a collection of manually drawn pathway maps representing the molecular interaction and reaction networks for metabolic and cellular processes, genetic and environmental information processing, human disease and drug development. KEGG provides the user with a visualisation of the components of these biochemical pathways that are present in a system under study, allowing hypotheses to be constructed about the possible reactions occurring and the functional capacity of the system as a whole.

However, when investigating the genetic basis of a complex polygenic trait (as is often the case with traits of economic importance in animal breeding) we may only know a few of the proteins involved. In this case, it is necessary to build up a picture of all possible genes involved from a small set of genes identified from the literature or as a result of genome wide association studies. Cross referencing genomic, experimental and/or literature data from multiple sources will identify possible functional partners to the genes in our original set. As this process is very time-consuming however, there are databases that provide this cross-referenced information pre-computed. One such database is STRING (search tool for the retrieval of interacting genes/proteins; http://string.embl.de/; Jensen *et al.*, 2009) which combines, scores and weights available information from multiple sources on protein–protein associations. This information is then further augmented with predicted interactions and the results of automatic literature-mining searches. The information on which STRING is built is classified into four categories, genomic information (i.e. synteny, conservation, etc.), high throughput experiments (i.e. protein-binding assays), conserved co-expression (from microarray experiments) and previous knowledge (from the literature). The result is pre-calculated functional associations for over 2.5 million proteins from 630 species.

With the knowledge of the biochemical pathways and proteins associated with a trait the next step would be to elucidate the interactions of the relevant proteins and any small molecules at a cellular level. Again, this information is available from different resources, but there are some attempts to collate them into a single resource. The database STITCH (search tool for interactions of chemicals) (http://stitch.embl.de/; Kuhn *et al.*, 2008) is one such tool, combining information on the interaction of chemicals and proteins from multiple sources. Chemicals are linked to other chemicals and proteins by evidence derived from experiments, databases and the literature, providing an overview of the part of the 'interactome' involved in the expression of the trait of interest.

Applications of 'omic' technologies and systems biology

To date most applications of 'omic' technologies in animal breeding have been through genomics, as either marker assisted selection (MAS) or genomic selection. Systems biology, however, through the use of bioinformatic analysis of data originating from different 'omic' disciplines has the potential to future augment the genetic gain achievable.

Marker assisted selection

MAS is the use in a breeding programme of an established linkage between the inheritance a trait of interest and

segregation of specific, measurable, genetic markers which are coupled with the trait of interest. These loci may be the causal mutation termed quantitative trait nucleotides (QTN; Ron and Weller, 2007) or chromosomal regions that may contain one or more genes that influence a multi-factorial trait termed quantitative trait loci (QTL; Mackay, 2001) or the QTL critical region (QTLCR; Sellner *et al.*, 2007).

Before MAS can be fully exploited in a commercial population, the QTN or QTL must be identified, validated and characterised. The main limitations to date in being able to accurately and rapidly identify QTN have been: (i) the use of inappropriate or inadequate mapping population designs (Sellner *et al.*, 2007); (ii) the lack of genotyping platforms to rapidly fine-map QTL regions (Sellner *et al.*, 2007); (iii) the inability to detect and implicate polymorphisms within QTL regions to the functional QTN (Sellner *et al.*, 2007) and (iv) the cost of genotyping large numbers of animals to generate sufficient statistical power. Fortunately, most of these disadvantages are surmountable. The application of MAS is discussed in detail elsewhere (Dekkers, 2004).

Identifying and detection of QTLs. Several approaches exist for QTL detection although they are generally not mutually exclusive. One approach uses associations between genome-wide dense markers and the trait of interest (Georges *et al.*, 1995). This approach is becoming increasingly feasible with the increasing number of SNPs represented on array chips. Statistical analyses of such data are detailed elsewhere (Weller, 2009). Traditionally, linkage analyses within families using daughter or grand-daughter designs were used to identify QTLs for generally routinely recorded traits, but now, predominantly because of a greater density of genetic markers (i.e. SNPs) available at a low cost, genome wide association studies utilising population-wide linkage disequilibrium are increasing in popularity to identify QTLs (Pryce *et al.*, 2010). A second approach to identifying QTLs is a more targeted strategy based on candidate genes. There are two general mechanisms used to identify candidate genes. The first relies on pre-existing knowledge on the function of a gene in the biological process governing the expression of the trait of interest (i.e. functional candidate gene). The second is based on knowledge of the position of a gene from comparative mapping and experiments in other species, or from screening of the genomic regions in previously identified QTLCR. Nonetheless, the ability of candidate gene approaches to detect genetic variants that affect a trait have been largely unsuccessful (Ron and Weller, 2007).

The choice of approach used in detecting QTLs will often depend on the phenotype under investigation. Genome-wide association studies, which generally require more records, are often used for the detection of QTL for routinely recorded traits while the candidate gene approach is used for traits difficult or expensive to measure and is usually performed on research herds with intensive, accurate, phenotypes.

Detection of QTN using genome wide scans or candidate genetic approaches can benefit greatly from having a systems biology approach incorporating data on the transcriptome, the proteome and the metabolome for narrowing down potential candidate regions as well as validating possible discoveries through expression arrays. Nevertheless, this approach assumes that the discovered QTN that is associated with the phenotype under investigation also affects gene, protein and metabolite expression levels. Alternatively, the mode of action of a QTN may be through a mechanism other than altered transcript abundance, such as affecting enzyme activity, binding kinetics, receptor affinity, etc. In addition, it is not always clear which gene transcript to measure, when it should be measured and in what tissue and under what environmental conditions. Furthermore, gene expression may be episodic further complicating the experimental design and resources required. Nevertheless, once a candidate gene is found to have a significant association with a given phenotype then expression studies on the gene in the population can be undertaken to better understand the role that gene plays in the biological process affecting that trait.

Gene expression studies can also be used for the detection of potential QTL. Jansen and Nap (2001) used the term 'genetical genomics' to describe the marrying of genetic mapping methodology with gene expression data by combining a genome-wide study of gene expression with a genome wide scan of loci controlling variation in gene expression. This approach can be used to dissect gene expression differences among animals into genetic and non-genetic components using populations and methods similar to those used for QTL mapping, but instead of actual phenotypes, the dependent variable is expression levels of multiple transcripts. Regions of the genome that affect gene transcription are called expression QTLs (eQTLS; Schadt *et al.*, 2003) and these can be used to elucidate the genetic basis of gene (co)regulation and transcriptional networks. Kadarmideen (2008) describes some statistical approaches for eQTL mapping. One current limitation of genetical–genomics approaches for QTL detection is the cost of undertaking genome wide expression profiles on a large number of animals to gain sufficient statistical power. However, this limitation will diminish as the cost of expression profiling reduces. In addition, some studies have reported strategies for selecting a subset of animals from QTL mapping populations (Jin *et al.*, 2004; Xu *et al.*, 2005).

Limitations of MAS. Although MAS has theoretically great potential such as the ability to increase the accuracy of selection for a given trait at a very young age, it also has, like traditional quantitative approaches, its shortcomings. Disadvantages of MAS include: (i) it is often difficult and resource-intensive to find the causative QTN or QTL in strong population wide linkage disequilibrium with the QTN although advances in genotyping platforms may help alleviate this issue, (ii) the QTL effects are often overestimated although statistical approaches that attempt to fit all genetic markers simultaneously in the statistical model will help reduce this, (iii) the QTL(s) rarely explain all of the genetic variance ('the missing heritability'; Maher, 2008) and therefore polygenic effects still need to be accounted for in the

analysis, (iv) linkages phases between the QTL and functional QTN may break down or be reversed in some families or breeds over time, and (v) epistatic interactions may exist and may be contributing to the estimated QTL effect in the original study and may subsequently breakdown over time or in different populations.

Genomic selection

One of the main shortcomings of MAS is that currently, and for the foreseeable future, identified QTL or QTN are not likely to explain all the genetic variance in quantitative traits (Maher, 2008). An alternative to identifying and selecting on a limited number of QTL is to select on all QTL simultaneously. Genomic selection, as it is currently known, was first described by Meuwissen *et al*. (2001) and has been described as 'the most promising application of molecular genetics in livestock populations since work began almost 20 years ago' (Sellner *et al*., 2007). It is based on the simultaneous selection for many thousands of genetic markers that densely cover the entire genome and is essentially a larger-scale version of MAS made possible by the development of arrays with many thousands of SNPs for fast throughput genotyping. The success of genomic selection is based on exploitation of linkage disequilibrium between the SNPs and the QTN and the association and linkage phase is assumed to persist across the population. Therefore, dense marker coverage is vital to ensure that all QTN are in linkage disequilibrium with an SNP or haplotype. de Roos *et al*. (2008) suggests that arrays with $\sim$ 300 000 SNPs would be required to find markers that are in linkage disequilibrium with QTN across breeds. The process of genomic selection and accuracy of genomic selection is reviewed elsewhere (VanRaden, 2008; Vanraden *et al*., 2009; Calus, 2010).

Further advancements through systems biology

There is a wealth of information available to build up a picture of the genetic mechanisms behind the expression of any phenotypic trait of interest (Figure 1) and the interactions between traits, but recently the most exciting development in bovine genetics has been the publication of the assembly of the *Bos taurus* genome (Liu *et al*., 2009; Zimin *et al*., 2009). This release, covering over 90% of the genome on all 30 chromosomes is made up of over 2.8 billion base pairs. As the vast majority of these positions correspond to non-coding regions, they represent a huge resource of sequence information on regulatory elements and possible SNPs for genomic selection. Furthermore, many resources that have been created to mine information from the other genomes (i.e. the human genome) can be relatively easily adapted to the bovine model, fuelling the speed in which regulatory, immunological and structural information will become available about the genetic makeup of *Bos taurus*. One such tool is InnateDB (Lynn *et al*., 2008), which contains genes, proteins, experimentally verified interactions and signalling pathways involved in the innate immune response of humans and mice to microbial infection. Crucially, this has also been expanded to include information from the entire human and mouse interactomes. While acting as a valuable resource for the identification of bovine orthologs of interest in humans and mouse and their possible interacting partners, the future release of a bovine-specific version of InnateDB will greatly enhance the results of a systems-level analysis of the bovine genome in an animal-breeding context.

Finally, with the advent of high throughput experiments and increasing computational power there is enough data to support another approach to building a bigger picture: simulation-based analyses. Simulation analyses test hypotheses using *in silico* experiments providing predictions that can be tested *in vitro* or *in vivo*. The aim is to create a computational model of the system under study with enough realism to perturb it and measure the effects (Ideker *et al*., 2001; Chang *et al*., 2005). While this should have great use in systems where there is much knowledge already, for systems with many unknowns (like the genetic regulation of phenotypic traits and the interaction between them) far more information than we currently have will be needed in order to exploit possibilities of this approach.

Systems biology offers much hope for the future of animal breeding. By understanding the intricate functional interrelationships of DNA, RNA, proteins and metabolites occurring within an animal, we hope to build up a picture of how variation in economically important traits occurs. These insights will allow the identification of novel candidate genes and biochemical pathways for study, possibly leading to more targeted selection whereby whole gene families or pathways and not individual SNPs are used for selection. Such knowledge may also be used to elucidate the mechanisms behind genetic correlations among phenotypes and thereby aid in resolving genetic antagonisms within breeding programs, one of the main limitations of current methods.

Challenges and future research

Intellectual property rights and funding

Probably the greatest challenge to animal breeding will not be the science itself, but rather how the innovations are handled by the creator(s). Although there will probably be a greater tendency for commercial organisations to avail of trade secrets rather than embarking on the process and expense of patenting, trade secrets will not prevent duplication of research efforts from different research groups and will lead to slower uptake of innovation both by scientists and animal breeders.

Much discussion in some countries is still on-going as to the proprietor of animal genotypes. Is it the body that owns or owned the animal or the person who incurred the expense of genotyping the animal? The answer is unclear. However, Ogden and Weigel (2007) state that because a DNA sequence is not an original work of authorship, it cannot be copyrighted. Not only do arguments over intellectual property stall research on increasing genetic gain, but approaches like the use of contractual licensing of germplasm such as semen by commercial companies can reduce genetic gain at a national level and generate monopolies.

A further challenge again unrelated to the science itself, is the shift in funding, and by default scientists and graduates, from quantitative genetics to molecular genetics and related disciplines. This is not to say that resources should not be expended in systems biology and its component disciplines, but a more balanced approach must be adopted. Without excellent statistical knowledge on the generation of the most appropriate and accurate phenotypes, as well as the most efficient way to exploit additional information into genetic evaluations and breeding schemes, the commercial benefit of investment in 'omic' disciplines and systems biology to increase genetic gain will not be realised.

Computational challenges

While a systems biology approach can combine data from multiple experimental sources to reveal biologically significant trends not obvious from any one source alone, the amount of data involved and how to integrate these data for analysis can be daunting (Hawkins *et al*., 2010). As an example, a single lane from an RNA-seq experiment, which gives a snapshot of the expressed protein-coding genes in a cell, can produce over 15 million short sequence reads which have then to be mapped back to the genome sequence. Tools (Langmead *et al*., 2009b; Trapnell *et al*., 2009) are emerging that are capable of carrying out such mappings in a time frame of hours, but making sense of this data after mapping is the real challenge. Equally, other 'omic' approaches, each providing a one-dimensional view of the genome function, are becoming increasingly high throughput and are producing equivalently large amounts of data. The real challenge to future systems biology research is not in producing the data, but in effectively carrying meaningful comparative analyses between all the sources of data available. Inroads have been made to address this problem, indeed precomputed databases of comparisons like those highlighted earlier are approaches which will become increasingly common place and analyses that take advantage of 'cloud computing' (Langmead *et al*., 2009a) are likely to be extremely useful to researchers in the future. However, it remains clear that for the present, such large-scale comparisons are only possible in research institutions that have the resources to invest heavily in the computing infrastructure and personnel necessary. The advantages of such integrative approaches and the approaches to integrative analyses are detailed elsewhere (Hawkins *et al*., 2010).

The emerging 'omics'

Epigenetics is an emerging frontier of science (Callinan and Feinberg, 2006), especially in livestock species and involves the study of changes in the regulation of gene activity and expression that are not dependent on gene sequence. Epigenetic modification is a dynamic response that can result from exposure to various environmental stimuli from nutritional levels and composition *in utero* to disease (Jirtle and Skinner, 2007). Evidence from studies in other species suggests that the epigenome may be as important as genetic variation to the pathogenesis of infection (Wilson, 2008). A wide range of traits have evidence linking them to epigenetic mechanisms and which have also been shown to be heritable, and thereby have implications for animal breeding programmes. It has become clear that epigenetics and epigenomics – the genomewide distribution of epigenetic changes – will be essential to an accurate understanding of the complete set of factors regulating the phenome.

Many types of epigenetic processes have been identified including methylation, acetylation, phosphorylation and ubiquitylation. These processes regulate gene imprinting – where the expression of a gene depends on its mode of inheritance (i.e. whether the allele is derived from the maternal or paternal line). The insulin-like growth factor 2 gene (IGF2) encodes an essential growth factor which is associated with both beef and milk production traits in cattle (Goodall and Schmutz, 2007; Flisikowski *et al*., 2007; Bagnicka *et al*., 2010). This gene is imprinted in a tissue specific manner (Gebert *et al*., 2006), indicating that epigenetic processes regulate the capture of these traits in any animal. Another well-studied example of an epigenetic process is chromatin modification. Chromatin is the complex of proteins (histones) and DNA that is tightly bundled to fit into the nucleus. The complex can be modified by acetylation, enzymes and some forms of RNA (microRNAs and small interfering RNAs), which alters the chromatin structure to influence gene expressions. In general, tightly folded chromatin tends to be shut down or not expressed, while more open chromatin is functional or expressed.

Relatively small differences in epigenetic patterns can have a large impact on the phenotype, and it is becoming increasingly clear that a single genotype does not result in a single phenotype (Fraga *et al*., 2005). Epigenetic factors may be a significant contributor to the 'missing heritability' issue in animal breeding and therefore understanding epigenomics is key to accurate manipulation of the animal's phenotype. Epigenetics will also have important implications for the application of advances in reproductive technologies, where *in vitro* manipulations have been shown to alter gene expression profiles (Tveden-Nyborga *et al*., 2008) as well as embryo survival (Perecin *et al*., 2009).

The forgotten 'omics'

Genetic gain is known to be a function of selection intensity, accuracy of selection, genetic variation present among selection candidates, and generation interval (Rendel and Robertson, 1950). However, what is missing from this simple definition is how the phenotype under investigation reflects the true trait of interest over and above that reflected in the heritability. The field of study concerned with the characterisation of a phenotype can be called phenomics. Unfortunately, this field is receiving less and less attention despite being one of the major pieces in unravelling the genetic architecture of complex traits. Furthermore, there is a lack of standards on how to define certain traits across studies with the exception of those traits covered by the International Committee on Animal Recording (http://www.icar.org). If the phenotype is not sufficiently characterised then QTL and

QTN will be identified for a trait, which may not be the true trait of interest. An example is obvious in the complex trait of feed efficiency (Berry, 2008). Do we really have a good definition of 'feed efficiency' or are we trying to identify QTLs associated with measurement error?

Many studies are preoccupied with single phenotypes or a selection of phenotypes. However, the phenome reflects the entire phenotypic profile of a given animal across all traits. The availability of such data is limited primarily due to the necessary resources required to obtain such an extensive list of phenotypes. Therefore, research must be undertaken on developing phenomic tools for rapid phenotyping at a low cost. An example is the use of mid-infrared spectroscopy of milk in the estimation of milk fatty acid content (Soyeurt *et al.*, 2006) which would normally require the resource intensive approach of gas chromatography. Considerable resources have been expended on comparative genomics, but there is also a justification for 'comparative phenomics' where phenotypic databases on different model organisms are available and comparable in a single database and the phenotypes of orthologous genes from different organisms can be directly compared. This was the motivation between the development of PhenomicDB (Kahraman *et al.*, 2005). Although standardisation of phenotype descriptions is a major hurdle to the success of such an endeavour, Kahraman *et al.* (2005) use the example that a gene associated with cancer in mammals which also exhibits a proliferation phenotype in lower organisms such as yeasts may warrant further investigation. Fraught with less complications would be the development of freely available phenotypic and genotypic databases within species, although the costs incurred in acquiring such phenotypes and the commercial sensitivities associated with them may hinder such an initiative, in the near future at least.

A second 'omics' often ignored is econ-omics. Few undertake cost-benefit analyses before embarking on studies in 'omics' and systems biology disciplines. This is despite relatively simplistic methodology being available to undertake such analysis, albeit assumptions need to be made (Weller, 1994). Weller (1994) suggested that the long-term profitability of a breeding prorgam is a function of the expected costs and returns of the breeding program as well as an appropriate discount rate and profit time horizon. Weller (1994) went on to describe alternative approaches for evaluating alternative breeding programs. The different approaches to quantify the economic ramifications of exploiting a new technology in a breeding program include (i) fixing the discount rate and profit horizon and cumulating the annual profit until the end of the time horizon, (ii) estimating the cumulative cost and returns for one cycle of selection with a fixed discount rate but an infinite time horizon; (iii) fixing the time horizon and estimating the discount rate necessary based on the expected costs and returns to yield a net profit of zero, and (iv) fixing the discount rate and, given the costs and returns, estimate the number of years required to achieve a net profit of zero. Some studies, nonetheless, have attempted to quantify the economic benefit of new technologies, more recently the benefit of introducing genomic selection into a national breeding program. Schaeffer (2006) compared a selection strategy using genomic selection to a traditional progeny testing scheme similar to that operated in Canada; he reported a twofold increase in genetic gain using genomic selection with the cost of proving a bull reduced by 92%. In that example, Schaeffer (2006) assumed that genomic selection gave an accuracy of prediction of 0.75. Included in the costs was that of generating the reference or training population. Selection of elite cows based on their genotypes (using the costs of the large SNP assay) and genotyping the bull calves were also included in the cost of the breeding program.

The authors are, however, unaware of any study that has quantified the potential long-term benefit in better defining phenotypes and refining the statistical methodology used in current genetic evaluations compared to allocating resources to 'omic' disciplines and systems biology research.

Future areas of research

Explaining the 'missing heritability' (Maher, 2008) will be arguably one of the most challenging areas for animal breeding in the future. Three studies (Gudbjartsson *et al.*, 2008; Lettre *et al.*, 2008; Weedon *et al.*, 2008) attempted to identify genomic variation associated with human height across a total of approximately 63 000 people. A total of 54 loci were identified following validation, but they explained just over 5% of the phenotypic variation despite the narrow sense heritability of human height being 80% (Visscher, 2008). Finding this 'missing heritability' will require stronger collaboration among the disciplines within systems biology as well as with quantitative geneticists. There are many possible reasons for the apparent inability of currently identified genetic markers to explain all the heritability. First, the estimate of the heritability itself may be inaccurate. This could happen if the experimental design used to estimate the heritability do not accurately account for systematic environmental effects. However, heritability estimates of individual height are relatively consistent across species and across morphological traits (Visscher *et al.*, 2008), and even though they may not be as large as 80%, they are, for sure, not as low as 5%. Height is a very objective measure and less influenced by measurement error which could deflate the heritability.

Another possible contributing factor to the 'missing heritability' is that the expression of a phenotype may be due to a few rare alleles with large effects, or indeed many common or rare alleles with very small effect. Addressing the former may require deep sequencing of a large number of individuals for candidate genes, identified through systems biology. This may help in identifying rare mutations. Animals to sequence could be identified from extremes of a distribution of estimated genetic merit using quantitative genetics. Genetic risk estimated on a liability distribution could be used for dichotomous traits and this may be superior to defining controls in case-control experiments as simply those individuals that are not diseased, thereby overcoming the reduced study power from the defined controls being 'near cases'

based on the liability threshold. Quantitative genetics is key to identifying individuals for sequencing. Identifying common alleles with very small effect can be achieved through using larger data sets in genome wide association studies. Considerable data sets are required however to detect loci with small effects. A sample size of 10 000 individuals is necessary to have 29% power to detect a locus that explains 0.2% of the variance of a trait, assuming a type I error rate of 0.0000005 (Purcell *et al.*, 2003). Nevertheless, larger data sets are currently available in cattle and other species such as pigs and poultry. Because of the availability of phenotypic information on often large paternal half-sib families the accuracy of the phenotype based on genetic merit, at the level of the sire, can be considerably higher than that of an individual animal's own phenotype thereby increasing the power of the study although usually limiting the extend of the available phenotypes. These animal models, coupled with comparative genomics and phenomics, as well as other bioinformatics tools, may also be extremely useful in unravelling the genetic architecture of phenotypic differences within human populations (e.g. for disease-related traits).

Genome wide association studies, using SNPchips, are unlikely to detect the causal mutation. However, by using bioinformatics, genes in the vicinity of the detected genetic markers, or overrepresented biological pathways may be detected which could aid in identifying putative candidate genes for further interrogation. Relaxing the stringency of the statistical tests may further help in identifying these pathways as well as utilising data from transcriptomic, proteomic and metabolomic experiments in a systems biology approach.

Copy number variants (CNVs) are stretches of DNA that are either deleted or repeated between individuals. They commonly arise *de novo* although if this is the case they are unlikely to contribute substantially to the 'missing heritability'. CNVs however, cannot be detected with current SNPchip technology although they can be detected with comparative genomic hybridisation. Epistasis, the interaction between genes, is for the most part, directly 'ignored' by quantitative genetics since it (mainly) contributes to the non-additive genetic variance, which is not completely passed on between generations. As most genes do not operate independently, a systems biology approach is needed to decipher how the entire network of genes and regulatory sequences lead to the expression of a phenotype. Although non-additive genetic variance is not explicitly assumed to be included in the numerator of the narrow sense heritability, chromosomal segments rather than genes are inherited and epistatic interactions between linked genes may be inherited together with the additive effects thereby contributing to the 'additive' genetic variance and the heritability.

It is unlikely, however, despite advances in system biology, that even within the next decade, QTNs identified in farm animals will explain more genetic variation in the traits of a breeding goal than the combination of the remaining polygenic effects. Therefore, it is vitally important that research in the definition and modelling of 'novel' traits and methods of collection of the relevant phenotypic data is not ignored as described previously. For example, Berry (2008) concluded that no satisfactory study has been undertaken on the definition of feed efficiency in dairy cattle, yet it is a topic of increasing interest in 'omic' disciplines and systems biology. Furthermore, accurate phenotypes for different health traits are not readily available in most countries.

The historical approach to animal disease has been fundamentally reductionist in nature, where focus has been on a single host immune parameter resulting from a single disease. However, the immune system is not the result of a single mechanism but rather results from the interactions of numerous genes, proteins, mechanisms and the external environment, to produce immune responses to fight disease. As such, the complex dynamic behaviour of the immune system, with its abundance of intricate intra- and inter-cellular interactions, provides an excellent subject for systems biology (Smith and Bolouri, 2005). The challenge for disease biologists in the future will be to understand how multiple simultaneous and interacting stresses disrupts the function of a system. Understanding how disease perturbs the system will be key to ultimately manipulating this response to reduce the burden of disease. Collation and incorporation of data from the different 'omic' disciplines into an analysis of the entire system will be key to achieving this.

Finally, permanent environmental effects (i.e. factors that effect the performance of an animal over its lifetime) usually account for more variation in performance among animals than additive genetic effects (Berry *et al.*, 2003). Currently little is known about the biological process(es) contributing to these permanent environmental effects, and little research has been done to attempt to exploit such effects in animal breeding. One potential mechanism contributing to the permanent environmental effect is epigenetics, which may be reflected in the transcriptome, proteome and/or metabolome. Although permanent environmental effects are not passed on to progeny, they can be useful nonetheless, when added with the additive genetic merit of the animal to estimate the productive potential of the animal and are therefore useful in making on-farm culling decisions.

Conclusions

Systems biology involves the bringing together of 'omic' technologies and bioinformatics to generate a more holistic understanding of the genetic architecture underlying the biological processes influencing the expression of phenotypes. While systems biology in farm animal species has a multitude of potential applications, its prospective benefits for animal breeding is overwhelming, especially for the identification and validation of QTL or QTN associated with economically important traits and how they interact. Traits that are lowly heritable, measurable in only one gender, are difficult, impossible or resource intensive to measure, can only be measured in adult or dead animals and are antagonistically correlated with other traits could benefit most of all from research in 'omic' disciplines and in particular systems biology. Opening the 'black box' underlying the response to selection

for traits of interest to animal breeders will facilitate a move toward elucidation of the hitherto unimaginable multi-dimensional and integrative complexity underlying biological responses. It will take a concerted management and research effort to ensure that its inherent potential is not lost.

Acknowledgements

The contribution of M.P. Mullen was funded by a grant from Science Foundation Ireland (07/SRC/B1156).

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410.

Anderson AL and Hunter CL 2006. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Molecular and Cellular Proteomics 5, 573–588.

Bagnicka E, Siadkowska E, Strzałkowska N, Zelazowska B, Flisikowski K, Krzyżewski J and Zwierzchowski L 2010. Association of polymorphisms in exons 2 and 10 of the insulin-like growth factor 2 (IGF2) gene with milk production traits in Polish Holstein-Friesian cattle. Journal of Dairy Research 77, 37–42.

Baxevanis AD 2008. Searching NCBI databases using Entrez. Current Protocals In Bioinformatics Chapter 1, Unit 1.3.

Berry DP 2008. Improving feed efficiency in cattle with residual feed intake. In Recent advances in animal nutrition (ed. PC Garnsworthy and J Wiseman), pp. 67–99. University of Nottingham Press, Nottingham, UK.

Berry DP, Buckley F, Dillon P, Evans RD, Rath M and Veerkamp RF 2003. Genetic relationships among body condition score, body weight, milk yield and fertility in dairy cows. Journal of Dairy Science 86, 2193–2204.

Bowman JC 1974. An introduction to animal breeding. Edward Arnold Ltd, London, UK.

Bulmer MG 1980. The mathematical theory of quantitative genetics. Clarendon, Oxford.

Callinan PA and Feinberg AP 2006. The emerging science of epigenomics. Human Molecular Genetics 15, R95–R101.

Calus MPL 2010. Genomic breeding value prediction: methods and procedures. Animal 4, 157–164.

Chang WC, Li CW and Chen BS 2005. Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics 6, 44.

Chicurel M 2002. Bioinformatics: bringing it all together. Nature 419, 751–753.

de Roos APW, Hayes BJ, Spelman RJ and Goddard ME 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179, 1503–1512.

Dekkers JCM 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. Journal of Animal Science 82, E313–E328.

Evans AC, Forde N, O'Gorman GM, Zielak AE, Lonergan P and Fair T 2008. Use of microarray technology to profile gene expression patterns important for reproduction in cattle. Reproduction on Domestic Animals 43, 359–367.

Falk R, Ramström M, Ståhl S and Hober S 2007. Approaches for systematic proteome exploration. Biomolecular Engineering 24, 155–168.

Flisikowski K, Adamowicz T, Strabel T, Jankowski T, Switonski M and Zwierzchowski L 2007. An InDel polymorphism in exon 6 of IGF2 associated with the breeding value of Polish Holstein-Friesian bulls. Biochemical Genetics 45, 139–143.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu Y-Z, Plass C and Esteller M 2005. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America 102, 10604–10609.

Gebert C, Wrenzycki C, Herrmann D, Gröger D, Reinhardt R, Hajkova P, Lucas-Hahn A, Carnwath J, Lehrach H and Niemann H 2006. The bovine IGF2 gene is differentially methylated in oocyte and sperm DNA. Genomics 88, 222–229.

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele M, Zhao X, Womack JE and Hoeschele I 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139, 907–920.

Goodall JJ and Schmutz SM 2007. IGF2 gene characterization and association with rib eye area in beef cattle. Animal Genetics 38, 154–161.

Graham DR, Elliott ST and Van Eyk JE 2005. Broad-based proteomic strategies: a practical guide to proteomics and functional screening. Journal of Physiology 563, 1–9.

Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A, Ingason A, Steinthorsdottir V, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Pedersen O, Aben KK, Witjes JA, Swinkels DW, den Heijer M, Franke B, Verbeek ALM, Becker DM, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Gulcher T, Kiemeney LA, Kong A, Thorsteinsdottir U and Stefansson U 2008. Many sequence variants affecting diversity of adult human height. Nature Genetics 40, 609–615.

Hawkins R, Hon GC and Ren B 2010. Next-generation genomics: an integrative approach. Nature Reviews 11, 476–486.

Hurd PJ and Nelson CJ 2009. Advantages of next-generation sequencing versus the microarray in epigenetic research. Briefings in Functional Genomics and Proteomics 8, 174–183.

Hutchison 3rd CA 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Research 35, 6227–6237.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292, 929–934.

IHGSC 2004. Finishing the euchromatic sequence of the human genome-international human genome sequencing consortium. Nature 431, 931–945.

Jansen RC and Nap JP 2001. Genetical genomics: the added value from segregation. Trends in Genetics 17, 388–391.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T and Bork P 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Research 36, D250–D254.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P and von Mering C 2009. STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37, D412–D416.

Jin C, Lan H, Attie AD, Churchill GA, Bulutuglo D and Yandell BS 2004. Selective phenotyping for increased efficiency in genetic mapping studies. Genetics 168, 2285–2293.

Jirtle RL and Skinner MK 2007. Environmental epigenomics and disease susceptibility. Nature Reviews Genetics 8, 253–262.

Kadarmideen HN 2008. Genetical systems biology in livestock: application to gonadotrophin releasing hormone and reproduction. IET Systems Biology 2, 423–441.

Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, Pohlenz H-D and Weiss B 2005. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 21, 418–420.

Kanehisa M and Goto S 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27–30.

Kitano H 2002. Computational systems biology. Nature 420, 206–210.

Kraly JR, Holcomb RE, Guan Q and Henry C 2009. Review: microfluidic applications in metabolomics and metabolomic profiling. Analytica Chimica Acta 653, 23–35.

Krishna RG and Wold F 1998. Post translational modifications. In Proteins: analysis and design (ed. RH Angeletti), pp. 121–206. Academic Press, San Diego, CA.

Kuhn M, von Mering C, Campillos M, Jensen LJ and Bork P 2008. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Research 36, D684–D688.

Langmead B, Schatz MC, Lin J, Pop M and Salzberg SL 2009a. Searching for SNPs with cloud computing. Genome Biology 10, R134.

Langmead B, Trapnell C, Pop M and Salzberg SL 2009b. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IR, Jacobs KB, Lyssenko V, Uda M, The Diabetes Genetics Initiative, FUSION, KORA, The Prostate, Lung Colorectal and Ovarian Cancer Screening Trial, The Nurses' Health Study, SardiNIA Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann H-E, Mohlke KL and Hirschhorn JN 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nature Genetics 40, 584–591.

Lippolis JD and Reinhardt TA 2008. Centennial paper. Proteomics in animal science. Journal of Animal Science 86, 2430–2441.

Liu Y, Qin X, Song XZ, Jiang H, Shen Y, Durbin KJ, Lien S, Kent MP, Sodeland M, Ren Y, Zhang L, Sodergren E, Havlak P, Worley KC, Weinstock GM and Gibbs RA 2009. Bos taurus genome assembly. BMC Genomics 10, 180.

Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N, Lo R, Naseer M, Que J, Yau M, Acab M, Tulpan D, Whiteside MD, Chikatamarla A, Mah B, Munzner T, Hokamp K, Hancock RE and Brinkman FS 2008. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular and Systems Biology 4, 218.

Mackay TF 2001. Quantitative trait loci in Drosophila. Nature Review Genetics 2, 11–20.

Maher B 2008. The case of the missing heritability. Nature 456, 18–21.

Maningat PD, Sen P, Rijnkels M, Sunehag AL, Hadsell DL, Bray M and Haymond MW 2009. Gene expression in the human mammary epithelium during lactation: the milk fat globule transcriptome. Physiology Genomics 37, 12–22.

Mardis ER 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics 24, 133–141.

Matlin AJ, Clark F and Smith CWJ 2005. Understanding alternative splicing: towards a cellular code. Nature Review in Molecular and Cell Biology 6, 386–398.

Meade KG, Gormley E, Park SD, Fitzsimons T, Rosa GJ, Costello E, Keane J, Coussens PM and MacHugh DE 2006. Gene expression profiling of peripheral blood mononuclear cells (PBMC) from Mycobacterium bovis infected cattle after in vitro antigenic stimulation with purified protein derivative of tuberculin (PPD). Veterinary Immunology Immunopathology 113, 73–89.

Meuwissen THE, Hayes BJ and Goddard ME 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Morozova O and Marra MA 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264.

Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 621–628.

Ogden ER and Weigel K 2007. Can you shrinkwrap a cow? Protections available for the intellectual property of the animal breeding industry. Animal Genetics 38, 647–654.

Perecin F, Méo SC, Yamazaki W, Ferreira CR, Merighe GKF, Meirelles FV and Garcia JM 2009. Imprinted gene expression in in vivo- and in vitro-produced bovine embryos and chorio-allantoic membranes. Genetics and Molecular Research 8, 76–85.

Pryce JE, Bolormaa S, Chamberlain AJ, Bowman PJ, Savin K, Goddard ME and Hayes BJ 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93, 3331–3345.

Purcell S, Cherny SS and Sham PC 2003. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19, 149–150.

Raes J and Bork P 2008. Molecular eco-systems biology: towards an understanding of community function. Nature Reviews Microbiology 6, 693–699.

Rendel J and Robertson A 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. Journal of Genetics 50, 1–8.

Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B and Wiley HS 2008. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. Bioinformatics 24, 2894–2900.

Ron M and Weller JI 2007. From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. Animal Genetics 38, 429–439.

Royer C 1999. Protein-protein interactions, outline of the thermodynamic and structural principles governing the ways that proteins interact with other proteins. Previously Published in the Biophysics Textbook Online (BTOL). Retrieved October, 2009, from http://www.biophysics.org/education/croyer.pdf.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM and Smith M 1977. Nucleotide sequence of bacteriophage X174 DNA. Nature 265, 687–695.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB and Friend SH 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302.

Schaeffer LR 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123, 218–223.

Sellner EM, Kim JW, McClure MC, Taylor KH, Schnabel RD and Taylor JF 2007. Board-invited review: applications of genomic information in livestock. Journal of Animal Science 85, 3148–3158.

Shendure J and Ji H 2008. Next-generation DNA sequencing. Nature Biotechnology 26, 1135–1145.

Shulaev V 2006. Metabolomics technology and bioinformatics. Briefings in Bioinformatics 7, 128–139.

Smith KD and Bolouri H 2005. Dissecting innate immune responses with the tools of systems biology. Current Opinion in Immunology 17, 49–54.

Soyeurt H, Dardenne P, Dehareng F, Lognay G, Veselko D, Marlier M, Bertozzi C, Mayeres P and Gengler N 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. Journal of Dairy Science 89, 3690–3695.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME and Dermitzakis ET 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, 848–853.

Tatusov RL, Koonin EV and Lipman DJ 1997. A genomic perspective on protein families. Science 278, 631–637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ and Natale DA 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold RB, Collins SJ, Hanlon WA and Hood LE 2004. Integrated genomic and proteomic analyses of gene expression in mammalian cells. Molecular and Cellular Proteomics 3, 960–969.

Trapnell C, Pachter L and Salzberg SL 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111.

Tveden-Nyborga PY, Alexopoluosa NI, Cooney MA, French AJ, Tecirliogluc RT, Holland MK, Thomsena PD and D'Cruzb NT 2008. Analysis of the expression of putatively imprinted genes in bovine peri-implantation embryos. Theriogenology 70, 1119–1128.

van Ommen B and Stierum R 2002. Nutrigenomics: exploiting systems biology in the nutrition and health arena. Current Opinion in Biotechnology 13, 517–521.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC and Sonstegard TS 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methodology 5, 247–252.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF and Schenkel FS 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

Visscher PM 2008. Sizing up human height variation. Nature Genetics 40, 489–490.

Visscher PM, Hill WG and Wray NR 2008. Heritability in the genomics era-concepts and misconceptions. Nature Reviews Genetics 9, 255–266.

Voelkerding KV, Dames SA and Durtschi JD 2009. Next-generation sequencing: from basic research to diagnostics. Clinical Chemistry 55, 641–658.

Walsh CT 2006. Posttranslational modification of proteins expanding nature's inventory. Robert and Company, CO., USA.

Waters KM, Pounds JG and Thrall BD 2006. Data merging for integrated microarray and proteomic analysis. Briefings in Functional Genomics and Proteomics 5, 261–272.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JRB, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CAN, Morris AD, Ouwehand WH, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI and Frayling TM 2008. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40, 575–583.

Weller JI 1994. Economic aspects of animal breeding. Chapman & Hall, London.

Weller JI 2009. Quantitative trait loci analysis in animals. CABI Publishing, London.

Wilson AG 2008. Epigenetic regulation of gene expression in the inflammatory response and relevance to common diseases. Journal of Periodontology 79, 1514–1519.

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly M-A, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, MacInnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ and Querengesser L 2007. HMDB: the human metabolome database. Nucleic Acids Research 35, D521–D526.

Xu Z, Zou F and Vision TJ 2005. Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. Genetics 170, 401–408.

Yates JR, Ruse CI and Nakorchevsky A 2009. Proteomics by mass spectrometry: approaches, advances, and applications. Annual Review of Biomedical Engineering 11, 49–79.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marcais G, Roberts M, Subramanian P, Yorke JA and Salzberg SL 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10, R42.