Symposium review: Omics in dairy and animal science—Promise, potential, and pitfalls*

J. D. Lippolis,¹† E. J. Powell,^{1,2} T. A. Reinhardt,¹ T. C. Thacker,³ and E. Casas¹

¹Ruminant Diseases and Immunology Research Unit, USDA Agriculture Research Service, National Animal Disease Center, Ames, IA 50010

²Oak Ridge Institute for Science and Education, Oak Ridge Associated Universities, Oak Ridge, TN 37830

³Bacterial Diseases of Livestock Research Unit, National Animal Disease Center, USDA Agricultural Research Service, Ames, IA 50010

ABSTRACT

Sequencing the first genome took 15 yr and \$3 billion to complete. Currently, a genome can be sequenced in a day for a few thousand dollars. Comparing the relative abundance of nearly every mRNA transcript and small RNAs from cells and tissues from different experimental conditions has become so easy that it can take longer to transfer the data between computers than to perform the experiment. Nucleotide sequencing techniques have become so sensitive that the greatest concern is not detecting a gene or transcript but rather, falsely identifying one. Better genome sequencing has led to more complete transcriptomic and proteomic databases and, combined with more sensitive instrumentation and separation techniques, is bringing us closer to detecting complete transcriptomes and proteomes. The promise of these powerful omics techniques is to lead us to new and unexpected connections between molecular processes in the context of animal health. This promise cannot be achieved without hypothesis-driven research that connects omics data with animal health experiments. Any researcher who wishes to invest the time and resources in omics experiments should be aware of the common pitfalls and limitations of these techniques so they can avoid these issues and maximize the use of these research tools. Several important questions must be asked: What is the quality of the databases and how they are annotated? Are the annotations based on experimental results or computational predictions? What assumptions are made by the analysis algorithms, and how will this affect the result? Finally, how can the research community use the vast amount of data being generated by omics experiments in ways to achieve the goals of better animal health and production (which

is the promise of omics technologies)? Until the observations shown in omics data sets are used to achieve the goals of better animal health and production, the potential of omics technology will not be fully realized. **Key words:** omics technology, sequencing, transcriptomics, proteomics, genome-wide association study (GWAS)

INTRODUCTION

Advances in molecular technologies and data processing have led to new and exciting means to ask scientific questions. Many of these advancements have their origin with the Human Genome Project, a 15-yr, \$3 billion international consortium that sequenced 92% of the human genome, discovered more than 21,000 genes, and led to significant improvements in sequencing technologies, software, and data handling abilities (Collins et al., 2003). At the conclusion of the project, the US National Human Genome Research Institute authored a vision for future genomics research that included a list of strategic goals. The goals for application of the newly “completed” genome were to identify gene variants that contribute to good health and resistance to disease and to use this new understanding of genes and their pathways to develop new therapeutic approaches (Collins et al., 2003). Animal research shares the same goals: to identifying gene variants that contribute to good health and to define gene and protein pathways to aid in the development of new therapeutics. Furthermore, animal research goals include efforts to increase the production and quality of food animal products. The benefit of the Human Genome Project on animal health research should not be underestimated. The same technological advances that helped sequence the human genome were later used to sequence genomes of agriculturally important animals (Hood and Rowen, 2013). Much of the functional annotation for animal genes and proteins comes from matching to human orthologs (Fang et al., 2010). The advances in the field of genomics have helped transcriptomics and proteomics to progress. With ever-increasing technical advances

Received June 22, 2018.

Accepted August 6, 2018.

*Presented as part of the Joint ADSA/NMC Symposium: Advances in Mammary Health and Immunology at the ADSA Annual Meeting, Knoxville, Tennessee, June 2018.

†Corresponding author: john.lippolis@ars.usda.gov

and a continually expanding list of omics technologies, it has become easy for researchers to add these methods to their experimental repertoire.

Scientists are taught the “scientific method,” which starts with a hypothesis based on observations. The hypothesis is a supposition that must then be experimentally tested. Based on the results, a conclusion of the validity of the hypothesis is made, new information causes the hypothesis to be refined, and the process repeats. A significant limitation to this approach has been the inability to assay all of the possibilities in the very complex system we call life. Omics provided a potential solution to the issue of complexity by attempting to analyze “everything” (e.g., all genes, all transcripts, and all proteins) that is going on at a single moment, with the promise of identifying the most relevant and important factors and grouping these factors into biological themes. This was part of the promise of omics experiments—but has the promise been fulfilled?

Omics methodologies (genomics, transcriptomics, metabolomics, and proteomics) are frequently used in animal sciences. These experimental approaches have the potential of leading to new basic and applied understandings of biological processes. However, even though progress is being made, omics has not achieved its full potential. For example, the use of genomics analysis (specifically SNP analysis) is being used by dairy producers as an indicator of a bull’s eventual daughter-based evaluation (Wiggans et al., 2017). Currently, omics data are not frequently followed by concomitant new therapeutics and animal production improvement strategies. Yet the promise of omics research is still valid. For the promise to be realized, we must overcome the pitfalls associated with omics technologies. If a researcher, attracted by the promise of the omics analysis, wishes to perform an experiment, they should be aware of the common pitfalls and limitations to maximize the experiment’s scientific potential. One needs a basic understanding of the technology, what results are expected, and what constitutes a good experimental design. Next, the strengths and weaknesses of the analysis and, in particular, the assumptions that are made, are essential for optimal results. It is critical to understand that omics experiments are data-driven analyses that result in observations, and these observations must be followed up by hypothesis-driven biological experiments for the vast omics data to have partially fulfilled the promise.

TYPES AND OUTCOMES OF OMICS EXPERIMENTS

Science has quickly moved from experiments that study a single gene, transcript, or protein to experiments that have the goal of studying all genes, all transcripts,

or all proteins at one time. Whole-genome sequencing (**WGS**) is available from various commercial entities. RNA sequencing (**RNA-Seq**) techniques are robust enough to detect the smallest of transcript numbers. Recently a “complete” human proteome was published, describing the ability to quantify 99.7% of the annotated human proteins (Kusebauch et al., 2016). This review will concentrate on genomics, transcriptomics, and proteomics experiments with an emphasis on identifying the utility, reasonable expectations, and limitation of each approach.

Genomics

The entire genetic material organized in chromosomes is known as the genome. Although the Human Genome Project remains the gold standard of what can be accomplished in terms of genome characterization, technology has made WGS a possibility for research objectives in animals. Whole-genome sequencing is used to identify genomic regions associated with specific health or production phenotypes in agricultural animals. These genomic regions associated with a trait of interest can be referred to as QTL. Specific QTL can be used to predict or select for desired health or production phenotypes. Originally, the search for QTL was done with labor-intensive technology using molecular markers known as microsatellites (Kappes et al., 1997). Although technological advances in sequencing are constantly improving sequence efficiencies and reducing costs, WGS continues to be an expensive consideration in most experimental designs. To combat this, many genomics research projects on production species turn to more concise sequencing methods that focus on smaller regions sprinkled across the genome, rather than sequencing the whole. The availability of multiple genome sequences for each livestock species has created reference genomes that can be analyzed and compared at the nucleotide level. Single nucleotide polymorphisms are variations in a single nucleotide at a specific position. They have been used to detect regions in the genome influencing traits of interest and can be used as molecular markers (Elsik et al., 2009). These SNP are easily automated into panels or “chips” composed of hundreds of thousands of SNP, allowing a highly efficient way to identify the genomic regions of interest (Matukumalli et al., 2009).

The search for QTL using SNP is known as a genome-wide association study (**GWAS**); GWAS experiments have identified genomic regions closely associated with a variety of production characteristics in different agricultural species, such as growth and carcass traits in beef (Snelling et al., 2010; Lindholm-Perry et al., 2011) and milk production in dairy cattle (Cole et al., 2011;

Jiang et al., 2016). In addition, GWAS experiments have identified genomic regions associated with host response to viral infections such as porcine reproductive and respiratory syndrome (PRRS), bovine viral diarrhea virus, and bacterial diseases such as Johne's disease (Neibergs et al., 2010; Boddicker et al., 2012; Casas et al., 2015). Identification of QTL can be useful at the level of precision of genomic regions; however, the identification of SNP associated with specific genes can also yield important information about disease or health processes. For example, 25-hydroxyvitamin D (25OHD) serum levels in beef cattle were shown to be heritable. Several SNP are localized to the *CYP2J2* gene, a member of the cytochrome P450 superfamily. Members of this superfamily are responsible for the enzymatic conversion of vitamin D to 25OHD. This example shows that genomic data can lead to specific mechanistic explanation of observations; in this case, the differences in 25OHD serum levels were due to a genetic difference in an enzyme that converts vitamin D into 25OHD (Casas et al., 2013). Although results can identify a potential region or even a candidate gene associated with a given phenotype, GWAS and similar marker-based selection methods have been faulted for a general lack of validation and difficulty implementing meaningful results for the livestock industries (Dekkers, 2004).

Genomics has revolutionized how we make genetic associations with phenotypes of interest, from production traits, such as milk yield and back fat, to animal health and disease resistance. However, as researchers seek to use these technologies to explain more complicated traits, it is paramount to assess genomics in the perspective of the whole biological system. Currently, the most advanced genomic technology can offer only a small examination of the genetic variations potentially associated with the phenotype of interest.

Transcriptomics

High-throughput sequencing technology has made it possible to study the dynamic expression of the genome, known as transcriptomics. The transcriptome comprises all RNA—that coding for genes and noncoding RNA. Transcriptomics experiments typically examine the expression of mRNA, the transcripts from genes ready to be translated into proteins. Additionally, transcriptomics can capture the expression of small RNA species, small nuclear RNA, small nucleolar RNA, microRNA, and transfer RNA, some of which are thought to have protein translational regulatory functions. The advantage of transcriptomics lies in the ability not only to quantify transcript expression but also to be able to study transcript variations, such as various isoforms

and splice site variants. Although the genome of an individual is constant, transcriptomics allows for the study of the manifestation of the genome in the form of RNA, permitting the comparison of transcript expression from different tissues and treatments, such as time, drugs, nutrients, infections, or production stressors.

Transcriptomics uses 2 technologies: microarrays and RNA-Seq. Microarrays were one of the first methods to investigate the expression of numerous genes in one assay (Gergen et al., 1979). Microarrays involve a hybridization system that uses gene-specific nucleotide probes coated on a plate. Microarrays produce sensitive, reproducible results, and have long been appreciated for their low cost and straightforward data analysis. Although microarrays are robust, they require knowledge of the nucleotide sequence of the transcript of interest. A microarray assay may not be able to distinguish the gene of interest or nontarget sequences if the similarity is too great, or conversely, may not be able to detect an allele if the sequence divergence is too great. Much like PCR, microarray specificity and sensitivity depend on oligonucleotide probe length and probe–target annealing conditions. Because RNA-Seq does not rely on the integrity of premade probes, it can capture novel SNP and does not suffer from some of the hybridization challenges that exist in techniques like microarrays (Zhao et al., 2014). Moreover, RNA-Seq technologies produce data that require more computational expertise to fully process and analyze than microarray data sets. The cost of microarrays compared with RNA-Seq is one of the major advantages of the former approach; however, constant advancements in RNA-Seq have made it more affordable in recent years.

Transcriptome studies using RNA-Seq were first conducted in humans (Pan et al., 2008; Sultan et al., 2008). Transcriptomic approaches have characterized numerous tissues and cell types to explore novel mechanisms of disease (Chaussabel et al., 2010; Zhao et al., 2014). In animal models, transcriptomics analysis has been used in cattle to establish differences in expression of RNA for growth and carcass composition traits (Pegolo et al., 2014; Pena et al., 2014; Costa et al., 2018), for dairy nutrition (Bionaz and Loor, 2012), and mastitis (Loor et al., 2011; Lawless et al., 2014). Reviews published to date focus on the potential of this methodology to improve animal production (Ferreira et al., 2013; Li et al., 2017), and have potential for impact in veterinary research and animal health.

Proteomics

The purpose of a proteomics experiment may be to detect as many proteins as possible (a survey) or to detect a select few (targeted). The most common proteomics

experiment is referred to as “bottom-up proteomics,” meaning a proteome is digested into peptide fragments before injection into a mass spectrometer. The survey identification of many proteins (thousands) is based on the sequencing of the peptide fragments from a data-dependent algorithm or alternatively, identifying small groups of proteins (hundreds) using targeted selected reaction monitoring (SRM).

Proteomics experiments can also be divided into categories of survey and expression, based on the general goal of the experiment. The goal of a survey proteomics experiment is to identify proteins in a sample and to organize them by their gene ontology (Lippolis and Reinhardt, 2010). In addition to identifying proteins in a sample, a survey proteomics experiment identifies each protein's associated peptides. These peptides can then be used to develop an approach to selectively monitor proteins in a complex mixture using SRM. Proteomics experiments can also be designed to determine expression differences in various samples. This can be accomplished by simple counting of peptides that the mass spectrometer detects from a protein. Comparison of protein expression in several samples can be accomplished with various molecular tags that label peptides in a way that can be quantitated by the mass spectrometer (Lippolis and Reinhardt, 2008). The results of an expression proteomics experiment will be lists of proteins whose expression increases, decreases, or does not change due to experimental or natural perturbations. These differences in protein expression can then be grouped to identify biological themes and may be the foundation for hypothesis-based experimentation.

Targeted proteomics experiments such as SRM have been used to quantify as many as 200 proteins per analysis (Vidova and Spacil, 2017). This experimental approach is often the result of a survey proteomics experiment where potential biomarkers are discovered. However, in this type of experiment, the mass spectrometer is of a different type and is set to detect specific peptide and peptide fragment masses. In the presence of radiolabeled synthetic peptides that act as internal standards, numerous peptides are quantified. Recently, a group challenged dairy cows with lipopolysaccharide or peptidoglycan to partially mimic the immune response to gram-negative or gram-positive bacteria, respectively. Milk from these animals was tested over several days for the presence of several immune-responsive proteins using SRM. The goal of their work was to identify a biomarker that could distinguish a gram-negative from a gram-positive infection (Kusebauch et al., 2018). The challenge will be to translate these observations into a quick and reliable test that can be used on farms.

What Makes a Good Omics Experiment?

First, a good omics experiment results in information that leads to one or more testable hypotheses. Often the justification for an omics project is that it will result in new and unexpected observations that will lead to new therapeutics and animal practices. However, the critical intermediate step between the omics experiment and implementation of a new treatment on farm is testing the hypothesis that resulted from observations generated by these powerful but limited methodologies. Finally, we must ensure that our omics experiments are set up and performed in a scientifically sound manner to maximize the probability that the observations will result in sound hypotheses to be tested. In other words, what makes a good omics experiment?

Quality of Sample

The adage “garbage in, garbage out” is certainly true for omics experiments. The quality of the starting material in any omics experiment determines the quality of resulting observations. The quality of the starting material can be determined using several methods. For example, RNA is especially susceptible to degradation. One common metric of RNA quality is the RNA integrity number (RIN), and studies have shown that differences in RIN value can affect gene expression quantitation (Gallego Romero et al., 2014; Reiman et al., 2017). Therefore, it is common practice to test RNA samples with a quality assessment test before the RNA sequencing methods. There is not an accepted quality control assessment for proteomic samples. Because part of the experiment is to digest the protein into peptides, there is less need for a quality assessment. However, sample preparation with protease inhibitors as well as careful and consistent sample handling are critical. If problems with the samples exist, a gross assessment of the sample on a SDS-PAGE gel before protease digestion can help evaluate the diversity and quality of the proteins in the sample.

Any interpretation of the results must be placed in context with the starting material. The conclusions from an omics experiment using a tissue would be different from those using a single cell type, cell organelle, or a biological fluid from various compartments. Researchers who experiment with tissues must acknowledge that 2 sections of the same tissue will not have the same composition of cells. With highly sensitive methodologies and instrumentation, this variation has the capacity to affect results. This can be especially problematic when working in an infection model. For

example, accounting for the changes in the number and types of immune cells entering a tissue would be critical to an omics experiment comparing pre- and postinfection samples.

What is a single cell type? In the case of an immune response, would isolation of T cells be sufficient for an omics experiment to be successful? Would isolation of T helper (Th) cells be better, and if so, would isolation of Th1 cells be better still? Ultimately, would isolation of antigen-specific Th1 cells not be the best (Phetsouphanh et al., 2015)? Of course, the more homologous the cell population analyzed, the easier any interpretation will be. However, there may be too few of those cells to isolate and obtain enough sample for analysis. Therefore, what is possible and what is practical are important considerations to be aligned with the overall goals and questions being asked.

Sample Pooling

One major reason to pool samples is simple—to reduce the cost of the experiment. Second, pooling may be necessary if individual sample size does not yield sufficient product to be detectable in current methods or by current instruments (Schisterman and Vexler, 2008). Some have made the argument that pooling can reduce biological variation due to an outlier, and thus reduce noise in the data set (Zhang and Gant, 2005; Schisterman and Vexler, 2008; Bruning et al., 2015). Others have argued that pooling samples will remove natural variation and generate an artificial “in-between” phenotype (Bruning et al., 2015). However, the consequences of sample pooling are the loss of statistical power and the ability to assess real biological effects and variation. It has been demonstrated that the best way to improve detection of differentially expressed genes is to add more replicates, with biological replicates being more valuable than technical replicates (Rapaport et al., 2013; Lowe et al., 2017). Therefore, pooling samples should be a choice of last resort. Genomic, transcriptomic, and proteomic samples are often combined after each individual sample has been labeled or tagged in some way. Thus, multiple samples are run on an instrument at once, but the labeling allows analysis of individual samples. Because of the sample labeling, this type of combining samples does not affect the biological sample number.

Sample Preparation

Careful and consistent sample preparation is critical to any omics experiment. However, typical genomics and transcriptomics sample preparation methods are

largely based on commercial kits, and once a particular method is chosen, it is a matter of consistency of execution. Proteomic sample preparation is much more varied, and choices about sample preparation will have a profound effect on the outcome of the experiment.

A major difference between proteomics and transcriptomics is the dynamic range of protein compared with mRNA concentrations. Protein abundances are thought to span 7 orders of magnitude, whereas mRNA span 3 or 4 orders of magnitude (Schwanhäusser et al., 2011). The very large protein dynamic range is a problem compounded by the fact that mass spectrometers can perform analysis on molecules with a difference of approximately 4 orders of magnitude. This problem is exacerbated by the fact that most proteomics experiments detect peptides from a proteome digestion, increasing the number and dynamic range of the problem (Zubarev, 2013). The difference in dynamic range of what the mass spectrometer can detect and what typically exists in most proteomes has required the development of separation methods to fractionate the proteome into less complicated peptide samples for mass spectrometry analysis. Therefore, many proteomics experiments consist of multiple runs per sample. Various separation methods exist, each with their advantages and disadvantages. The choice of separation method must be aligned with the goals and expectations of the project (Lippolis and Nally, 2018).

What Is Significant?

It is important to understand how omics data are validated. Genomics, transcriptomics, and proteomics have different requirements to validate experimental results. There are validation methods for identification of nucleotide sequence or protein, and methods for determining significant expression differences for genes or proteins. A basic understanding of various confidence metrics will help a researcher evaluate the quality of the experimental results. These metrics include accounting for sequencing errors, ambiguous database matches, technical variation, and biological variation. Genomic sequencing techniques have an inherent error rate. By increasing sequence depth or redundancy, the inherent error rates in sequencing methods can be corrected. However, repetitive regions in a genome can inhibit complicate sequencing, and high-quality genome sequencing projects now use multiple sequencing approaches to achieve the greatest level of genome coverage (Sims et al., 2014). When the goal of the sequencing project is to identify SNP and small insertions or deletions (indel), one study suggested a 50× average mapped depth to allow SNP and indel detection in 95%

of a human genome (Ajay et al., 2011). Transcriptomic expression data are typically expressed as reads per kilobase per million reads (RPKM) or fragments per kilobase of transcript per million reads (FPKM), to account for the fact that longer genes will have more associated reads than shorter genes. It has been suggested that 80% of transcripts that are expressed with >10 RPKM can be accurately quantified using ~36 million mapped reads coverage in a eukaryotic genome (Sims et al., 2014).

Determination of the sequence of a nucleotide fragment generated in the methodologies required for DNA and RNA sequencing progresses sequentially from one end, nucleotide by nucleotide, to the other. In contrast, peptide identification by mass spectrometry typically uses a 2-step process. The first step is to determine the accurate mass of the peptide, and the second is to fragment the peptide. The observed peptide mass is then matched to peptides from a database of choice, resulting in hundreds of potential matches. The fragmentation profile of the peptide is then compared with theoretical profiles of the potential matches, and the one with the best fit is chosen. Protein scores are therefore a product of the number of peptides matched and their quality scores. For that reason, protein identification scores based on a single peptide, referred to as “one-hit wonders,” are discouraged in many proteomics journals without additional documentation. However, one report suggested that protein identification with 2 mediocre peptide matches is worse than 1 high-scoring peptide match (Gupta and Pevzner, 2009). If one-hit wonders are used, they should be identified and used with care and should not be used in differential expression proteomics. In addition, some nucleotide sequences or peptide sequences can match 2 or more distinct genes or proteins. Using principles of parsimony, processing software attempts to correct for this issue to accurately assign sequences to their proper gene or protein. Spurious results, especially expression data, may be due to parsimony problems and can be assessed with various software tools.

In a study to determine how many biological replicates were needed to show differential expression changes in a transcriptomics experiment, the researchers compared various tools and biological replicates. The conclusions from their findings were that 3 replicates could be sufficient to identify genes with a significant differential expression level of 2.0-fold change or greater. However, due to sample variation in their system they recommended 6 replicates for gene expression experiments, and 12 replicates if all differentially expressed genes are critical (Schurch et al., 2016). No similar study to determine best replicate numbers has been published

for proteomics data. However, significant differential gene or protein expression as determined by a statistical software does not infer biological importance or functional change. The magnitude of expression changes that are biologically significant would be unique to each gene and dependent on the context of the experimental study. The choice of an expression threshold is somewhat arbitrary. A suggested approach would be to choose a conservative threshold (2-fold change), determine the genes or proteins observed to have expression differences, and analyze to genes or proteins or their pathways of interest. This would be followed by a gradual reduction of the fold-change threshold to specifically look for additional genes or proteins that are associated with a researcher's interest. Confirmation of transcriptomic or proteomic observations that are important to the investigator should be independently confirmed by a secondary method such as real-time quantitative PCR, Western blots, or ELISA.

BIOINFORMATICS

Omics experiments generate huge data files that need to be processed into a usable format. On the most basic level, the good and usable information in the data set needs to be separated from artifacts that are produced with these technologies. Sequence information is aligned to nucleotide or protein databases. Various metrics that determine confidence in the gene/protein identification and expression values are calculated. Genes, transcripts, and proteins can then be analyzed for known associations with biological processes, pathways, and networks. All these steps are part of bioinformatics and systems biology and can be accomplished with various software packages. Initial omics processing software was developed using small computer programs, each performing specific data analyzing tasks and accomplished in a specific order. The sequential running of these computer programs is referred to as a pipeline because the output of one step was used as the input for the second step and so forth. These computer programs require programming skills along with precision and understanding of the data modification undergone at each step to manage the variables required to run the scripts successfully and may require collaboration with bioinformaticists. Conversely, bundled software packages do the same tasks as the pipelines and do not require advanced programming skills. Many people prefer the control and speed of these pipelines of sequentially run small programs, rather than software packages that bundle many of these same small scripts into a more user-friendly format. However, although many researchers prefer to have their data processed by

others more experienced with omics data, it would be helpful for every researcher to have a basic understanding of the experimental design, assumptions, thresholds, and limitations inherent in omics data processing.

After omics data have been generated by the instrument, the first step in data processing is quality control. For proteomics data, exclusions can be made to eliminate spectra that are considered too small or do not have a good signal-to-noise ratio. Generally, the next step in the process is to map the nucleotide reads or peptide spectra to a reference database. Mapping peptide spectra to a sequence in a protein database is a prediction based on acceptable tolerances and variation. Some of these tolerances and variation can be due to instrument noise, incomplete processing reactions, or biological modifications such as posttranslational modifications.

Similar to the proteomics data, the first step of nucleotide sequencing data processing is quality control. Linker sequences used in the process of nucleotide sequencing are trimmed from the reads. If a reference sequence is available, the transcriptomics data can be mapped to the genome or transcriptome. Genome mapping requires accounting for gene splicing for eukaryotic samples (Conesa et al., 2016). If a reference database is not available, one can be generated from both genomic and transcriptomic data.

Databases

Key elements of any omics experiment are the databases of known nucleotide and protein sequences. This critical component is largely out of the control of the researcher. When the human genome was considered “complete” in 2003, it reported ~99% coverage of the euchromatin interrupted by 341 gaps (International Human Genome Sequencing Consortium, 2004). In addition, about 90% of genomic structural variants that exist between individuals were found in the areas of the genome that were not sequenced in the initial human genome project, indicating significant differences between individuals and the reference genome (Huddleston et al., 2017). This observation suggests that every individual in an out-bred population of animals could have a genome that varies significantly from a given reference genome. This problem is compounded for the genomes of species sequenced after the human genome. For instance, the first step in sequencing the cattle genome was to anchor it to the previously completed human and mouse genomes (Larkin et al., 2003).

After raw omics data have undergone the quality control steps, the next most common step is to align the nucleotide or peptide sequences to a relevant database. The assumption is that the nucleotide or protein

sequence exists in the database. If the assumption that the nucleotide sequence observed is not in the database used, most software will create a separate file for these unmapped reads. Largely, the unmapped reads go ignored unless the extra work of de novo assembly and matching nucleotide reads to more general nucleotide databases is performed. In contrast, if the peptide observed in the raw data is not in the database, the software will find the closest match, an ortholog, or simply an incorrect peptide match. These incorrect matches are indicated by a lower value from a scoring algorithm. The quality metrics of protein identifications therefore depend on the completeness of databases. For example, an experiment to determine the proteome of bovine neutrophils in 2004 was initially analyzed with a non-species-specific protein database available at that time (Lippolis and Reinhardt, 2005). The results showed a larger number of protein matches to human and mouse than to cattle. However, when the same data set was reanalyzed using an improved protein database from 2007, the number of bovine proteins matched was greater than the combined number of mouse and human proteins, and the average protein score of the bovine proteins increase indicated that the matches were of higher quality (Lippolis and Reinhardt, 2008). The positive side of this is that as database quality increases, so will the resulting omics data sets, and older data can be reanalyzed when better databases are available.

We have compared the gross level of functional annotation of protein databases from agricultural animals to the human database. A UniProt protein database for each species is divided into 2 sections, reviewed and unreviewed or, formally, Swiss-Prot or TrEMBL, respectively. Reviewed proteins have undergone manual curation and annotation, whereas the unreviewed are computer annotated. In Table 1, the number of reviewed and unreviewed proteins are indicated for cattle, chicken, pig, sheep, and human, with the percent of unreviewed included. The number of reviewed proteins for the best of the agricultural species is one-quarter of the number of proteins from the human database. The unreviewed proteins were sub-divided using descriptions that were “uncharacterized” or “open reading frame,” or descriptions that include “-like” or “-containing.” The number of uncharacterized and open reading frame proteins is much higher in species with a higher percentage of unreviewed proteins. These numbers reflect proteins that do not have a computer-generated match to an annotated orthologous protein. Proteins with the words “-like” or “-containing” reflect those proteins that have similarities to other structural motifs or functions (e.g., asparaginase-like protein 1 or asparagine synthetase domain-containing protein 1). These design-

Table 1. Assessing the level of confidence in the functional annotation of proteins from databases of several agricultural animals compared with the human protein database¹

Species	Number reviewed	Number unreviewed (% of total)	Unreviewed			
			“Uncharacterized”	“-like”	“-containing”	“ORF”
Cattle	5,999	17,970 (75)	4,371	905	1,972	228
Chicken	2,288	27,192 (92)	8,344	532	1,806	173
Pig	1,422	39,291 (97)	10,793	439	2,780	360
Sheep	457	22,654 (98)	5,930	195	1,625	272
Human	20,245	51,540 (72)	1,852	6,719	7,536	147

¹The UniProt protein database (<https://www.uniprot.org/uniprot/>) for each species is divided into 2 sections, reviewed and unreviewed (formally Swiss-Prot and TrEMBL, respectively). The number of reviewed and unreviewed proteins are indicated for cattle, chicken, pig, sheep, and human with the percent of unreviewed included. The unreviewed proteins were subdivided using descriptions that were “uncharacterized” or “open reading frame (ORF),” or descriptions that contain the words “-like” or “-containing.”

nations are of higher value that uncharacterized and open reading frame, but still fall short of a computer annotation match. Furthermore, there is greater confidence in manually curated and annotated (reviewed) matches than computer annotation matches. Another way to assess the confidence of the protein databases in UniProt is to look at the type of evidence that supports the existence of the protein. The types of evidence include experimental evidence at the protein level, at the transcript level, inferred from homology, or predicted. Table 2 shows the number of proteins that fall into different “type of supporting evidence” categories for cattle, chicken, pig, sheep, and human. Compared with humans, for which over 75% of their proteins with clear experimental evidence for the existence of the protein, cattle have under 10%.

In addition to databases that contain gene and protein sequences are those that contain 3-dimensional structures; enzyme and biological pathways; organization by families or by functional or structural domains; polymorphisms and mutations; protein–protein interactions; posttranslational modifications; and gene and protein ontology (Chen et al., 2017). All of these databases are intended to organize and make biological sense of the huge volume of data obtained from an

omics experiment. They do this by grouping genes and proteins into biological themes, and then allowing the researcher to find the meaning of those themes in their field of study.

Biological Themes

In addition to simply aligning the raw genomics, transcriptomics, and proteomics to identify genes, transcripts, and proteins, respectively, databases are used to identify their functions and to group those functions into biological themes or other meaningful associations. The format of the early transcriptomics and proteomics publications was to list the genes or proteins that were observed or whose expression changed, and then discuss 3 or 4 of those genes or proteins. The decision of which proteins to discuss might be based on the researcher’s knowledge of how these proteins would be predicted to react in their experimental design. Alternatively, the choice of which protein to discuss was based on a search of the literature to find candidate proteins that fit the experimental design or the researcher’s hypothesis or bias. For example, in an early proteomics paper to determine expression differences between neutrophils from prepartum and peripartum dairy cows, we showed

Table 2. Number of proteins from different species by protein existence¹

Value	Cattle	Chicken	Pig	Sheep	Human
Experimental evidence at protein level	1,852	662	11,446	221	53,910
Experimental evidence at transcript level	7,084	2,707	1,442	357	3,836
Protein inferred from homology	4,666	6,261	7,178	6,028	1,589
Protein predicted	10,362	19,848	20,644	16,504	11,705

¹“Experimental evidence at protein level” indicates that there is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray, or nuclear magnetic resonance (NMR) structure, good quality protein–protein interaction, or detection of the protein by antibodies. “Experimental evidence at transcript level” indicates that the existence of a protein has not been strictly proven but that expression data (such as existence of cDNA, reverse transcription-PCR or Northern blots) indicate the existence of a transcript. “Protein inferred from homology” indicates that the existence of a protein is probable because clear orthologs exist in closely related species. “Protein predicted” is used for entries without evidence at protein, transcript, or homology levels. Category definitions are from the UniProt website (http://www.uniprot.org/help/protein_existence). Numbers were obtained by first finding each species reference proteome and adding in the advanced search type of “Protein Existence [PE]” and entering the appropriate term from the drop-down menu.

a downregulation of histone in periparturient cows (Lippolis et al., 2006a). A literature search for histones and neutrophils revealed a new publication that described a novel neutrophil killing mechanism called neutrophil extracellular traps (**NET**; Brinkmann et al., 2004). Associated with these NET were histones. This early proteomics experiment found just over 40 differentially expressed proteins, and literature searches for all 40 proteins could be accomplished manually. Newer and improved transcriptomics and proteomics techniques now result in thousands of genes or proteins identified, and more automated curation methods were needed and developed. Thus, functional information has been added to genomic, transcriptomic, and proteomic databases, with the intention to group polymorphisms or expression changes into larger functional categories and, from those larger function categories, make interpretations about the larger biological context of the changes observed.

There are many ways to identify and group potentially associated genes and proteins. Figure 1 is an example of a visual means to group genes or proteins by associated function. In this type of analysis, each gene or protein is represented by a node (circle), and other genes or proteins with a known functional association to the primary gene or protein are typically close in proximity and connected with a line. In Figure 1, data from a transcriptomics experiment to determine gene expression differences when multiple *Escherichia coli* strains were tested under conditions comparing swimming or swarming motility types. Color differences indicate relative up- or downregulation comparing gene expression from transcripts isolated from swimming or swarming bacteria. The circle indicates a cluster of associated genes, most with functions involving iron transport or utilization. Addition of iron chelators or FeCl_3 to motility assays showed significant effects, confirming the association between iron transport or utilization genes and motility (Lippolis et al., 2016).

Gene ontology (**GO**) is perhaps the most widely used means to systematically organize and prioritize omics data into biological themes; GO terms are divided into 3 categories: cellular components (e.g., where gene products are active), molecular functions (e.g., molecular activities of gene products), and biological process (e.g., pathways and larger processes made up of the activities of multiple gene products; The Gene Ontology Consortium, 2015; Hosack et al., 2003). The number of GO terms increased from 18,000 in 2004 to more than 40,000 in 2014 (The Gene Ontology Consortium, 2015), and GO terms have a hierarchical organization. For instance, the GO term leukocyte migration (GO:0050900) has parent terms of cell motility (GO:0048870) and locomotion (GO:0040011) as part

of the biological process category. Leukocyte migration can be divided into numerous daughter categories such as dendritic cell migration, lymphocyte migration, myeloid leukocyte migration, and so forth. Each of these terms is further defined to some final and very specific GO term. Utilization of GO terms requires a database linking the gene or protein sequence with numerous GO terms. Each gene or protein can have many associated GO terms that have been manually or predictively associated. However, the annotation of agriculturally important species for GO terms falls behind databases for mice and humans. For example, by searching the www.geneontology.org website, we found 894 *Homo sapiens* genes associated with the GO term “leukocyte migration.” In contrast, 528 genes are associated with leukocyte migration in *Bos taurus*, 466 in *Sus scrofa*, 302 in *Gallus gallus*, and 5 in *Ovis aries*. Of course, this does not mean that there are only 5 genes involved in leukocyte migration in sheep, but rather it reflects the poorer annotation of the sheep database. One common method to enhance biological theme data from agricultural species with poorly annotated functional databases is to match the data to orthologs from better annotated species such as human or mice. One such example is that goat proteomics data were matched to human orthologs to generate conclusions about the data (Hernández-Castellano et al., 2016). However, caution in data interpretation is warranted when orthologs are used because a human protein’s function may not completely be identical to its agricultural animal’s ortholog.

Proteins are typically associated with a single function. However, an increasing number of proteins have multiple functions. These proteins are sometimes referred to as “moonlighting proteins” (Huberts and van der Klei, 2010). Many of these moonlighting proteins are enzymes, which can have functions that involve signal transduction, transcriptional regulation, apoptosis, and motility in addition to their catalytic function. For example, myeloperoxidase is a major neutrophil enzyme that catalyzes the production of hypochlorous acid and other reactive oxidants (Winterbourn et al., 2000). Recently, myeloperoxidase has been shown to be a cell surface ligand for E-selectin, which helps target immune cells to the site of infection (Silvescu and Sackstein, 2014). In humans and in cattle, administration of granulocyte colony-stimulating factor (G-CSF) causes upregulation of cell surface myeloperoxidase on neutrophil and monocytes, suggesting an important role of myeloperoxidase in immune cell localization and response to infections (Silvescu and Sackstein, 2014; Powell et al., 2018).

One approach utilizing GO terms is to determine which terms are enriched within a given sample set, meaning what proportion of the genes or proteins

matched to a specific GO term differs statistically compared with a reference (Manzoni et al., 2018). The goal of identifying enriched GO terms is to identify interesting biological themes resulting from a data set. For example, changes in proteins from milk fractions (exosomes, milk fat globular membranes, and whey) were analyzed for changes in expression before and during mastitis, and these data were analyzed for enriched GO terms. All 3 milk fractions had the general biological processes GO terms of “defense response to bacterium,” whereas only whey had GO terms association with “acute inflammation” (Reinhardt et al., 2013). In this way, general classes of proteins and their potential functions can be segregated by milk fraction. This type

of association can then be further analyzed by more targeted and hypothesis-driven experimentation.

Although databases like Gene Ontology are constantly being updated, they are not complete. In the experiment assessing the changes in the proteome of milk fractions during a mastitis discussed above, researchers found a large number of proteins that were associated with NET but had no associated GO term with an immunological function. Their solution was to generate their own database of proteins identified in the literature to have been associated with NET and compare that with the proteomic data set (Reinhardt et al., 2013). Generation of new or unique databases with unique functions or annotations is a valid pos-

Figure 1. Graphical representation of gene and protein functional associations, as an example of a visual means to group genes or proteins by associated function. In this type of analysis, each gene or protein is represented by a node (circle), and other genes or proteins with a known functional association to the primary gene or protein are typically close in proximity and connected with a line. The blue circle indicates a cluster of associated genes that are highly interconnected. Green (red) shapes indicate significantly upregulated (downregulated) genes in swarming bacteria compared with swimming.

sibility when standard databases fail to provide needed information.

BIOLOGICAL SIGNIFICANCE

Omics experiments are powerful tools to ask biological questions that were previously very difficult to formulate. They allow insight into the effects of various biological stimuli on a whole system and provide a uniquely global view. Furthermore, omics experiments are straightforward, and many institutions have access to core facilities that can provide these techniques as a service. As animal researchers, our goal for omics experiments is to improve the health and productivity of agriculturally important species. It is unlikely that results from omics data will directly lead to improved animal health and production. More likely, omics data will result in information that will generate new hypotheses and ideas and fulfill the promise of question-based research leading to hypothesis-based research. However, if the hypotheses generated from omics data are never tested, of what value were the omics data? With limited funding available, questions about the proper balance between question-based research and hypothesis-based research are critical in animal science research. Those of us who use omics methodology need to consider what our responsibility is to ensure that the effort and resources expended on omics are followed by hypothesis-based research.

Examples of hypothesis-based research that was the result of an omics experiment is difficult to document if both did not come from the same laboratory, which may be less common because those skill sets can be very divergent. In a Nature podcast from May 16, 2018 (<https://www.nature.com/articles/d41586-018-05172-3>), microbiologist Adam Duetschbauer said, “We have millions and millions of different gene sequences and virtually none of those genes have been studied experimentally.” This quote seems to indicate that the transition from omics experimentation to hypothesis-driven research is rare. A major pitfall of omics data is the ease with which the data can be generated. The temptation is to generate these wonderful vast data sets, and instead of following up on the interesting observations from the data and generating new hypotheses, we just move on to generating the next omics data set. We suggest that our community of animal scientists help move our omics data set from observational research to hypothesis-driven research. We offer a few examples from our experience to illustrate where omics data have been followed up with hypothesis-based experiments.

Neutrophils are known to have suppressed function during the periparturient period. In experiments to determine the change in protein expression in neutrophils

isolated from cows before and during the periparturient period, we found that histone H2A.1 was downregulated in periparturient neutrophils (Lippolis et al., 2006a). As stated above, histones have recently been associated with a new neutrophil-killing mechanism called NET. We formed a hypothesis that bovine neutrophils could form NET and showed that NET formation in fresh milk was uninhibited after hours of incubation in milk, unlike other neutrophil-killing mechanisms (Lippolis et al., 2006b). Recently, we have described the genomic, transcriptomic, and proteomic difference between *E. coli* strains that cause a transient versus a persistent infection (Lippolis et al., 2014, 2017; Thacker et al., 2017). The result of those experiments was the identification of motility, iron-regulation, and capsid differences between strains that cause persistent and transient infections. We then tested the hypotheses that *E. coli* strains that result in transient infections were less motile, reacted to iron differently, and were more sensitive to complement due to lack of capsid proteins compared with the strains that cause persistent infections (Lippolis et al., 2016, 2017). These are just a few examples of how we have tried to form hypotheses from our omics data and test those hypotheses experimentally.

SUMMARY

Has the promise of omics technologies been fulfilled? If we were to put our answer in the form of a grade, we would give the promise of omics technology an “incomplete,” albeit not due to failure of the technical aspects of omics experiments. Technical advances in nucleotide sequencing techniques and instrumentation and protein and peptide detection by mass spectrometer are allowing faster and better omics data to be generated. The promise of omics is incomplete in part because it is a rather new set of technologies that is still undergoing improvements in the tools and processes to interpret the data sets. However, the promise of omics can only be fulfilled as we strive to complete and strengthen the functional databases and to follow the omics observations with hypothesis-driven research. We believe that the deficiencies of functional databases can best be corrected by “old-fashioned” biochemical and biophysical experiments, likely at the single gene or protein level. Importantly, the hypotheses for the old-fashioned biochemical and biophysical experiments will likely come from observations from omics experiments. The functional databases need to reflect that many proteins can have multiple functions. Critically, these databases should be annotated to a level of detail such that species-specific functional databases are available. Recently, a consortium has been formed to “accelerate genome-to-phenome” knowledge in agricultural animals

called the Functional Annotation of Animal Genomes (FAANG) project (FAANG Consortium et al., 2015). Efforts of the FAANG project will be essential to truly understanding the biology that is critical to health and production of agricultural animals. Last, the animal science community must resolve to ensure that the ever-growing omics datasets progress from question-based research to hypothesis-based research. Omics data already in the literature should be mined for important hypotheses, and these hypotheses should be explored and tested. Current and future omics experiments should be conducted with the anticipation that the results will yield exciting hypotheses that will be tested. With limited resources for research of animal health issues, we need to ensure that our omics data are used to the fullest potential.

ACKNOWLEDGMENTS

The USDA was the sole funder for this research (Project Number: 5030-32000-115-00-D). The USDA determines the research priorities of all research that it funds. Publication of research is subject to review of USDA officials. Specific study design, data collection, data analysis, and preparation of the manuscript are under the control of the individual researchers. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendations or endorsement by the USDA. USDA is an equal opportunity provider and employer.

REFERENCES

- Ajay, S. S., S. C. J. Parker, H. O. Abaan, K. V. F. Fajardo, and E. H. Margulies. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* 21:1498–1505. <https://doi.org/10.1101/gr.123638.111>.
- Bionaz, M., and J. J. Loo. 2012. Ruminant metabolic systems biology: Reconstruction and integration of transcriptome dynamics underlying functional responses of tissues to nutrition and physiological state. *Gene Regul. Syst. Bio.* 6:109–125. <https://doi.org/10.4137/GRSB.S9852>.
- Boddicker, N., E. H. Waide, R. R. Rowland, J. K. Lunney, D. J. Garrick, J. M. Reedy, and J. C. M. Dekkers. 2012. Evidence for a major QTL associated with host response to porcine reproductive and respiratory syndrome virus challenge. *J. Anim. Sci.* 90:1733–1746. <https://doi.org/10.2527/jas.2011-4464>.
- Brinkmann, V., U. Reichard, C. Goosmann, B. Fauler, Y. Uhlemann, D. S. Weiss, Y. Weinrauch, and A. Zychlinsky. 2004. Neutrophil extracellular traps kill bacteria. *Science* 303:1532–1535. <https://doi.org/10.1126/science.1092385>.
- Bruning, O., H. Rauwerda, R. J. Dekker, W. C. de Leeuw, P. F. K. Wackers, W. A. Ensink, M. J. Jonker, and T. M. Breit. 2015. Valuable lessons-learned in transcriptomics experimentation. *Transcription* 6:51–55. <https://doi.org/10.1080/21541264.2015.1064195>.
- Casas, E., B. E. Hessman, J. W. Keele, and J. F. Ridpath. 2015. A genome-wide association study for the incidence of persistent bovine viral diarrhea virus infection in cattle. *Anim. Genet.* 46:8–15. <https://doi.org/10.1111/age.12239>.
- Casas, E., R. J. Leach, T. A. Reinhardt, R. M. Thallman, J. D. Lippolis, G. L. Bennett, and L. A. Kuehn. 2013. A genome-wide association study identified CYP2J2 as a gene controlling serum vitamin D status in beef cattle. *J. Anim. Sci.* 91:3549–3556. <https://doi.org/10.2527/jas.2012-6020>.
- Chaussabel, D., V. Pascual, and J. Banchereau. 2010. Assessing the human immune system through blood transcriptomics. *BMC Biol.* 8:84. <https://doi.org/10.1186/1741-7007-8-84>.
- Chen, C., H. Huang, and C. H. Wu. 2017. Protein bioinformatics databases and resources. *Methods Mol. Biol.* 1558:3–39. https://doi.org/10.1007/978-1-4939-6783-4_1.
- Cole, J. B., G. R. Wiggins, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics* 12:408. <https://doi.org/10.1186/1471-2164-12-408>.
- Collins, F. S., E. D. Green, A. E. Guttmacher, M. S. Guyer, and US National Human Genome Research Institute. 2003. A vision for the future of genomics research. *Nature* 422:835–847. <https://doi.org/10.1038/nature01626>.
- Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. <https://doi.org/10.1186/s13059-016-0881-8>.
- Costa, A. S. H., P. Costa, S. P. Alves, C. M. Alfaia, J. A. M. Prates, V. Vleck, I. Cassar-Malek, J.-F. Hocquette, and R. J. B. Bessa. 2018. Does growth path influence beef lipid deposition and fatty acid composition? *PLoS One* 13:e0193875. <https://doi.org/10.1371/journal.pone.0193875>.
- Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim. Sci.* 82(E-Suppl):E313–E328. https://doi.org/10.2527/2004.8213_supplE313x.
- Elsik, C. G., R. L. Tellam, K. C. Worley, R. A. Gibbs, D. M. Muzny, G. M. Weinstock, D. L. Adelson, E. E. Eichler, et al. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324:522–528. <https://doi.org/10.1126/science.1169588>.
- FAANG Consortium, L. Andersson, A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess, D. W. Burt, E. Casas, H. H. Cheng, et al. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16:57. <https://doi.org/10.1186/s13059-015-0622-4>.
- Fang, G., N. Bhardwaj, R. Robilotto, and M. B. Gerstein. 2010. Getting started in gene orthology and functional analysis. *PLOS Comput. Biol.* 6:e1000703. <https://doi.org/10.1371/journal.pcbi.1000703>.
- Ferreira, A. M., S. L. Bislev, E. Bendixen, and A. M. Almeida. 2013. The mammary gland in domestic ruminants: A systems biology perspective. *J. Proteomics* 94:110–123. <https://doi.org/10.1016/j.jprot.2013.09.012>.
- Gallejo Romero, I., A. A. Pai, J. Tung, and Y. Gilad. 2014. RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biol.* 12:42. <https://doi.org/10.1186/1741-7007-12-42>.
- Gergen, J. P., R. H. Stern, and P. C. Wensink. 1979. Filter replicas and permanent collections of recombinant DNA plasmids. *Nucleic Acids Res.* 7:2115–2136.
- Gupta, N., and P. A. Pevzner. 2009. False discovery rates of protein identifications: A strike against the two-peptide rule. *J. Proteome Res.* 8:4173–4181. <https://doi.org/10.1021/pr9004794>.
- Hernández-Castellano, L. E., A. M. Ferreira, P. Nanni, J. Grossmann, A. Argüello, J. Capote, G. Cai, J. D. Lippolis, N. Castro, and A. M. de Almeida. 2016. The goat (*Capra hircus*) mammary gland secretory tissue proteome as influenced by weight loss: A study using label free proteomics. *J. Proteomics* 145:60–69. <https://doi.org/10.1016/j.jprot.2016.03.030>.

- Hood, L., and L. Rowen. 2013. The Human Genome Project: Big science transforms biology and medicine. *Genome Med.* 5:79. <https://doi.org/10.1186/gm483>.
- Hosack, D. A., G. Dennis, B. T. Sherman, H. Lane, and R. A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4:R70. <https://doi.org/10.1186/gb-2003-4-10-r70>.
- Huberts, D. H. E. W., and I. J. van der Klei. 2010. Moonlighting proteins: An intriguing mode of multitasking. *Biochim. Biophys. Acta Mol. Cell Res.* 1803:520–525. <https://doi.org/10.1016/j.bbamcr.2010.01.022>.
- Huddleston, J., M. J. P. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C.-S. Chin, J. Korlach, R. K. Wilson, and E. E. Eichler. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27:677–685. <https://doi.org/10.1101/gr.214007.116>.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945. <https://doi.org/10.1038/nature03001>.
- Jiang, J., Y. Gao, Y. Hou, W. Li, S. Zhang, Q. Zhang, and D. Sun. 2016. Whole-genome resequencing of Holstein bulls for indel discovery and identification of genes associated with milk composition traits in dairy cattle. *PLoS One* 11:e0168946. <https://doi.org/10.1371/journal.pone.0168946>.
- Kappes, S. M., J. W. Keele, R. T. Stone, R. A. McGraw, T. S. Sonstegard, T. P. Smith, N. L. Lopez-Corrales, and C. W. Beattie. 1997. A second-generation linkage map of the bovine genome. *Genome Res.* 7:235–249.
- Kusebauch, U., D. S. Campbell, E. W. Deutsch, C. S. Chu, D. A. Spicer, M.-Y. Brusniak, J. Slagel, Z. Sun, J. Stevens, B. Grimes, D. Shteynberg, M. R. Hoopmann, P. Blattmann, A. V. Ratushny, O. Rinner, P. Picotti, C. Carapito, C.-Y. Huang, M. Kapousouz, H. Lam, T. Tran, E. Demir, J. D. Aitchison, C. Sander, L. Hood, R. H. Aebersold, and R. L. Moritz. 2016. Human SRMatlas: A resource of targeted assays to quantify the complete human proteome. *Cell* 166:766–778. <https://doi.org/10.1016/j.cell.2016.06.041>.
- Kusebauch, U., L. E. Hernández-Castellano, S. L. Bislev, R. L. Moritz, C. M. Rontved, and E. Bendixen. 2018. Selected reaction monitoring mass spectrometry of mastitis milk reveals pathogen-specific regulation of bovine host response proteins. *J. Dairy Sci.* 101:6532–6541. <https://doi.org/10.3168/jds.2017-14312>.
- Larkin, D. M., A. Everts-van der Wind, M. Rebeiz, P. A. Schweitzer, S. Bachman, C. Green, C. L. Wright, E. J. Campos, L. D. Benson, J. Edwards, L. Liu, K. Osogawa, J. E. Womack, P. J. de Jong, and H. A. Lewin. 2003. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Res.* 13:1966–1972. <https://doi.org/10.1101/gr.1560203>.
- Lawless, N., T. A. Reinhardt, K. Bryan, M. Baker, B. Pesch, D. Zimmerman, K. Zuelke, T. Sonstegard, C. O'Farrelly, J. D. Lippolis, and D. J. Lynn. 2014. MicroRNA regulation of bovine monocyte inflammatory and metabolic networks in an in vivo infection model. *G3 (Bethesda)* 4:957–971. <https://doi.org/10.1534/g3.113.009936>.
- Li, S., Q. Wang, X. Lin, X. Jin, L. Liu, C. Wang, Q. Chen, J. Liu, and H. Liu. 2017. The use of “omics” in lactation research in dairy cows. *Int. J. Mol. Sci.* 18:983. <https://doi.org/10.3390/ijms18050983>.
- Lindholm-Perry, A. K., A. K. Sexten, L. A. Kuehn, T. P. L. Smith, D. A. King, S. D. Shackelford, T. L. Wheeler, C. L. Monrell, T. G. Jenkins, W. M. Snelling, and H. C. Freetly. 2011. Association, effects and validation of polymorphisms within the NCAPG-LCORN locus located on BTA6 with feed intake, gain, meat and carcass traits in beef cattle. *BMC Genet.* 12:103. <https://doi.org/10.1186/1471-2156-12-103>.
- Lippolis, J. D., B. W. Brunelle, T. A. Reinhardt, R. E. Sacco, B. J. Nonnecke, B. Dogan, K. Simpson, and Y. H. Schukken. 2014. Proteomic analysis reveals protein expression differences in *Escherichia coli* strains associated with persistent versus transient mastitis. *J. Proteomics* 108:373–381. <https://doi.org/10.1016/j.jprot.2014.06.008>.
- Lippolis, J. D., B. W. Brunelle, T. A. Reinhardt, R. E. Sacco, T. C. Thacker, T. P. Looft, and T. A. Casey. 2016. Differential gene expression of three mastitis-causing *Escherichia coli* strains grown under planktonic, swimming, and swarming culture conditions. *mSystems* 1:e00064-16. <https://doi.org/10.1128/mSystems.00064-16>.
- Lippolis, J. D., D. B. Holman, B. W. Brunelle, T. C. Thacker, B. L. Bearson, T. A. Reinhardt, R. E. Sacco, and T. A. Casey. 2017. Genomic and transcriptomic analysis of *Escherichia coli* strains associated with persistent and transient bovine mastitis and the role of colanic acid. *Infect. Immun.* 86:e00566-e17. <https://doi.org/10.1128/IAI.00566-17>.
- Lippolis, J. D., and J. E. Nally. 2018. Considerations for farm animal proteomic experiments: An introductory view gel-based versus non-gel-based approaches. Pages 7–16 in *Proteomics in Domestic Animals: From Farm to Systems Biology*. A. M. de Almeida, D. Eckersall, and I. Miller, ed. Springer International Publishing, New York, NY.
- Lippolis, J. D., B. D. Peterson-Burch, and T. A. Reinhardt. 2006a. Differential expression analysis of proteins from neutrophils in the periparturient period and neutrophils from dexamethasone-treated dairy cows. *Vet. Immunol. Immunopathol.* 111:149–164. <https://doi.org/10.1016/j.vetimm.2005.12.001>.
- Lippolis, J. D., and T. A. Reinhardt. 2005. Proteomic survey of bovine neutrophils. *Vet. Immunol. Immunopathol.* 103:53–65. <https://doi.org/10.1016/j.vetimm.2004.08.019>.
- Lippolis, J. D., and T. A. Reinhardt. 2008. Centennial paper: Proteomics in animal science. *J. Anim. Sci.* 86:2430–2441. <https://doi.org/10.2527/jas.2008-0921>.
- Lippolis, J. D., and T. A. Reinhardt. 2010. Utility, limitations, and promise of proteomics in animal science. *Vet. Immunol. Immunopathol.* 138:241–251. <https://doi.org/10.1016/j.vetimm.2010.10.003>.
- Lippolis, J. D., T. A. Reinhardt, J. P. Goff, and R. L. Horst. 2006b. Neutrophil extracellular trap formation by bovine neutrophils is not inhibited by milk. *Vet. Immunol. Immunopathol.* 113:248–255. <https://doi.org/10.1016/j.vetimm.2006.05.004>.
- Loor, J. J., K. M. Moyes, and M. Bionaz. 2011. Functional adaptations of the transcriptome to mastitis-causing pathogens: the mammary gland and beyond. *J. Mammary Gland Biol. Neoplasia* 16:305–322. <https://doi.org/10.1007/s10911-011-9232-2>.
- Lowe, R., N. Shirley, M. Bleackley, S. Dolan, and T. Shafee. 2017. Transcriptomics technologies. *PLOS Comput. Biol.* 13:e1005457. <https://doi.org/10.1371/journal.pcbi.1005457>.
- Manzoni, C., D. A. Kia, J. Vandrovceva, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari. 2018. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* 19:286–302. <https://doi.org/10.1093/bib/bbw114>.
- Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. <https://doi.org/10.1371/journal.pone.0005350>.
- Neibergs, H. L., M. L. Settles, R. H. Whitlock, and J. F. Taylor. 2010. GSEA-SNP identifies genes associated with Johne's disease in cattle. *Mamm. Genome* 21:419–425. <https://doi.org/10.1007/s00335-010-9278-2>.
- Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40:1413–1415. <https://doi.org/10.1038/ng.259>.
- Pegolo, S., F. T. Cannizzo, B. Biolatti, M. Castagnaro, and L. Bargeloni. 2014. Transcriptomic profiling as a screening tool to detect trenbolone treatment in beef cattle. *Res. Vet. Sci.* 96:472–481. <https://doi.org/10.1016/j.rvsc.2014.03.020>.
- Pena, R. N., R. Quintanilla, A. Manunza, D. Gallardo, J. Casellas, and M. Amills. 2014. Application of the microarray technology to

- the transcriptional analysis of muscle phenotypes in pigs. *Anim. Genet.* 45:311–321. <https://doi.org/10.1111/age.12146>.
- Phetsouphanh, C., J. Zaunders, and A. Kelleher. 2015. Detecting antigen-specific T cell responses: from bulk populations to single cells. *Int. J. Mol. Sci.* 16:18878–18893. <https://doi.org/10.3390/ijms160818878>.
- Powell, E. J., T. A. Reinhardt, E. Casas, and J. D. Lippolis. 2018. The effect of pegylated granulocyte colony-stimulating factor treatment prior to experimental mastitis in lactating Holsteins. *J. Dairy Sci.* 101:8182–8193. <https://doi.org/10.3168/jds.2018-14550>.
- Rapaport, F., R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 14:R95. <https://doi.org/10.1186/gb-2013-14-9-r95>.
- Reiman, M., M. Laan, K. Rull, and S. Söber. 2017. Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples. *FASEB J.* 31:3298–3308. <https://doi.org/10.1096/fj.201601031RR>.
- Reinhardt, T. A., R. E. Sacco, B. J. Nonnecke, and J. D. Lippolis. 2013. Bovine milk proteome: Quantitative changes in normal milk exosomes, milk fat globule membranes and whey proteomes resulting from *Staphylococcus aureus* mastitis. *J. Proteomics* 82:141–154. <https://doi.org/10.1016/j.jprot.2013.02.013>.
- Schisterman, E. F., and A. Vexler. 2008. To pool or not to pool, from whether to when: Applications of pooling to biospecimens subject to a limit of detection. *Paediatr. Perinat. Epidemiol.* 22:486–496. <https://doi.org/10.1111/j.1365-3016.2008.00956.x>.
- Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22:839–851. <https://doi.org/10.1261/rna.053959.115>.
- Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. 2011. Global quantification of mammalian gene expression control. *Nature* 473:337–342. <https://doi.org/10.1038/nature10098>.
- Silvescu, C. I., and R. Sackstein. 2014. G-CSF induces membrane expression of a myeloperoxidase glycovariant that operates as an E-selectin ligand on human myeloid cells. *Proc. Natl. Acad. Sci. USA* 111:10696–10701. <https://doi.org/10.1073/pnas.1320833111>.
- Sims, D., I. Sudbery, N. E. Illott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Rev. Genet.* 15:121–132. <https://doi.org/10.1038/nrg3642>.
- Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDanel, T. P. L. Smith, T. S. Sonstegard, R. M. Thallman, and G. L. Bennett. 2010. Genome-wide association study of growth in crossbred beef cattle. *J. Anim. Sci.* 88:837–848. <https://doi.org/10.2527/jas.2009-2257>.
- Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321:956–960. <https://doi.org/10.1126/science.1160342>.
- Thacker, T. C., J. D. Lippolis, B. W. Brunelle, T. A. Casey, T. A. Reinhardt, R. E. Sacco, and D. B. Holman. 2017. Genome sequences of *Escherichia coli* strains that cause persistent and transient mastitis. *Genome Announc.* 5:e00775-e17. <https://doi.org/10.1128/genomeA.00775-17>.
- The Gene Ontology Consortium. 2015. Gene Ontology Consortium: Going forward. *Nucleic Acids Res.* 43:D1049–D1056. <https://doi.org/10.1093/nar/gku1179>.
- Vidova, V., and Z. Spacil. 2017. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal. Chim. Acta* 964:7–23. <https://doi.org/10.1016/j.aca.2017.01.059>.
- Wiggans, G. R., J. B. Cole, S. M. Hubbard, and T. S. Sonstegard. 2017. Genomic selection in dairy cattle: The USDA experience. *Annu. Rev. Anim. Biosci.* 5:309–327. <https://doi.org/10.1146/annurev-animal-021815-111422>.
- Winterbourn, C. C., M. C. Vissers, and A. J. Kettle. 2000. Myeloperoxidase. *Curr. Opin. Hematol.* 7:53–58.
- Zhang, S. D., and T. W. Gant. 2005. Effect of pooling samples on the efficiency of comparative studies using microarrays. *Bioinformatics* 21:4378–4383.
- Zhao, S., W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 9:e78644. <https://doi.org/10.1371/journal.pone.0078644>.
- Zubarev, R. A. 2013. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* 13:723–726. <https://doi.org/10.1002/pmic.201200451>.