© 2010 Heather Ann Adams

ENHANCED TRANSCRIPTOME PROFILING AND BIOMARKER DISCOVERY USING META-ANALYTICAL TECHNIQUES

BY

HEATHER ANN ADAMS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Animal Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Sandra L. Rodriguez-Zas, Chair
Professor Harris A. Lewin
Professor Gene E. Robinson
Professor Roger D. Shanks

ABSTRACT

This is a comprehensive study of the application of meta-analytical techniques to gene expression data to detect differentially expressed genes across experiments and conditions, and as a means of classifying unknown samples. Three chapters were created to: (1) investigate three novel meta-analysis approaches as a means of combining and analyzing gene expression experiments for differential expression across experiments, (2) develop a novel meta-classification analysis to accurately classify samples to groups, and identify candidate transcript biomarkers, and (3) apply novel meta-analysis approaches to experiments with related but different conditions. The novelty of the approaches is seen with the inclusion of estimates of differential expression (effect size) as well as the sign of expression (up- or down-regulation) for genes, where previous meta-analysis studies have combined results based solely on P-values or effect sizes alone.

Aim 1: The objectives were 1) to use different types of information and develop and compare different meta-analyses approaches to addresses limitations of previous microarray meta-analysis work and 2) to apply these approaches to enhance the knowledge on the expression patterns between honey bees representing different age and behavior groups. A novel study-level meta-analysis approach was developed that combined estimates of fold change between groups because in many cases, only fold change results are available. The novelty of this approach stemmed from the adjustment of the fold change estimates by their precision and weighting of each study by the individual variation. Recognizing the importance of considering within- and between-study variation from the study-level meta-analysis, a sample-level meta-analysis model was extended to include heterogeneity between studies. The sample-level meta-analysis

combined normalized gene expression values and the comparison of study and sample-level meta-analyses allowed for the assessment of the relative advantages of using two types of information. Lastly, the new study- and sample-level approaches were compared to the traditional approach of identifying overlapping genes among the list of differentially expressed genes between two or more studies.

Aim 2: The objectives were to 1) develop and test complementary meta-classification approaches to integrate different types of information across studies and gain precision on the identification of gene expression biomarkers to classify samples, 2) to apply the new approaches to identify gene expression profiles that can accurately classify two honey bee behavior groups, and 3) to compare different methods to identify a subset of gene profiles that can provide accurate and generalizable sample classification. In the context of using gene expression information from multiple microarray experiments to classify honey bee samples into age or behavior groups, the response and explanatory variables were gene expression and group respectively in the previous chapter, meanwhile they correspond to group and gene expression, respectively in the present chapter. A linear mixed-effects model was used to meta-analyze the Normally distributed gene expression variable from Aim 1, and a binary logistic mixed-effects model was used to meta-classify samples in this aim. A novel three-stage strategy that allowed the identification, selection and combination of accurate and generalizable biomarkers was developed. In the first stage, three approaches that integrate information across studies within gene were implemented: a) individual-study analysis, b) study-level meta-classification analysis, and c) sample-level meta-classification analysis. Promising classifier genes to use in the second stage of the classification strategy were identified

based on the statistical significance and consistency of the association with the odds of behavior across analyses. In the second stage, results from the first stage were incorporated into multiple classification approaches to develop classifier functions and identify classifier genes. Four classification techniques based on the simultaneous consideration of multiple genes were developed, and selection methods were used to address potential over-parameterization of the model. In the third stage, additional biomarker gene sets with comparable classification performance were uncovered from all genes in the studies using a score selection method, which allowed for flexibility to arrive at models that produced great classifiers without generating an excessive number of models with low performance. Classification functions were trained on numerous datasets, allowing for the improved ability to detect biomarkers that are strong indicators of conditional class using a larger sample of observations from the population. These biomarkers were tested on an additional dataset to assess their performance in prediction of classes.

Aim 3: The objectives were designed to address the situation of integrating information from studies that do not share conditions, and included: 1) development of a meta-analysis approach that allowed to combine information across partially overlapping studies, 2) implementation of the approach to characterize gene expression profiles in cattle embryos spanning a variety of conditions, 3) the comparison of results of this novel approach against other standard meta-analysis approaches. To accomplish the first objective, the model developed for the sample-level meta-analysis approach in the second aim was re-parameterized. To accomplish the second objective, four cattle microarray experiments that partially overlapped on the conditions studied were meta-analyzed. The

samples included 7 day embryos, 25 day embryos and extra-embryonic tissue, and term (or near term, approximately 280 day) extra-embryonic tissue and a reference sample. In addition to developmental age and tissue source, samples pertained to one of two reproductive technologies, artificial insemination or somatic cell nuclear transfer. The availability of conditions not present in all studies allowed testing of the novel meta-analysis model. The availability of the two reproductive technologies across all studies allowed the comparison of standard and novel meta-analysis techniques. Results from this novel extension allowed the characterization of comprehensive expression profiles that extend conditions not studied within a single experiment.

To my family, who have shown me the beauty of life, of love, of faith, and the importance of family and friends – each member a blessing, and each memory we share a treasure. Grateful thanks for the strong and positive presence all have had, and will continue to have, in my life's endeavors. I hope that each one realizes how much they are appreciated and loved.

ACKNOWLEDGEMENTS

I am very grateful to the many people who have been such great inspiration, support and help to me during this project, and my graduate career as a whole. First and foremost, thank you to my doctoral advisor Dr. Sandra Rodriguez-Zas, who has been a vital part to all of my doctoral successes, and a great role model and inspiration for all women in academia. Thank you to Dr. Roger Shanks for continuing to support me beyond my Master's project, serving on my doctoral committee, and entrusting me with instructing his ANSC 444 Applied Animal Genetics course while on sabbatical (a priceless opportunity!). Thank you to my doctoral committee members, Dr. Harris Lewin, Dr. Gene Robinson, and Dr. Roger Shanks, who provided me with crucial advice and direction during my preliminary exam and throughout the doctoral process. Thank you to Dr. Bruce Southey for his programming support, advice, and numerous manuscript corrections. Thank you to my bioinformatics lab colleagues Younhee Ko, Nicola Serao, Kristin Delfino, Dr. Kyuil Cho, and departmental colleague Kathleen Barry, for their support (and coffee breaks!). Thank you to Dr. David and Mrs. Norraine Baker for awarding me the Baker Graduate Fellowship in Animal Sciences, which has provided me with great financial support towards the completion of my project. Thank you to Dr. Bruce Schatz for funding support through the BeeSpace project. And last but certainly not least, my family – my supporting foundation – who are with me always, and have been my inspirations in life.

TABLE OF CONTENTS

Page

List of Figures ..... **xi**

List of Tables ..... **xiii**

1. Chapter 1: Introduction **1**

1.1 Motivation 1

1.2 Background 3

2. Chapter 2: Meta-analysis of genome-wide expression patterns associated with behavioral maturation in honey bees **49**

2.1 Abstract 49

2.2 Background and Motivation 50

2.3 Materials and Methods 53

2.4 Results 61

2.5 Discussion 71

2.6 Conclusions 84

3. Chapter 3: Development of meta-classification techniques to identify gene biomarkers that accurately classify samples **86**

3.1 Abstract 86

3.2 Background and Motivation 88

3.3 Materials and Methods 92

3.4 Results 104

3.5 Conclusions 122

4. Chapter 4: Meta-analysis of gene expression studies provides insights into cattle embryo development across reproductive techniques, developmental stages and sources 124

4.1 Abstract 124

4.2 Background and Motivation 125

4.3 Materials and Methods 126

4.4 Results and Discussion 133

4.5 Conclusions 154

References 158

Appendix A: Chapter 2 Gene ontology and fruit fly annotation information for sample-level meta-analysis results 204

Appendix B: Chapter 4 Results for individual studies comparing AI and NT 205

Appendix C: Chapter 4 Results for study-level and sample-level meta-analyses comparing AI and NT 206

Appendix D: Chapter 4 Results of contrasts of developmental stage within reproductive technology and tissue source 207

Appendix E: Chapter 4 Results of contrasts of tissue source within reproductive technology at 25 days gestation 208

Appendix F: Chapter 4 Results of contrasts of reproductive technology for 25 day and 280 day extra-embryonic samples 209

Appendix G: Chapter 4 Results of contrasts of reproductive technology for 8 day and 25 day embryonic samples 210

Author's Curriculum Vitae 211

LIST OF FIGURES

Figure **Page**

Chapter 1

1.1 The three main steps to build a spotted microarray..... 5

1.2 Basic steps to obtain microarray gene expression measurements including sample collection and preparation, co-hybridization to microarray slide, and laser scanning 6

1.3 Reference (A) and loop (B) designs in microarray experiments..... 8

1.4 Morphologies of honey bees 39

1.5 Diagram of oocyte and embryo development 45

1.6 Steps in somatic cell nuclear transfer cloning 46

Chapter 2

2.1 Venn diagram of the number of differentially expressed transcripts detected by at least two individual-study analyses (Ind), study-level (Study), and sample-level (Sample) meta-analyses..... 174

2.2 Funnel plots of differential expression estimates and 95% confidence interval limits for chosen transcripts 175

Chapter 3

3.1 Principal component analysis of the top 25 transcripts by individual sample..... 176

Chapter 4

4.1 Venn diagram of the number of differentially expressed transcripts detected by one or more individual-study analyses (Ind), the study-level (Study), and the sample-level (Sample) meta-analyses. 176

4.2 Funnel plots of differential expression estimates and 95% confidence interval limits for chosen genes. 177

4.3 Venn diagrams of the number of differentially expressed genes detected within and between technologies, stages and sources. 178

4.4 Expression of prostaglandin D2 synthase gene within reproductive technology for different combinations of tissue source and developmental stage. 179

LIST OF TABLES

Table **Page**

Chapter 1

1.1 Estimated number of genes in the genome of the fruit fly, human, honey bee and cattle genomes. 3

1.2 Relationships among observed and predicted classification of outcomes into two classes (A and B), and related indicators. 31

Chapter 2

2.1 Number of transcripts with differential expression within individual analyses, study-level standardized and sample-level meta-analysis 180

2.2 Number of transcripts with positive and negative differential expression within individual analysis, study-level standardized and sample-level meta-analysis 181

2.3 Gene Ontology information for 12 transcripts significantly differentially expressed in four or more studies.....182

2.4 Number of non-significant and significant transcripts within Group.....185

Chapter 3

3.1 Transcripts identified as having association with behavior..... 186

3.2 Top twenty-five transcripts identified by meta-analyses, and associated gene information..... 187

3.3 Results of logistic analysis of top 25 transcripts individually.....190

3.4 Results of linear combination of top five transcripts..... 191

3.5 Pearson correlation coefficients for top 25 transcripts 192

3.6 Results of logistic analysis of top 25 transcripts considered simultaneously 194
3.7 Results of stepwise logistic regression of top 25 transcripts 194
3.8 Linear discriminant functions for behavior of the top 25 transcripts 195
3.9 Results of discriminant analysis of top 25 transcripts after selection 195
3.10 Results of stepwise discriminant analysis (stepdisc) of top twenty-five transcripts for three selection methods 198
3.11 Results of leave-one-study-out analysis 199
3.12 Results of leave-one-study-out analysis, using three selection methods per study 200

Chapter 4

4.1 Number of differentially expressed genes between samples created using artificial insemination (AI) and somatic cell nuclear transfer (NT) 201
4.2 Number of differentially and non-differentially expressed genes between samples created using artificial insemination (AI) and somatic cell nuclear transfer (NT) with combinations of individual studies, study-level and sample-level meta-analyses 202
4.3 Gene information for 8 cattle genes with significant differential expression in two individual studies with opposite signs 202
4.4 Differentially expressed genes between conditions within developmental stages, sample sources, and reproductive technologies 203

CHAPTER 1

INTRODUCTION

1.1 Motivation

Microarray gene expression experiments aim to characterize the expression level of transcripts in a variety conditions. The integration of gene expression information from multiple microarray experiments can enhance the characterization of gene expression profiles that are consistently expressed across experiments, minimize the impact of experiment-specific factors, and support a more accurate identification of gene biomarkers, functional categories, and pathways associated with the process of interest than results from individual experiments. Meta-analysis encompasses a series of techniques to integrate multiple independent microarray experiments that vary on the information used from each experiment and the approach to combine the information. Examples of meta-analysis include overlap of genes detected to be differentially expressed across conditions to combination of indicators of differential expression (e.g. P-value, fold change) across experiments. The overall objective of the proposed research was to develop and evaluate novel approaches to integrate information from microarray experiments that will offer further and more accurate understanding of the transcriptome. Three specific aims (or research chapters) supported the accomplishment of this objective:

Aim 1. *Comparison of meta-analytical approaches to detect differentially expressed genes across experiments.* This aim focused on investigating the

complementary advantages of three meta-analysis approaches that use different types of information to detect genes associated with behavioral maturation in honey bees.

Aim 2. *Development of meta-classification techniques to accurately classify or assign samples to classes or groups, and identify transcriptome biomarkers.* A novel integration of classification and meta-analysis techniques is proposed to develop predictive equations and classify new samples based on the gene expression values. The capability of the meta-classification approach to assign honey bees to maturation stages based on their gene expression levels and identify genes that can be used to classify honey bees is investigated. Genes that exhibit (or fail to exhibit) differential expression across conditions will be identified.

Aim 3. *Assessment of meta-analysis approaches to integrate evidence of differential expression from experiments of related but different conditions*. Gene expression information from experiments that characterize the transcriptome of different components influencing embryo development (e.g. reproductive technique used to obtain the associated embryo, placentome) at different stages was integrated. Information from microarray experiments that do not share conditions was analyzed using a re-parameterized sample-level meta-analysis to characterize gene expression profiles in cattle embryos spanning a variety of conditions.

1.2 Background

1.2.1 Microarray Technology

The genome of a number of organisms (**Table 1.1**) has been (e.g. human, mouse, fruit fly, honey bee, cattle), or is in the process of being sequenced (e.g. pig, horse).

Table 1.1. Estimated number of genes in the genome of the fruit fly, human, honey bee and cattle genomes.

Organism	**Number of Genes**	**Genome size in base pairs (bp)**	**Reference**
D. melanogaster	~13,600[1]	~200 Mbp[2]	[1]Adams *et al.*, 2000 [2]Kimball, 2008
H. sapiens	~25,000[3]	~3 Gbp[3,4]	[3]IHGSC, 2004 [4]NHGRI, NIH, 2008
A. mellifera	~12,000[5]	~300 Mbp[6]	[5]Honeybee Genome Sequencing Consortium, 2006 [6]HGSC, 2005
B. taurus	~25,000[7]	~3 Gbp[7]	[7]BGSC, 2007

Sequencing and cDNA projects have been crucial in the development of microarray gene expression technology that allows the profiling of the expression of thousands of genes simultaneously and understanding of gene regulation on a genome-wide scale (Everts *et al.*, 2005; Schena *et al.*, 1995; Whitfield *et al.*, 2002).

A Deoxyribonucleic acid (DNA) microarray is used to detect the presence and abundance of labeled nucleic acids (i.e. measuring gene expression levels) in samples. Gene expression describes the transcription of information contained within the DNA into messenger RNA (mRNA) that is then translated into the proteins critical for the functions of cells. Gene expression measures the amount of mRNA concentration found in a cell after transcription. By examining the kinds and amounts of mRNA that are produced by a cell, scientists are able to identify which genes are expressed, which in turn reveals how cells adapt to changes within and outside of (environmental changes)

the organism. Of the many microarray platforms available, a review of two-dye spotted microarrays is presented. **Figure 1.1** depicts the process of manufacturing a spotted microarray slide.

A spotted DNA microarray is made up of a solid glass slide, typically a microscope slide, with DNA molecules chemically bonded to the surface (Stekel, 2003). The DNA probes are either nucleotide products of polymerase chain reaction (PCR), or are synthesized. Briefly, PCR is a technique used to amplify one or more copies of a short sequence of DNA into millions of copies. These copies are then purified and are used as probes on the microarray (Hoy, 2003). Oligonucleotides are short segments of RNA or DNA synthesized by cleaving longer segments, or by polymerizing (joining of) individual nucleotide precursors (Stekel, 2003). The probes are organized in microtiter well plates and deposited onto the slide surface by a spotting robot. Chemical linkers are attached to the slide in preparation for the bonding of the probe to the slide. The probes bond to the microarray slide surface through covalent (aliphatic amine group added to probe) or non-covalent (amine group added to slide) bonds. A spotting robot, consisting of numerous pins (solid or hollow), grabs liquid from each well of the plate to spot on the slide. Once the liquid is attached to the slide, the pins are washed, and the process repeats again, using a second set of wells corresponding to another batch of DNA. The process repeats until the microarray is complete (Stekel, 2003).

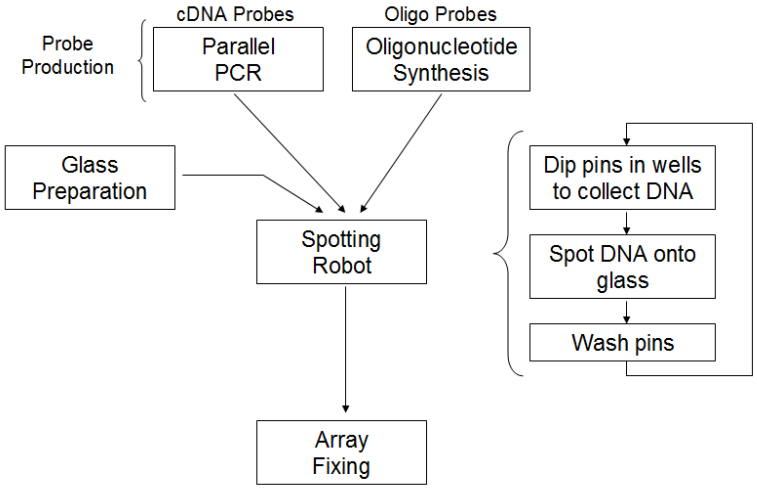


Figure 1.1. The three main steps (probe production, spotting robot, and microarray fixing) to build a spotted microarray (Stekel, 2003).

Figure 1.2 shows the basic steps of a spotted microarray experiment from sample collection (e.g. cells) to user scanning and subsequent measurement of fluorescence intensities of each spot in each microarray. The preparation of the samples for hybridization to the microarray slides starts with the extraction of mRNA, reverse transcription followed by fluorescent labeling of the complementary DNA (cDNA). In spotted microarray platforms, typically two samples are hybridized to the slide and a competitive reaction takes place. To differentiate the samples, the mRNA from each sample is labeled with a different dye that when scanned, fluoresce at two wavelengths (Cy3 or green, and Cy5 or red). Two common methods of labeling are direct and indirect labeling. Both incorporate the dye by reverse transcriptase. In direct labeling, a fluorescently labeled cytosine nucleic acid (dC) is used in addition to the usual nucleic acids added to the transcription reaction, whereas for indirect labeling an amino-allyl-modified dC is used (Stekel, 2003). The latter molecule is smaller, and allows for reverse transcription to occur more efficiently (Stekel, 2003). After labeling, the samples are co-

hybridized to the microarray slide and the slides are washed to remove any excess solution. The final step is to acquire an image of the surface of the microarray. A scanner with two wave lengths is used to excite the two dyes of the two samples that were co-hybridized. The dyes fluoresce, and this fluorescence is recorded as digital image pixels. Each dye is excited during a separate scan and red and green images are obtained. The intensity of each dye represents the level of hybridization of the mRNA in the sample (target) to the probe, or expression level of the DNA. High intensity suggests high expression, meanwhile low intensity suggests low expression or abundance. Both images are then combined. Assuming that sample 1 (i.e. cells in condition 1) was labeled with red dye and sample 2 was labeled with green dye, then green spots indicate over-expression of the genes represented by the probes in sample 1 relative to sample 2 and red spots represent over-expression of the gene in sample 2 relative to sample 1. Yellow spots are observed when the gene is similarly expressed in both samples. The information from the image is transformed into numerical intensities and these intensities are the basis for the statistical analysis of the microarray.

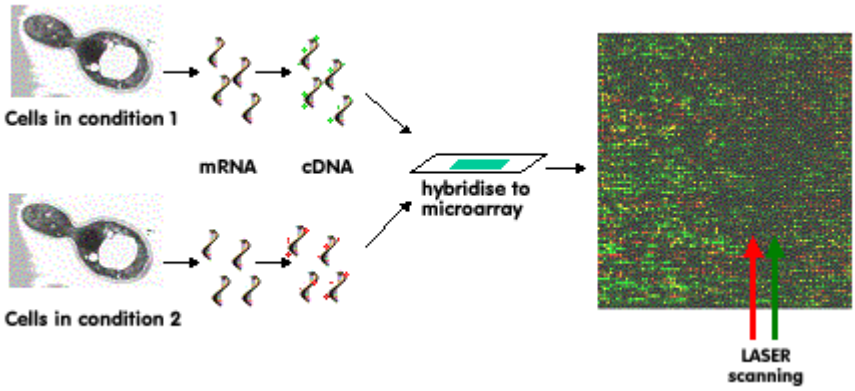


Figure 1.2. Basic steps to obtain microarray gene expression measurements including sample collection (e.g. cells) and preparation (obtain mRNA and fluorescence labeling of cDNA), co-hybridization to microarray slide, and laser scanning (Brazma *et al.*, 2008).

1.2.2 Analysis of Microarray Gene Expression Data

The analytical workflow of a microarray gene expression experiment can be partitioned into five main steps 1) experimental design, 2) data preprocessing, 3) inference (detection of differentially expressed genes), 4) classification (supervised and unsupervised) and 5) validation of results.

1.2.2.1 Experimental design

The first step of the microarray experiment analytical workflow, experimental design, involves the development of a plan to assign samples to microarrays maximizing the quality and quantity of information obtained from the experiment and minimizing the resources required. Technical and biological sources of variation and confounding effects (extraneous or undesirable variables or effects) must be considered in the design of experiments. Biological variation encompasses sample-to-sample variation observed in the population being studied. Technical variation in the experimental setup encompasses variation due to, for example, dye, microarray, pin tip and scanner (Stekel, 2003). Confounding effects are those effects in which the researcher is unable to separate as to which is responsible for differences in gene expression (Stekel, 2003). If in an experiment, more than one technician runs the experiments and different batches are used, it can be unclear as to which effect is responsible for expression differences. In this case, treatment, technician and batch are confounded effects.

In two-dye spotted microarray experiments, two experimental designs are typically used: reference and loop designs.

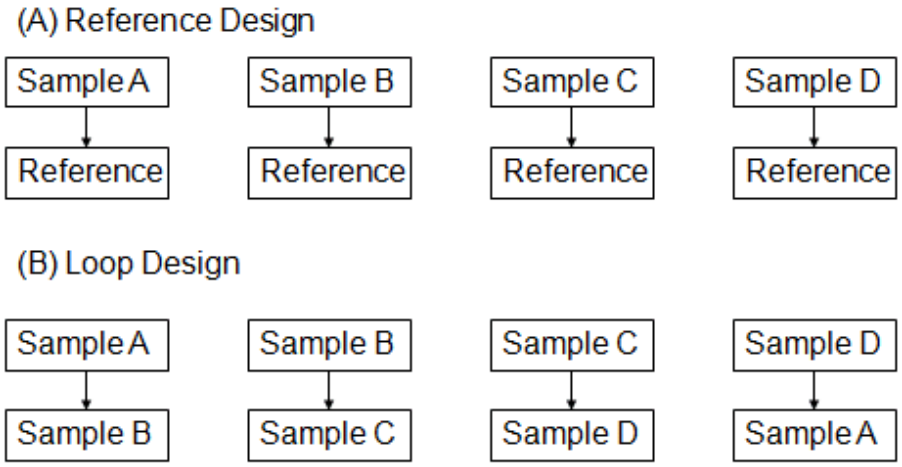


Figure 1.3. Reference (A) and loop (B) designs in microarray experiments.

Figure 1.3 depicts two types of microarray experimental designs. **Figure 1.3A** represents a reference design experiment, where all samples are directly compared to a reference sample. Indirect comparisons can be made between all samples. **Figure 1.3B** represents a loop design where direct comparisons can be made between samples A and B, samples B and C, samples C and D, and samples D and A. In this type of design, indirect comparisons can be made between samples A and C, and samples B and D. Reference designs are the most straight-forward and easy to implement designs, however, no direct comparisons can be made between the samples. Therefore, the variability of the comparisons between samples may be large. When loop designs are used, both direct and indirect comparisons reduced the variability of the comparison between samples relative to reference designs. Therefore, to decrease the variability of the comparison between samples, it is best to use a combination of direct and indirect comparisons (Rodriguez-Zas, 2005b). The design of an experiment also requires the consideration of the number of samples needed to detect a minimum differential expression for a particular standard error, type I error and statistical power, and use of individual *versus* pooled samples.

1.2.2.2 Data preprocessing

The second step of the microarray experiment analytical workflow, data preprocessing, aims at removing unreliable observations and systematic sources of variation and includes data filtering, subtraction of background from foreground intensity measurements, transformation, normalization, and collapsing. In data filtering, the intensity values of spots that have been flagged by the scanning software (i.e. for being based on few pixels or extend to other spots), and low intensity values (i.e. lower than a fixed or microarray-dependent threshold such as median of negative control spots or the background of all spots in the microarray) are removed from analysis because they are deemed unreliable. Background intensity values indicate potential intensity artifacts and subtracting the background from the foreground intensity aims at removing intensity that is not due to hybridization of the target to the probe. Fluorescence intensity values tend to exhibit a skewed, non-normal distribution. The use of a logarithmic transformation makes the fluorescence intensity data exhibit a more Normal distribution. A base 2 logarithmic transformation (log2) is typically used to facilitate the interpretation of results. Consider the log2 of the ratio between the foreground minus background value of a sample labeled with the red dye *versus* a sample labeled with the green dye for a spot. A log2 ratio equal to 1 indicates that the gene was two-fold over expressed on the red sample relative to the green sample and a value of -1 indicates that the gene was two-fold over expressed on the green sample relative to the red sample. A log2 ratio value equal to 0 indicates that the gene measured by the probe on the spot was equally expressed on both samples.

Data normalization allows for joint analysis of gene expression measurements from different dyes, microarrays, etc (Allison *et al.*, 2006). Normalization is used to minimize technical variations in gene expression profiles of two mRNA samples allowing the differences due to the factor of interest to be more accurately estimated (Yang *et al.*, 2002). Dye effects arise as the red (Cy5) and green (Cy3) dyes can have different expression levels even if pertaining to the same mRNA sample. Typically when this occurs, intensities are higher for the green dye (Yang *et al.*, 2002). These differences may stem from differences in the physical properties of the two dyes, scanner settings, or the efficiency of dye incorporation (Yang *et al.*, 2002). Normalization can be done globally (i.e. on the entire data set) or locally (i.e. on a physical subset of the data set). Global normalization methods assume that red (Cy5) and green (Cy3) intensities are related by a constant, and the center of the distribution of log ratios is shifted to equal zero. This is represented as,

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG) \qquad [1.1]$$

where k is the constant, c is the shift parameter usually representing the median or mean of the log intensity ratios, R is the intensity value associated with the red (Cy5) dye, and G is the intensity value associated with the green (Cy3) dye (Yang *et al.*, 2002). An equivalent model was proposed by Kerr *et al.* (2002),

$$\begin{aligned} Z_{rk} &= \log_2(Y_{rk} + C) \\ Z_{gk} &= \log_2(Y_{gk} + C) \end{aligned} \qquad [1.2]$$

where C is the constant found from minimizing the deviation of each log ratio (Z_{rk} - Z_{gk}) from the median ratio of the microarray, and Y is the intensity value for the red (r) and green (g) channels. Yang *et al.* (2002) subtracts a constant from a single dye, whereas

Kerr *et al.* (2002) adds a constant to both dyes, but in both cases, they are adjusting by a constant to make the dyes equivalent.

Global normalization assumes there is a single shift between R and G intensities in the microarray. However, the systematic difference between R and G may vary with the intensity of the spot, or the pin of the robotic arm. Thus, local normalizations that remove systematic noise on a more refined or specific (less global) manner are typically necessary.

A wide range of local normalization approaches have been developed to remove systematic sources of variation within and across microarrays. The main potential sources of variation within microarray platform include dye, pin, and microarray area effects. Pin effects may become evident as one region or block of the microarray may have more intense or brighter spots than another. Loess is a curve fitting local dye normalization method, where a local regression line is fitted to the ratio R/G (or M) *versus* the average of the R and G intensities (or A), and the data are re-centered (Cui *et al.*, 2003). The spelling of this method hints as to the type of function used in the transformation. When a linear function is used to fit the regression line, the spelling is "lowess" *with* a "w", but if a quadratic function is used to fit the regression line, the spelling is "loess" *without* the "w" (Cui *et al.*, 2003). Loess is a local normalization approach because the dye adjustment is not uniform across the microarray. Furthermore, a single loess can be applied to each microarray or different loess adjustments can be implemented by block.

Lastly, when platforms include multiple spots of the same probe and particularly if these have been printed by the same print tip, probes are from the same well or replicate wells, or the replicate spots are located next to each other, then the observations

from these spots tend to be highly correlated and are not independent. To address the typical model assumption that the observations are independent, the average or median of the normalized fluorescence intensities from replicate spots are analyzed instead of the individual values, and this strategy is also known as collapsing the data. In platforms including more than one spot per probe for all probes, one gene is typically represented by two spots on a microarray, and thus by collapsing those two spots into a mean or median value, one observation per gene is now available for the analysis. Once the raw data have been preprocessed, the numerical information about gene expression abundance is ready for analysis and inference.

1.2.2.3 Inference

The third step of the microarray experiment analytical workflow encompasses the analysis of the preprocessed measurements with the goals of identifying differentially expressed genes (DEG), or in some cases identifying genes not differentially expressed, and characterizing the expression profile across the conditions evaluated in the experiment. This step of the analytical workflow also includes adjustment of statistical significance for multiple testing, and can include the analysis of the representation of biological themes (e.g. Gene Ontology groups, pathways) among a list of genes.

An important step in data analysis is fitting an appropriate model. Two common types are fixed-effects and mixed-effects models. In fixed-effects models, all factors in an experiment have predetermined sets of levels (Ott and Longnecker, 2001). A more appropriate model may be the mixed-effects model, where some factors in the experiment are predetermined, and others have been randomly selected from a population of all possible levels. In the case of microarray experiments, dye and treatment are considered

fixed effects, whereas microarray and sample are considered random effects. A fixed effects model including dye and treatment effects would be written as,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad [1.3]$$

where y_{ijk} is the gene expression measurement for dye i (i = red or green), treatment j and observation k, μ is the overall mean response, α is the effect of dye i, β is the effect of treatment j, and ε is the error term (Ott and Longnecker, 2001). An interaction between dye and treatment could also be included in the model if the interaction was of interest in the experiment. A mixed effects model would be written as,

$$\begin{aligned} &y_{ijklm} = \mu + D_i + T_j + S_k + A_l + e_{ijklm} \\ &S \sim (0, \Sigma) \\ &A \sim (0, \Delta) \end{aligned} \quad [1.4]$$

where y_{ijkm} is the gene expression measurement for dye i (i = red or green), treatment j, sample k, and microarray l, μ is the overall mean response, D is the effect of dye i, T is the effect of treatment j, S is the effect of sample k, A is the effect of microarray l, and ε is the error term.

After fitting a model, hypothesis testing is performed to identify genes that are differentially expressed across conditions. For experiments including two conditions, the most common statistical test used is the t-test. Two versions of the t-test can be used, depending on whether the data are paired or unpaired. For a paired (one-sample) t-test, each sample has two observations. The observations of each sample are combined by subtracting them to remove the sample effect, and in the case of microarrays, log ratio is obtained for each sample. The common formula for the paired t-test is,

$$t = \frac{\bar{x}}{s/\sqrt{n}} \quad [1.5]$$

where $\bar{x}$ is the mean of the log ratios for each sample, s is the standard deviation for the group of samples (or standard error), and n is the number of samples in the experiment (Stekel, 2003). The test statistic value obtained from the formula is compared to a t-distribution using the appropriate degrees of freedom (i.e. independent observations, or n-1). A p-value is generated, and if the value is less than the threshold p-value (or confidence level), the given gene or measurement is significant. An unpaired (two-sample) t-test is similar to the paired, except that the observations of each sample are analyzed separately. Therefore, there are two separate groups of observations to compare. The formula for an unpaired t-test is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad [1.6]$$

where $\bar{x}_1$ and $\bar{x}_2$ are the two group means, s_1 and s_2 are the standard deviations of the groups, and n_1 and n_2 are the sample sizes for the two groups (Stekel, 2003). The test statistic is compared against a t-distribution is conducted to test significance as in the paired t-test. Although t-tests are good simple tests to apply to data sets, there may be microarray data sets of higher complexity, for instance with more than two levels of a factor or more than two groups. When this case arises, an analysis of variance (ANOVA) test would be more appropriate. A one-way ANOVA allows the comparison of multiple groups. The null hypothesis in an ANOVA model states that all group means are the same. When performing an ANOVA, the sums of the squared deviations from the means are calculated between groups and within groups, and an F-test is computed. The formula for the F-test is,

$$F = \frac{s_B^2}{s_W^2} \qquad [1.7]$$

where s_B^2 is the between group variance, and s_W^2 is the within group variance (Ott and Longnecker, 2001). The value computed from this ratio is compared to an F-distribution with the appropriate numerator (number of groups) and denominator (number of observations - number of groups) degrees of freedom, and the null hypothesis is rejected. If the p-value is significant, then one or more groups is different from the others (Stekel, 2003). For a 2 group scenario, the F test is equivalent to the t^2 test for an unpaired case. For more complex general linear models with one or multiple class factors and covariates, F tests are used to test factors, and t-tests are used for covariates.

Least squares, likelihood, or Bayesian approaches can be used to implement these or more complex models and obtain parameter estimates (e.g. differences between groups; error variances). These estimates are used to compute the test statistics. In parametric approaches these statistics are compared to theoretical distributions (e.g. t or F in least squares, chi-square or likelihood ratio test) to test the null hypothesis. The null hypothesis in microarray gene expression experiments states that there is no difference in gene expression across groups or conditions, i.e. the gene is not differentially expressed. When analyzing the results obtained from applying a model, an important factor to consider is the selection of a suitable significance threshold. For instance, a t-test will indicate how different is the expression of a gene in two groups, but whether or not the hypothesis is accepted or rejected depends on the significance level. Different significance thresholds can lead to varying numbers of genes that are found to be important to the question at hand (Li and Yang, 2002). Given the sample size and error

variance, the significance threshold tends to be a subjective choice, that relies on the selected type I (false positive) and consequent type II (false negative) error rates. There is a compromise between both error rates. More stringent (lower) type I error rates (lower α value) are associated with higher error II rates. Depending on the objective of the investigation, one error rate may be favored over the other. For example, if the confirmation or further studies of the genes identified based on the α-threshold is costly, then it is imperative to be highly certain about the selected genes, and thus a stringent α-threshold to compare the p-values must be used. In addition to a minimum statistical significance, a minimum fold change between conditions may be required for a gene to be considered differentially expressed. This additional requirement can address biological in addition to statistical considerations.

In microarray experiments hypothesis testing is performed on hundreds or thousands of probes or genes, thus increasing the probability of type I error rate or false positive rate. To overcome a possible rise in false positives by chance, multiple test adjustment techniques such as False Discovery Rate (FDR), Bonferonni adjustment, or permutation tests are used (Allison *et al.*, 2006). False Discovery Rate is the expected proportion of null hypotheses that have been rejected that are actually false positives (type I error). In an FDR adjustment, gene p-values are ranked from smallest to largest, $p_1 < p_2 < ... < p_k < ... < p_m$, where k is the ranking and m is the number of genes. The FDR-adjusted α value for test k is calculated by [k/m]*α (Rodriguez-Zas, 2005a). Therefore, if $p_{(k)} < [k/m]\alpha$, then the treatment effect is significant for the given gene (Rodriguez-Zas, 2005a). If the adjustment of consecutively ranked genes is such that the lower rank gene has a less significant adjusted p-value than the higher rank, then the less

significant p-value is assigned to the higher rank gene. In accordance with this, all genes with ranking < k are significant by default. In a Bonferroni adjustment, p-values calculated by the chosen statistical test are multiplied by the number of tests being performed, e.g. in the case of a gene expression analysis, p-values are multiplied by the number of tested genes (Stekel, 2003), or the threshold significance α is adjusted by dividing its value by the number of genes (Rodriguez-Zas, 2005a). This can be a very stringent method of adjustment with a high false negative rate, and may not be appropriate for microarray experiments. In permutation tests, data are permuted, or shuffled, to create new pseudo-sets. The statistic of interest is evaluated to see if its departure from the null hypothesis is greater in the actual observed data than in a large proportion of statistics in the pseudo-sets (Allison *et al.*, 2006). Permutation test is a non-parametric approach in that there is no need to assume a test statistic distribution, and this is particularly beneficial in small data sets that may depart from the Normal distribution assumption. The permutation within gene addresses the small sample-assumption situation. The permutation across genes addresses the multiple tests situation.

Once genes have been identified as differentially expressed, a Gene Ontology (GO) analysis can be conducted to characterize the molecular functions, biological processes or cellular components of the genes, aiding in the interpretation of the data from a genomic standpoint. Molecular functions are defined as activities that occur at a molecular level, such as catalytic activity, transporter activity or binding (Gene Ontology Consortium, 2000). Biological processes are series of events accomplished by assemblies of molecular functions, such as a cellular physiological process or signal transduction (Gene Ontology Consortium, 2000). Finally, cellular components are

particular areas of the cell, such as the nucleus or ribosome (Gene Ontology Consortium, 2000). Using this tool, probabilities are calculated for specific GO terms using a hypergeometric formula,

$$P(n_i) = \frac{\binom{m_i}{n_i}\binom{M - m_i}{N - n_i}}{\binom{M}{N}} \qquad [1.8]$$

where M is the total number of genes identified (i.e. with significant variation across conditions) from the analysis (the total on the microarray), m_i is the subset of genes belonging to GO term i, N is the set of reference genes in the microarray (or sometimes genome), and n_i is the subset of reference genes pertaining to GO term i (Sun *et al.*, 2006). If the p-value is less than the set threshold (e.g. 0.01), then it is considered a statistically significant finding (Sun *et al.*, 2006). Here multiple adjustments are also necessary. Another tool for interpreting gene expression data is Gene Set Enrichment Analysis (GSEA), which focuses on sets of genes to find groups that share similar biological functions, chromosomal locations, or regulation (Subramanian *et al.*, 2005). The tool was developed to overcome analytical challenges that accompany many gene expression experiments. Subramanian *et al.* (2005) outlined four challenges. The first challenge deals with the output after multiple test adjustments are made. Frequently, no individual genes meet the threshold significance level because biological differences are outweighed by microarray noise, and therefore no genes are available for a functional analysis. The second challenge is the reverse of the first. An analysis can reveal a long list of genes identified as statistically significant, of which no biological theme seems to be prominent. A third challenge is that single-gene analysis can miss important pathway effects, where

sets of genes are found to be synergistically important. Finally, two separate groups studying the same data can find lists of significant genes that show little overlap, largely due to model differences. GSEA allows genes to be grouped into gene sets using prior knowledge, such as pathway information or gene relationships previously published. The GSEA approach ranks all genes analyzed according to estimated log2 fold change from positive to negative. Then, moving through the list, it assigns a constant if the gene pertains to the biological theme of interest or zero if not. The accumulation of these coefficients is compared to a theoretical distribution using the Kolmogorov-Smirnov (K-S) statistic to assess for differential expression of the theme.

1.2.2.4 Classification

The fourth step of the workflow can also include statistical and computational mining approaches to further facilitate the interpretation of results including, supervised classification (e.g. clustering of genes or samples based on gene expression profiles) and unsupervised classification with the goal of developing predictive equations and identify gene biomarkers that can be used to assign samples into groups or conditions. This step may be done on a selected group of genes found to be differentially expressed at a particular minimum significance and fold change, and/or on genes of interest based on prior biological knowledge or previous experiments.

Classification approaches aim at identifying distinctions between groups of observations (e.g. genes, samples, treatments) based on a set of measurements (Stekel, 2003). To predict the class in which an observation belongs to, models are built using, in the case of microarray experiments, gene expression measurements. In unsupervised learning classification approaches like clustering, observations are assigned to groups

without using any knowledge of the true class that the observation pertains to. In supervised learning classification approaches, like K-nearest neighbor, the information of the class to which the observation pertains is used to develop a classification predictive function. Here, the training data set used to develop the predictive equation can be the same or different to the test data set used to evaluate the predictive equation. Common supervised classification methods include k-nearest neighbor, neural networks and support vector machines. K-nearest neighbors (KNN) is the easiest way to classify samples. The class of the known sample(s) closest to the unknown sample will be assigned to as the class of the unknown sample (Stekel, 2003). The distance measurement is based on the expression measurements of the samples. Logistic regression is a generalizable linear model that uses a theoretical framework to obtain estimates and develop a predictive equation that can be used to classify samples. Logistic regression is explained in more detail in section 2.3.2 below. In neural networks, a network (model) is trained on samples of known classes, and then new observations are classified based on the predictive model 'learned' from the training observations (Stekel, 2003). Logistic functions are often used in neural networks that can accommodate for non-linearly separable data. Support vector machines (SVMs) project gene expression data into a higher dimensional space, which is then separated using a linear method (Stekel, 2003). The process allows for non-linear separation of data, a feature shared by neural networks.

An analysis of the relationship between samples, genes or treatments can be conducted using clustering, or unsupervised learning, techniques. Hierarchical clustering is the most common method of clustering, where genes or samples are arranged in a tree

pattern so similar samples are closer together and dissimilar samples are farther apart on the tree. Genes or samples are grouped based on distance or similarity (correlation) The algorithm for K-means clustering consists of six steps: 1) select a number of clusters [k], 2) randomly assign samples to one of k clusters, 3) calculate the centroids of the clusters, 4) calculate the distance between each sample and the centroids of all clusters, 5) if the sample is closest to a cluster different from the one it is in, then move the sample to a new cluster and update the centroids for those clusters, and 6) repeat steps 4 and 5 for all samples (Stekel, 2003). A third clustering method is self-organized maps (SOM). Similar to K-means clustering, the number of clusters must be predefined. The properties of SOM are that 1) clusters relate to each other in a grid setup, 2) the size of the grid is decided before the analysis (i.e. the number of clusters), and 3) genes are assigned to the clusters at random, and therefore different runs of SOM can give rise to varying results (Stekel, 2003). Once samples have been classified and clustered based on related profiles, the results are then validated to ensure that both methods provided the correct results.

1.2.2.5 Validation

The final step on the analytical workflow is to validate the findings of the analysis. For example, some genes that appear to be differentially expressed even after multiple test adjustment may be false positives. Additional data (a confirmatory experiment or more samples) can be analyzed separately to confirm the findings, but this solution may not be feasible. In several studies, the same samples that were analyzed using a microarray platform are analyzed using a different technique, real-time PCR, to confirm the differential expression of selected genes (Loor *et al.*, 2006, 2007; Smith *et al.*, 2005, 2008; Everts *et al.*, 2008). Real-time PCR, also known as quantitative real time PCR

(qPCR), is a technique used for amplifying, as well as quantifying, a targeted DNA molecule. The key aspect of real-time PCR is its quantification of amplified DNA as it amasses in the reaction after each amplification cycle (Kubista *et al.*, 2006). A literature review could also help confirm findings but few studies (or none) that evaluated the same or similar conditions may have been reported. Confirmation of results based on literature review can be considered an informal application of meta-analysis of gene expression information across studies. The overall objective of this proposal is to investigate different approaches to integrate information from multiple studies within a formal and probabilistic framework.

1.2.3 Meta-analysis

Large numbers of microarray experiments have been reported in the literature and in some cases the data are available in public repositories (e.g. NCBI Gene Expression Omnibus (Barrett *et al.*, 2007), EMBL ArrayExpress (Parkinson *et al.*, 2007)). The integration of information from multiple related experiments can increase the precision to estimate parameters (e.g. differences in gene expression across conditions), augment the power to test the hypothesis (e.g. detect truly differentially expressed genes) and identify study-particularities. However, the integration of information across studies is not always straightforward and there are many ways to accomplish this integration. For example, the same type of information (e.g. raw data, estimates) may not be available in all the studies of interest and the studies may have evaluated identical, similar, and different conditions. Meta-analysis is the term used to describe the integration of information across studies, and encompasses many approaches. A review of meta-analysis is presented, and

limitations of the reported applications of meta-analysis to microarray experiments are discussed because these are motivators for the work proposed.

Meta-analysis can be defined as the use of statistical methods to combine results from multiple independent research studies (Egger *et al.*, 1997; Hedges, 1992). The scope and tools of meta-analysis have broadened greatly as many diverse fields have found new applications for its use while requiring particular considerations. Meta-analysis has ventured from its use solely for literature reviews/searches, to its utilization in analyzing data in numerous fields of modern science. Nevertheless, the basic idea of meta-analysis has remained the same from its origins through today. The first scientific meta-analysis was conducted by Karl Pearson in 1904 in an effort to account for the reduced statistical power in enteric fever inoculation studies that had very small sample sizes (Pearson, 1904).

Essentially, meta-analysis is considered 'survey research' that examines different studies (i.e. interview each study) and creates an overall coding scheme to aid in accounting for the study characteristics (e.g. variation) and results (i.e. estimates) while keeping the individual study identities (Lipsey and Wilson, 2001). The information from all studies can be analyzed together using statistical techniques that weight the information from each study by the sample size or standard error of each study. Fixed and random effects meta-analyses are two main approaches to implement meta-analysis. In fixed effects meta-analysis, the effect under study is tested against the within-study variability. The assumption is that the observations across studies can be considered from the same population and thus the differences between studies are only due to within-study variation (Hunter and Schmidt, 2000). Random effects meta-analysis accounts for both

between-study and within-study variability because the assumption is that studies are a random sample from the population of all possible studies (Choi *et al.*, 2003). In the presence of substantial heterogeneity of variance across studies, the between-study variance is higher than the within study. Studies of different sample size receive relatively similar weight in fixed effects meta-analysis, but the weight is inversely proportional to the study variance in the random effects meta-analysis. When there is homogeneity of variance across study, the effect of sample size dominates over between and within study variance, and both fixed and random effects meta-analysis models give similar results. Random effect meta-analysis is therefore more conservative and the tests are less likely to reject the null hypothesis than a fixed effects model (Hunter and Schmidt, 2000). Typically, a Chi-square test is used to test homogeneity of variances. If the test is non-significant, then it cannot reject the null hypothesis of homogeneity. However, this non-significant result does not prove homogeneity, as the Chi-square test has a low power to detect variation in study parameters, unless the number of studies is large (Hunter and Schmidt, 2000).

An alternative to meta-analysis is the mega-trial, where rather than analyzing multiple small studies, one large study is conducted (Shrier *et al.*, 2007). Although a mega-trial would be a solution to the random errors associated with environmental effects when combining many smaller studies, a definite disadvantage is the cost and feasibility associated with performing a very large study. As many pertinent microarray studies have been published, meta-analysis of these published studies constitutes an effective way to reuse the information and gain more precise understanding of study-to-study variation, compared to the meta-trial alternative.

1.2.4 Meta-analysis of microarray studies

The usefulness of meta-analysis in numerous fields including econometrics, psychology and epidemiology has been long and well-established (Hunter and Schmidt, 1990a,b; Schmid *et al.*, 2004). The application of meta-analytical approaches to microarray gene expression data is an area of active development (Rodriguez-Zas *et al.*, 2008; Romualdi *et al.*, 2006). A basic assumption of integrated analysis is that independently generated datasets may share common information despite systematic variations between experiments.

A key issue in meta-analysis is the statistic that will represent the information contributed by each experiment. One main line of approaches to implement meta-analysis of gene expression microarray experiments, is to calculate the effect size (a measure of the magnitude of the difference between conditions) for each gene in the individual experiments, and then integrate these indicators into a single measurement (Hu *et al.*, 2006). The effect size indicators can be fold changes, t-test statistics, Z-scores, and P-values. Ghosh *et al.* (2003) identified genes consistently differentially expressed across four microarray datasets using a weighted average of the t-test statistics with weights equal to the sample size of each dataset. This approach is similar to a random-effects meta-analysis. Another line of approaches centers on combining lists of genes that have been detected to be significantly associated with the conditions tested across experiments. This strategy can also include consideration of the ranking of the genes within study lists, and the pattern of expression.

Some meta-analyses of microarray experiments aim at identifying biomarkers with particular expression patterns across experiments (Fishel *et al.*, 2007; Rhodes *et al.*,

2004; Yang and Sun, 2007). Fishel *et al.* (2007) developed a support vector machine classifier based on 80% of the data from three lung cancer data sets considered together and applied the classifier to the remainder 20%. Rhodes *et al.* (2004) proposed a non-parametric meta-analytical approach, ranking genes based on the number of significant gene lists across experiments in which these genes appeared, and used a permutation test to assess the statistical significance of the consistency across experiments. In the non-parametric approach of Yang and Sun (2007), similarities of ordered gene lists are used to detect significant prognostic marker genes from multiple microarray datasets based on consistent changes of multiple genes. Studies ranking genes based on significance are combining studies based on results, and therefore they do not take into account overall estimates of the magnitude of differential expression for genes (Hong *et al.*, 2006), which can produce biased results. Combining studies based on raw data, and conducting one analysis would aid in correct estimation of differential expression. Other meta-analysis implementations compare the list of top ranking predictive genes using parametric and non-parametric rank aggregation approaches (DeConde *et al.*, 2006; Hong *et al.*, 2006). DeConde *et al.* (2006) converted pair-wise preferences between list elements across studies into a stationary distribution that represents an aggregate ranking in a Markov chain framework. Similarly, Hong *et al.* (2006) developed a tool based on the non-parametric rank product method to detect elements that are consistently highly ranked in a number of lists. Expression values are transformed into ranks, and therefore Hong *et al.* (2006) states that genes can be consistently found as up-regulated or down-regulated across studies regardless of the microarray platform being used (e.g. Affymetrix oligonucleotide microarrays, two-color cDNA microarrays). To address the

consideration of expression profiles by non-parametric rank-based approaches, Zhang and Fenstermacher (2005) proposed the identification of biomarker reporter genes using a rank-sum test statistic and the combination of expression levels across studies using a linear index that is trained and tested across datasets. A similar approach implemented by Schneider *et al.* (2006) confirms the expression signature of selected genes found in one dataset on other datasets. This two-stage approach may result in loss of information across stages and accurate training and validation requires a large number of experiments. Conlon *et al.* (2007) proposed to use an indicator variable for differential expression that is a function of the total number of genes. However, the number of genes can vary between experiments and the indicator did not take into account the sign of the differential expression, only whether the genes are differentially expressed or not.

Rank-based approaches, although simple and helpful in overcoming the consideration of distributional assumptions across experiments or concerns of incomparable expression levels across experiments, could result in substantial loss of power and information because equally ranked genes across experiments may have substantially different statistical significance or pattern of gene expression across conditions. Furthermore, random or mixed-effects meta-analytical approaches can accommodate model assumptions such as heteroscedasticity (heterogeneity of variances) across experiments. Summarizing past methods of applying meta-analysis to microarray studies reveals that many of the combination techniques do not take into account the sign of differential expression, which explains whether the genes are up- or down-regulated. In many cases, the magnitude (sign/direction) of expression is also ignored because the

process of combining studies is based on ranking genes by significance, or simply combining studies based on p-values.

Meta-analysis based on simple comparison of lists of genes differentially expressed within experiment is a useful first step to combine information across experiments (Assou *et al.*, 2004). However, overlap of lists of genes can result in potentially biased conclusions for two reasons. First, genes that may exhibit the same pattern across experiments but do not surpass the minimum threshold within one or multiple experiments may not be detected by this approach. Second, genes that may exhibit differential expression in more than one experiment may not reach differential expression when all the data across experiments is considered because the variation across experiments is higher than the variation within experiment. Model-based meta-analysis, in which the combination of information across studies is done using a model, is an ideal framework to conduct an objective, integrative, and comparative study of multiple related microarray gene expression experiments and help better understand the transcriptome and genomic basis of complex traits. Traditional meta-analytical approaches, which combine estimates or observations across experiments in one single step, offer a comprehensive solution to the simultaneous consideration of multiple experiments. These approaches have a solid and extensive theoretical framework, jointly model expression patterns and associated measures of uncertainty, and can be used to detect differentially expressed genes across experiments or to identify biomarkers associated with the conditions of interest. Few applications of model-based meta-analysis to the expression of thousands of genes have been reported, and none have compared the performance of alternative approaches for different gene expression scenarios or across

experiments with different levels of relationship. More research into the application of meta-analysis to microarray experiments remains necessary to help in identifying the best approach for the varying gene expression scenarios.

The proposed work centers on the examination of alternative model-based approaches to integrate information from multiple microarray experiments. In the first research aim, the relative advantages of mixed-effects meta-analysis of test statistics and normalized gene expression measurements under different circumstances are evaluated. The second research aim addresses the meta-analysis of conditions, instead of gene expression measurements, as a strategy to develop reliable predictors of condition based on gene expression information from multiple experiments. In the third research aim, alternative ways to combine gene expression statistics across experiments that do not include the same conditions will be investigated. The following section reviews classification approaches that can be used to develop predictors of condition based on gene expression information and to identify gene signatures or biomarkers of conditions.

1.2.5 Classification approaches

Classification of samples with unknown affiliation into groups using predictive functions trained on gene expression data from samples with known affiliation can aid in the diagnosis and prognosis of numerous diseases and disorders (Ghosh *et al.*, 2003, Rhodes *et al.*, 2004). Classification functions are typically trained on a dataset and tested on an independent dataset to accurately assess the performance of the classification function. The training and testing of classification functions can follow a cross-validation approach such that the dataset is divided into data subsets, the classification function is trained in all-minus-one subsets and tested on the remaining subset and this strategy is

repeated for all test data subsets. Several approaches to obtain classification functions or rules with complementary characteristics are available, including discriminant analysis, logistic regression and artificial neural networks. The performance of the classification function on the test dataset(s) is evaluated by comparing the predicted and known affiliation of the observations to the conditions or groups. From this comparison, the number of true positives (correct assignment to one group, or TP), true negatives (i.e. correct assignment to the other group, or TN), false positives (incorrect assignment to one group, or FP) and false negatives (incorrect assignment to the other group, or FN) are computed. More comprehensive measurements of model performance that are functions of the number of true and false positives and negatives, including sensitivity, specificity, correct classification rate, and area under the Receiver Operation Curve, can be computed. **Table 1.2** summarizes the different criteria that can be used to evaluate classification functions based on the comparison between two classes or groups (A and B) of observed and predicted outcomes. For example, A can represent one behavioral maturation group in honey bees (e.g. nurse) or embryo developmental stage (e.g. blastocyst) and B can represent another behavioral maturation group (e.g. forager) or developmental stage (e.g. 25-day embryo) group. An observation pertaining to group A and predicted to be in group A is a true positive if A is considered the positive result and B a negative result. Disagreement between observed and predicted classes constitutes false results.

Table 1.2. Relationships among observed and predicted classification of outcomes into two classes (A and B), and related indicators.

		Observed condition	
		A	**B**
Predicted condition	**A**	True Positive (TP)	False Positive (FP) Type I Error
	B	False Negative (FN) Type II Error	True Negative (TN)

Indicators	**Sensitivity**	= TP / (TP + FN)
	Specificity	= TN / (TN + FP)
	False Positive Rate (α)	= 1 - specificity
	False Negative Rate (β)	= 1 - sensitivity

Sensitivity and specificity are two criteria that can further be used to analyze the performance of the model that was fitted to the data. Sensitivity is the fraction of true positives among all true positives and false negatives. Specificity is the fraction of true negatives among all true negatives and false positives. Because sensitivity addresses one type of observations (e.g. A) and specificity the other (e.g. B), in some cases it may be necessary to settle for a lower specificity in order to achieve very high sensitivity, or vice versa. A receiver operator characteristic (ROC) curve can be used to select the targeted levels of each of these criteria, depending on the goals of the experiment. A ROC curve reveals the performance of the model across a range of specificities and thresholds (Hastie *et al.*, 2001). To construct a ROC curve, sensitivity values are plotted on the y-axis, and 1-specificity values on the x-axis. The graph ultimately reveals the sensitivity rate as specificity increases. In interpreting the graph, the area under the ROC curve (AUC) is calculated. The AUC value is a function of the predicting ability of the model. If AUC is close to 1, the model predicts well, but if AUC is close to 0, the predicting ability is poor. A final calculation can be made to assess the overall prediction model. Correct classification rate (CCR) is the fraction of true predictions among *all* predictions

(TP + TN + FP + FN), which is also known as accuracy. If the accuracy, or CCR, is calculated as 100%, then all observations in the model were correctly classified into the corresponding groups based on the expression of one or multiple genes. All criteria above can be useful to ensure that analyses are run correctly and that the results are accurate.

1.2.5.1 Discriminant analysis

Among the approaches to classify observations into categories or classes, discriminant analysis is considered the reverse of a one-way multivariate analysis of variance (MANOVA). The independent variables from MANOVA become the categories of the dependent variable for discriminant analysis. Likewise, the dependent variables from MANOVA become the predictors for discriminant analysis (Press and Wilson, 1978). Discriminant analysis is commonly used when only two categories exist in the dependent variable, but it can be used for multi-level categories as well. In discriminant analysis, gene expression coordinates are estimated with the maximum distance between groups, and the minimum distance within groups. Coefficients are calculated, which represent the center of the group for each gene (μ). The class of a new sample is predicted with the coordinate closest to the gene expression profile of the new sample.

1.2.5.2 Logistic regression

Logistic regression models are a flexible approach to describe a binary response variable (e.g. affiliation of a sample to a condition or group, like development stage, embryo type, and health status) with numerous explanatory or predictor variables, including the expression levels of multiple genes and other factors such as technical

sources of variation (Liao and Chin, 2007). Response variables with more than two levels can be described similarly using a multinomial logistic regression model (Zhu and Hastie, 2004).

A basic fixed effects logistic regression model is:

$$\log\left(\frac{\pi}{(1-\pi)}\right) = \beta_0 + \boldsymbol{\beta_1}\mathbf{x} \qquad [1.9]$$

where β_0 is the intercept parameter, β_1 is the vector of solution parameters, x is the vector of explanatory variables (e.g. expression values for multiple genes), π is the response probability being modeled (probability of affiliation to condition or group 1 or $\pi = \Pr(Y = 1 \mid x)$), and $\log\left(\frac{\pi}{(1-\pi)}\right)$ is the log (natural base) of the odds ratio or ratio that the probability of the sample pertains to group 1 relative to the probability of pertaining to group 0.

The dimensionality typical of microarray datasets poses a challenge to the implementation of logistic regression models because the potential number of explanatory variables, or genes, far exceeds the number of samples. Two major lines of approaches to constrain the number of explanatory variables can be used: gene (or variable) selection, and multidimensional reduction approaches that combine the information from multiple genes into fewer composite variables. The reduced number of explanatory variables (genes or combinations of genes) can then be used as explanatory variables in a logistic regression model.

Forward, backward and stepwise are three variable selection methods. These iterative methods allow for only a few genes strongly associated with behavior to be selected from a group of genes, and fitted in the model for analysis. In forward selection,

the model starts with no explanatory variables. P-values are computed for each variable not in the model, and if the most significant surpasses a user threshold, this variable is added to the model. The significance of each variable not in the model, given the variable(s) in the model, is computed and again the next significant variable is entered if it surpasses the user-defined threshold. The step of analyzing the remaining variables is repeated until no variables meet the significance threshold for entry into the model. When a variable is added to the model using forward selection, it is never removed. In backward selection, the model starts including all possible explanatory variables. At each iteration, the least significant variable that does not meet the specified threshold is removed from the model, and is never re-added. The process is repeated until no remaining variables meet the criteria. Stepwise selection can be considered a combination of both forward and backward methods. Variables previously entered into the model can be removed in another iteration or step, and those removed previously have a chance to be re-entered into the model in subsequent steps. The process continues until no variables can be added, or if the last variable added is also the last one removed in the next step. Other selection methods such as the combination of the previous methods can be used to reduce the input parameters in a model. A Wald test is used to examine the significance of the explanatory variables. The Wald statistic is calculated using the formula,

$$\frac{\left(\hat{\theta}-\theta_0\right)^2}{\text{var}(\hat{\theta})} \qquad [1.10]$$

where $\hat{\theta}$ is the maximum likelihood estimate of the parameter (solution for the expression of each gene), θ_0 is the proposed value, and var($\hat{\theta}$) is the variance of the

maximum likelihood estimate. This statistic is then compared to a Chi-squared distribution.

Li and Yang (2002) re-analyzed a data set in which the original publication used 50 out of more than 7,000 gene expressions from 38 samples to discriminate between samples. Li and Yang (2002) found, using logistic regression models, that upper limit for the amount of classification genes of 38 samples was 25. The calculations were based on information criteria obtained from the model results. Li and Yang (2002) warn that the addition of too many genes "may not improve the data-fitting performance enough to compensate for the increase of model complexity." Therefore, when applying classification analysis to studies of small sample sizes, the process of adding additional genes to the model must be carefully considered.

Multidimensional reduction techniques identify combinations of the original gene expression variables that account for much of the variation in the data. Two common methods used to reduce the dimensional space are Principal Component Analysis (PCA) and Partial Least Squares (PLS). Both PCA and PLS are capable of projecting a high dimensional space onto a lower dimensional space, while still capturing as much of the variability from the original data as possible. In an experiment, there are p components, or genes, which can be represented graphically with p axes. Using PCA, the number of components can be reduced to k. Eigenvalues are calculated for each of the k components, or axes, which are related to the amount of variation explained by the axis (Johnson and Wichern, 2007). The sum of the eigenvalues equals the number of components. These eigenvalues are expressed as a percentage of the total variation. For example, if an experiment produces eigenvalues (in percentages) of 63%, 33%, and 4%

for 3 axes, the interpretation would be that the first axis explained 2/3 of the variation in the entire dataset, and the second axis explained a majority of the remaining variation. Therefore the first two principal components may be the ones worth interpreting. How PCA and PLS differ, in general, is that PCA produces a weight matrix reflecting the covariance structure between the predictor variables ("Partial Least Squares", 2008), whereas PLS produces a weight matrix reflecting the covariance structure between the predictor and response variables (Abdi, 2007). The PLS approach directly considers information involving group separation when redefining the space (Liu and Rayens, 2007). Because PCA develops principal component functions without consideration of the affiliation of samples to groups, the resulting functions are not necessarily highly correlated with the conditions under experiment (Liu and Rayens, 2007). Nguyen and Rocke (2002) analyzed four microarray datasets consisting of a total of nine different tumor types for classification. Both PCA and PLS were tested as possible dimensional reduction methods, and PLS was identified as the preferred method. For classification of samples, both logistic discrimination and quadratic discriminant analysis were found to effectively classify normal and tumor samples. Results from the dimensional reduction of microarrays can then be used as predictors to classify samples using such methods as logistic regression.

There is no single method to select explanatory or predictor variables that guarantees the most accurate prediction across datasets, and thus multiple variable selection approaches must be considered. Extrapolation of results from meta-analysis of gene expression measurements suggests that training of classification methods based on information from multiple studies will result in more accurate classification. This

hypothesis is the basis for the proposed development of logistic classification approaches in a meta-analysis framework.

1.2.5.3 Artificial neural networks

An artificial neural network is a data mining technique similar to the biological signaling between neurons in the central nervous system (Wu and McLarty, 2000). In biological neural networks, a dendrite sends signals through synapses to an axon, which in turn produces output. Artificial neural networks have nodes which function as input parameters, or dendrites, sending signals to output nodes, or axons, which produce the output variable. Artificial neural network is a classification method in which input data are used to determine possible classes. For instance, input data may be known genes and their relationships to each other, and the output would be the classification of the unknown sample. If a specific input parameter is related to the sample of interest, an output signal would result, suggesting the proper classification of the sample.

1.2.5.4 Classification considerations

Application of classification approaches to develop predictive equations that assign samples to groups or conditions based on gene expression data from individual experiments have been reported (Nguyen and Rocke, 2002; Zhu and Hastie, 2004). In this proposal, a logistic regression approach to classify samples using information from multiple experiments within a meta-analysis framework will be investigated. This meta-logistic approach will be applied to the classification of honey bees into two groups that correspond to different maturation stages or groups, one-day old nurse (0-12 hours after eclosion from pupa) and forager based on gene expression data from multiple experiments. A description of the honey bee maturation process and review of gene

expression microarray experiments that have helped to characterize the transcriptome of honey bees is hereby presented.

1.2.6 Honey bee maturation

Honey bees are insects classified under the *Apis* Genus, and research suggests that honey bee species diverged about 8-11 million years ago, migrating to the North and West of Eurasia naturally and via human transport (Arias and Sheppard, 2005). Approximately 10 honey bee species that differ on size, nesting habit, location and worker "tempo" have been identified (Arias and Sheppard, 2005; Sen Sarma *et al.*, 2007). The Western honey bee, *Apis mellifera,* is the most familiar and widely studied species, with a sequenced and published genome (The Honeybee Genome Sequencing Consortium *et al.*, 2006).

Honey bees are notable for their production of honey, beeswax and royal jelly. Along with a few avian species, they are essential in their role as pollinators of numerous agricultural crops and natural flora. *Apis mellifera,* in particular, has been used for commercial pollination of crops and other flora valuing billions of dollars (Check, 2006). Despite having brains five orders of magnitude smaller than that of a human, honey bees demonstrate mesmerizing social skills and have been used as one of the model organisms to understanding social maturation (Robinson *et al.*, 2005).

Adult worker honey bees display a complex pattern of behavioral maturation that involves working in the hive when young and foraging when older. This progression gives rise to an age-related division of labor that contributes to their success as highly social insects (Robinson, 2002; Sen Sarma *et al.*, 2007). On average, honey bees have a short lifespan of 6 to 7 weeks and based on the maturation process, worker honey bees

can be classified by their behavior as either nurse or forager honey bees. For the first 2 to 3 weeks after eclosion, nurse honey bees work inside the hive receiving nectar and pollen from foragers, feeding larvae, queen and drones, cleaning, building and repairing the hive, and in ventilation of the hive. During the following 4 to 7 weeks, forager honey bees work mainly outside of the hive collecting and returning nectar and pollen to the hive, while also defending the hive (Whitfield *et al.*, 2003). Honey bees are renowned for their symbolic dance language by which successful foragers communicate the location of attractive food sources to their nest mates (Sen Sarma *et al.*, 2007). Workers also possess the unique ability to respond to the needs of their colony by accelerating, delaying or reversing their maturation from nurse to forager depending on the age distribution of the colony (Whitfield *et al.*, 2006).

A honey bee colony typically consists of one reproducing female called the queen, a few thousand drones (or males), and tens of thousands of semi-sterile female workers (Grozinger *et al.*, 2007). **Figure 1.4** illustrates the differences between a worker, drone and queen honey bee, relative to size. The remainder of the colony consists of immature forms of honey bees called the brood - eggs, larvae and pupae.



Figure 1.4. Morphologies of honey bees.
From left: worker, drone and queen bee (Huang, 2007).

As the behavior of an organism is typified by its actions and responses to their surroundings, behavioral differences are largely associated with age and environmental interaction. Previous studies have shown the transition from hive to forager honey bees to be associated with changes in endocrine activity, metabolism, circadian clock activity, brain structure, neurochemistry, and gene expression (Whitfield *et al.*, 2006). Gene expression differences have been found to have strong associations with caste and species (Grozinger *et al.*, 2007; Sen Sarma *et al.*, 2007). In Sen Sarma *et al.* (2007), species showed differences in dance language, habitat, nesting habit, body size and worker "tempo".

Recently, honey bees have been the "buzz" of many news stories, as Colony Collapse Disorder (CCD) has become a widespread problem in North America. The phenomenon CCD defines the mysterious disappearance of worker honey bees from honey bee hives, leading to the death of many honey bee colonies (National Honey Board, 2008). While disappearances have occurred in the past, the drastic rise in disappearances in late 2006 (Gill, 2007) has caused quite a stir, not only among apiculturists and honey bee researchers, but among the farmers and agriculturists whose crops depend on the pollination efforts of honey bees. Reports of additional CCD cases overseas add to the significance of the problem. Many proposed reasons for the disorder have been environmental stresses (pesticides), malnutrition, pathogens, and mites (Kaplan, 2008). Research on the cause behind CCD is ongoing, and hopefully solutions are in the near future.

1.2.7 Transcriptome profiling of honey bee maturation

The honey bee is a well-established model organism to study the genomic architecture of physiological, neurological, and behavioral maturation (Robinson, 2002). Gene expression experiments targeting specific honey bee genotypes (e.g. strains or subspecies), time points, and environmental conditions (e.g. host colony composition) have demonstrated that behavioral maturation in honey bees is associated with simultaneous changes in expression of thousands of genes in the brain (Cash *et al.*, 2005; Rodriguez-Zas *et al.*, 2006; Sen Sarma *et al.*, 2007; Whitfield *et al.*, 2003; Whitfield *et al.*, 2006). A study by Whitfield *et al.* (2003) showed 2,670 out of 6,878 genes in 72 microarrays to be associated with the age-related transition of honey bees from the hive to forager stage. A leave-one-out cross validation of the top 50 genes predicted the behavioral class correctly in 95% of the honey bees. Whitfield *et al.* (2003) identified this class of predictor genes to be successful in discriminating between classes in unknown samples. Whitfield *et al.* (2006) identified over 85% of approximately 5,500 with brain differences, and PCA of these genes found age, behavior, genotype, environment, and experience to distinctly influence these differences. These genes differed due to at least one factor (e.g. ontogeny, subspecies, colony, and interactions between factors), and 25% differed due to more than one factor (Whitfield *et al.*, 2006). The analyses identified candidate genes for the regulation of behavioral maturation, which are affected by both hereditary and environmental factors (Whitfield *et al.*, 2006).

Sen Sarma *et al.* (2007) revealed 1,772 out of 4,432 genes to have significant differences in brain gene expression between foragers and one-day-old honey bees in four species, however only 218 genes exhibited differences in at least two of the four species.

Sen Sarma *et al.* (2007) used principal component analysis (PCA) to show clear differences between the four species analyzed. A notable finding was that the four species studied were comparable on the *Apis mellifera* microarray that was used, which was in keeping with the understanding of the *Apis* phylogenetic makeup (Sen Sarma *et al.*, 2007). An enrichment analysis identified several categories that may be involved in ecological differences, including genes involved in defense response, metabolism, circadian rhythms and response to biotic stimulus (Sen Sarma *et al.*, 2007).

Denison and Raymond-Delpech (2008) presented an extensive review of changes in gene expression that accompany the transition to foraging. They mention three genes associated with behavioral changes in honey bees: *foraging*, *malvolio* and *vitellogenin*. The study of these three genes has revealed that increased sensitivities to stimuli, such as light and sucrose, are key factors in the transition from nurse to forager (Dennison and Raymond-Delpech, 2008). *Amfor*, a homologue of the foraging gene, is expressed highly in the brain, and contributes to a change in sensory processing (Denison and Raymond-Delpech, 2008). As young bees tend to avoid light and dwell in the dark inner parts of the hive, foragers fly to well-lit places during foraging, and remain around the entrance of the hive (Fry and Wehner, 2002; Southwick and Moritz, 1987). The *malvolio* gene was found to have an increased response to sucrose in fruit flies (Rodrigues *et al.*, 1995). A higher responsiveness to sucrose in foragers indicates the age at which the honey bees start to forage, as well as whether the honey bee collects pollen or nectar (Pankiw and Page, 1999). Honey bees collecting pollen have greater sensitivity to sucrose than bees collecting nectar (Pankiw and Page, 1999). The *vitellogenin* gene has been found to be associated with the regulation of foraging behavior as it keeps honey bees in the nursing

phase (Nelson *et al.*, 2007). Nelson *et al.* (2007) demonstrated that silencing this gene accelerated the transition to foraging, and Amdam *et al.* (2006) revealed that down-regulation of *vitellogenin* increased sucrose responsiveness, suggesting a sudden transition of the honey bees into foraging behavior. With the recent sequencing of the honey bee genome, further studies will help to uncover potentially additional genes, or possibly gene networks, associated with the transitioning phase of honey bees from nurse to forager. The function and biological processes of other honey bee genes can be inferred from the extensive research on fruit flies.

The analysis of individual microarray experiments has already offered insights into the changes in gene expression across maturation stages in the honey bee. The data sharing policy of the Robinson lab at the University of Illinois has resulted in the availability of multiple microarray data sets that compared gene expression at the same maturation stages on the same or similar conditions. In the first two research aims of this proposal, this unique data resource was used to develop and evaluate meta-analytical approaches that can support precise characterization of gene expression patterns and accurate classification of samples based on consistent gene biomarkers. Multi-fold outcomes from the first two research aims are expected. First, profiling and classification meta-analyses will offer additional understanding of the gene expression patterns related to maturation in honey bees. Second, the comparison and evaluation of the performance of the multiple meta-analysis approaches considered across conditions will offer a foundation for effective application of meta-analytical approaches. Lastly, evaluation of the methodologies across genes and datasets will support the effective investigation of statistical methodologies to combine information from experiments of related, yet

different conditions proposed in the third research aim. For the last aim, microarray data profiling the expression of genes in cattle embryos obtained using different reproductive technologies, at different developmental stages, and from placentomes were provided by the Lewin lab at the University of Illinois, the Yang and Tian labs at the University of Connecticut, and the Hue and Renard labs at INRA, Jouy en Josas, France. For this meta-analysis, the difference in the conditions studied across experiments required the consideration of additional meta-analyses techniques.

1.2.8 Embryo development in cattle

Mammalian embryos are created upon fusion of a sperm cell to a mature oocyte (**Figure 1.5**) in the reproductive tract of a maternal source in a fertilization step (Van Soom and de Kruif, 1998). Both natural and artificial insemination (AI) fall into this category as a means of creating embryos. However, since the introduction and development of assisted reproductive technologies, new fertilization methods have been used to create embryos, allowing for the practice of selectively producing individuals with elite genotypes and phenotypes. These methods include in vitro fertilization (IVF) and somatic cell nuclear transfer cloning (SCNT, or NT).

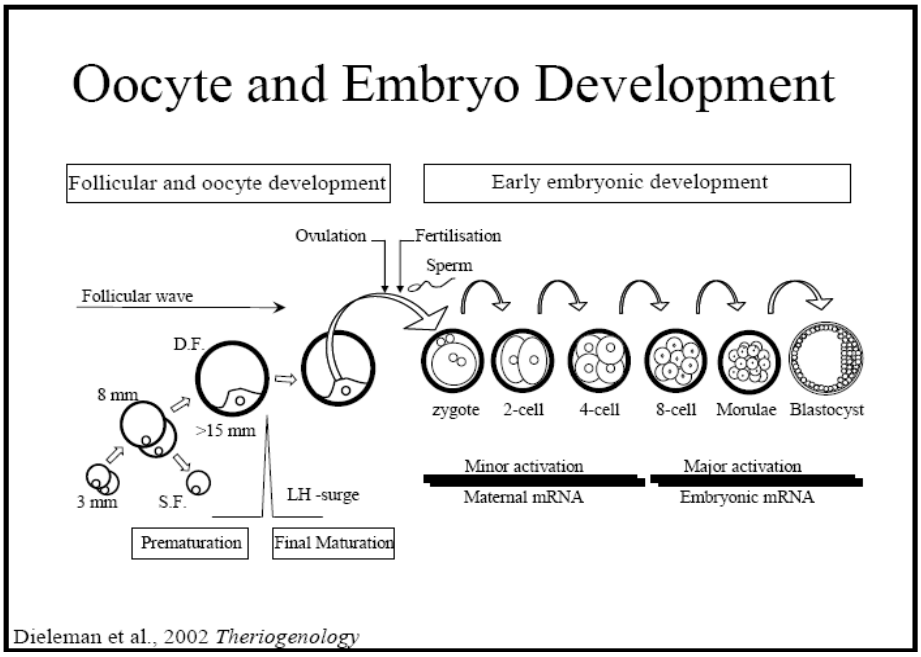


Figure 1.5. Diagram of oocyte and embryo development.

In vitro fertilization (IVF) is an "in laboratory" fertilization method. The process includes oocyte retrieval and insemination, culturing of the newly formed embryo, and transferring of the embryo into the uterus (Bing and Ouellette, 2009). Somatic cell nuclear transfer, commonly known as cloning, occurs when the nucleus of a donor somatic cell is transferred into an enucleated and unfertilized recipient oocyte (oocyte without a nucleus), and a new, genetically identical individual is generated (Tian *et al.*, 2003). Cloning presents numerous advantages not only in the world of research, but also in preserving endangered species, increasing the number of elite producers in a livestock herd, and possible uses of clones in xeno-transplantation (Tian *et al.*, 2003). **Figure 1.6** briefly outlines the steps of SCNT, where the genome of the cells are manipulated in vitro and used to produce transgenic animals.

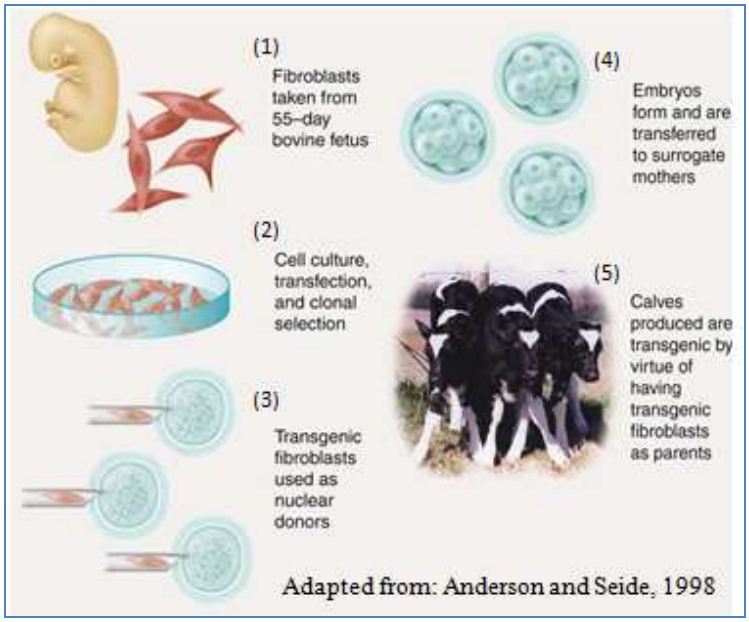


Figure 1.6. Steps in somatic cell nuclear transfer cloning.

The process outlines fibroblast cells taken from the donor individual and maintained in a lab, where the nucleus is removed and fused with an empty egg. Upon fusion, an embryo is created and implanted into a surrogate mother that will produce a cloned animal born with DNA identical to the donor individual. In addition to fibroblast cells, other cell types can also be used for the cloning process, including mammary cells, fetal muscle tissue, and fetal or neonatal skin (Anderson and Seide, 1998). Successful clones have been created in sheep, cattle, goats, horses, mice, rats, dogs, and several other mammalian species (Palmieri *et al.*, 2008). However, disadvantages to cloning exist as well. Currently, less than 5% of embryos produced using NT technology survive to full term with normal phenotypes (Heyman *et al.*, 2002). Numerous abnormalities have been documented in fetuses or calves produced using NT as well as IVF embryo transfer methods. These include low pregnancy rates, increased abortion rates, oversized calves, musculoskeletal deformities, respiratory problems, and placental abnormalities of the maternal source (Chavatte-Palmer *et al.*, 2002; Farin and Farin, 1995; Sinclair *et al.*,

1995). Collectively, these abnormalities describe what is commonly known today as large offspring syndrome (LOS). Abnormalities in embryos produced regardless of reproductive technology can occur as well, such as deficiency in uridine monophosphate synthase (DUMPS), which can lead to embryonic mortality in fetuses homozygous recessive for the condition (Shanks *et al.*, 1992) and decreased activity of uridine monophosphate synthase in fetuses heterozygous for the condition (Shanks and Robinson, 1989). Understanding the underlying mechanisms and genes involved in embryo development are key to improving herd economics through increased pregnancy and embryo and fetal survival rates.

1.2.9 Gene expression profiling of bovine embryo development

Hundreds of genes influence embryo development in human and reproduction models including mouse and cattle (Beyhan *et al.*, 2007; Humpherys *et al.*, 2002; Kues *et al.*, 2008; Somers *et al.*, 2006). Microarray gene expression studies have profiled thousands of genes in cattle embryos from different conditions such as tissue sources (embryo or extra-embryonic), developmental stages and reproductive technologies used to create the embryo (Everts *et al.*, 2007a; Everts *et al.*, 2007b; Everts *et al.*, 2008; Pfister-Genskow *et al.*, 2005; Smith *et al.*, 2005; Smith *et al.*, 2009; Somers *et al.*, 2006). These individual studies typically focus on a particular source or developmental stage or technology. The limited overlap among the list of genes generated from these individual microarray studies has hindered the identification of genes and gene pathways that are critical for successful embryo development. One factor contributing to the low consistency of results across studies is the limited size of some studies, consequence of

the extensive resources required to obtain cattle embryo and extra-embryonic samples and the cost of microarray technology. Another factor challenging the understanding the pathways critical to embryo development is the dynamic nature of these pathways due to capability of the embryo-maternal system to regulate gene expression to compensate for different experimental conditions between studies. Thus, it is not only of interest to identify genes differentially expressed across all conditions, but also genes differentially expressed at particular conditions in order to gain a comprehensive understanding of the transcriptome associated with the development of full-term calves.

In addition to differences in gene expression across developmental stages and sample sources, microarray studies have aimed to detect differences between different reproductive technologies because these findings can shed insights into the low efficiency of some technologies. The NT technology has substantially lower efficiency (measured in the number of transferred embryos that develop to full-term) than AI, the standard reproductive technique in cattle to produce healthy calves (Dinnyes *et al.*, 2008; Edwards *et al.*, 2003; Everts *et al.*, 2007a; Everts *et al.*, 2007b). Various studies have reported differential gene expression between samples from different reproductive technologies at specific developmental stages or sources (Everts *et al.*, 2007a; Everts *et al.*, 2007b; Everts *et al.*, 2008; Marjani *et al.*, 2009; Smith *et al.*, 2007).

CHAPTER 2

META-ANALYSIS OF GENOME-WIDE EXPRESSION PATTERNS ASSOCIATED WITH BEHAVIORAL MATURATION IN HONEY BEES

Adams HA, Southey BR, Robinson GE, Rodriguez-Zas SL
BMC Genomics 2008, **9**:503

2.1 Abstract

Background The information from multiple microarray experiments can be integrated in an objective manner *via* meta-analysis. However, multiple meta-analysis approaches are available and their relative strengths have not been directly compared using experimental data in the context of different gene expression scenarios and studies with different degrees of relationship. This study investigates the complementary advantages of meta-analysis approaches to integrate information across studies and further mine the transcriptome for genes that are associated with complex processes such as behavioral maturation in honey bees. Behavioral maturation and division of labor in honey bees are related to changes in the expression of hundreds of genes in the brain. The information from various microarray studies comparing the expression of genes at different maturation stages in honey bee brains was integrated using complementary meta-analysis approaches.

Results Comparison of lists of genes with significant differential expression across studies failed to identify genes with consistent patterns of expression that were below the selected significance threshold or identified genes with significant yet

inconsistent patterns. The meta-analytical framework supported the identification of genes with consistent overall expression patterns and eliminated genes that exhibited contradictory expression patterns across studies. Sample-level meta-analysis of normalized gene-expression can detect more differentially expressed genes than the study-level meta-analysis of estimates for genes that were well described by similar model parameter estimates across studies and had small variation across studies. Furthermore, study-level meta-analysis was well suited for genes that exhibit consistent patterns across studies, genes that had substantial variation across studies, and genes that did not conform to the assumptions of the sample-level meta-analysis. Meta-analyses confirmed previously reported genes and helped identify genes (e.g. *Tomosyn, Chitinase 5, Adar, Innexin 2, Transferrin 1, Sick, Oatp26F*) and Gene Ontology categories (e.g. purine nucleotide binding) not previously associated with maturation in honey bees.

Conclusions This study demonstrated that a combination of meta-analytical approaches best addresses the highly dimensional nature of genome-wide microarray studies. As expected, the integration of gene expression information from microarray studies using meta-analysis enhanced the characterization of the transcriptome of complex biological processes.

2.2 Background and Motivation

The integration of information from studies can increase the precision to estimate parameters (e.g. differences in gene expression across conditions and studies), augment the power to test the hypothesis (e.g. detect truly differentially expressed genes), and

identify study-particularities (study-specific properties). However, integration is not always straightforward, and there are many ways to combine studies.

Many types of meta-analytical approaches to integrate information from multiple microarray studies have been reported and these approaches have strengths and limitations. The basic approach is to compare lists of genes that exhibit differential expression (as indicated by the P-value and/or fold change). Examples of this approach can be found in Assou *et al.* (2004), Pancaldi *et al.* (2010), DeConde *et al.* (2006), and Hong *et al.* (2006). This approach has the advantage of identifying consistently highly expressed genes across studies (for ranks based on P-values), and the disadvantage that genes that have similar expression patterns across studies but do not surpass the threshold set within one or more studies are not detected. As opposed to a simple gene-list comparison approach, the Fisher's test can be used to combine study *P*-values (of treatment effects) obtained from the individual studies into a Fisher's χ^2 test statistic. This method is simple to apply and does not require additional analysis, however by using the P-values only, there is no way to estimate the average magnitude of differential expression. Pyne *et al.* (2006), Rhodes *et al.* (2002), and Smith *et al.* (2008) used the Fisher's test to combine microarray experiments. Integration of information across studies based on the combination of significance P-values using Fisher's test lose valuable information in regards to the biological magnitude or fold change and direction or expression profile (over or under-expressed) between the conditions or treatments studied among studies. Multiple ways to integrate measures of the expression patterns between treatments across studies have been proposed. Rhodes *et al.* (2004) and Yang and Sun (2007) use the ranking genes based on significance to combine information across studies.

Rhodes *et al.* (2004) proposed a non-parametric meta-analytical approach, ranking genes based on the number of significant gene lists across experiments in which these genes appeared, and used a permutation test to assess the statistical significance of the consistency across experiments. In the non-parametric approach of Yang and Sun (2007), similarities of ordered gene lists are used to detect significant prognostic marker genes from multiple microarray datasets based on consistent changes of multiple genes. This approach can result in a considerable loss of information because equally ranked genes across studies may have substantially different patterns of expression across conditions.

Ghosh *et al.* (2003) identified genes consistently differentially expressed across four microarray datasets using a weighted average of the t-test statistics with weights equal to the sample size of each dataset (a standardization approach), however, the method misses information on directionality of the genes (up- or down-regulation).

Comparisons of meta-analytical approaches have been conducted, showing benefits and limitations to the different methods being analyzed. Hong and Breitling (2008) compared t-based, ranking and Fisher's meta-analysis methods, and Hu *et al*. (2006) compared methods of P-value and estimate (non-standardized) combinations.

Different types of meta-analysis approaches combine different types of information (e.g. P-values, fold change) across studies. However, most studies only consider one type of information (Assou *et al.*, 2004; Pancaldi *et al.*, 2010; Popovici *et al.*, 2009). This situation hinders the assessment of the performance of the approach per se (e.g. ranked versus observed P-values) separately from the impact of the type of information. Conlon *et al.* (2007) proposed to use an indicator variable for differential expression that is a function of the total number of genes. However, the number of genes

can vary between experiments and the indicator did not take into account the sign of the differential expression, only whether the genes are differentially expressed or not. The comparison of meta-analyses that use different approaches and different information types is particularly critical in microarray studies because either or both, pre-processed (e.g. fold change) and raw gene expression data may be available from public repositories, literature review and personal communications.

The goals of this study are 1) to demonstrate the complementary advantages of the meta-analytical approaches to objectively integrate information across studies and identify genes with consistent expression profiles across studies, and 2) to fully mine information from eight microarray studies that have characterized differences in brain gene expression between one-day-old and forager honey bees. The performance of three complementary approaches (overlap of lists of genes, study- and sample-level meta-analyses) to detect consistent differential expression on a genome-wide level were compared. Functional analysis provided further confirmation of the results from meta-analysis. From a biological perspective, the objective assessment of the degree of agreement between studies can help identify common regulatory genes and pathways that could be responsible for simultaneous (parallel or orthogonal) fluctuation in the expression of some genes and the inalterability in the expression of other genes.

2.3 Materials and Methods

2.3.1 Data sets

Eight microarray gene expression studies from two experiments were available for the meta-analysis. The design of the two experiments allowed for separation of

species into separate studies for analysis. Four studies, hereby denoted Group 1 studies, are described in Sen Sarma *et al.* (2007) and encompass the comparison of one-day-old and forager honey bees from four distinct species; *Apis (A.) mellifera* (AM), *A. cerana* (AC), *A. dorsata* (AD) and, *A. florea* (AF) raised in colonies of the same species. The remainder four gene expression studies, hereby denoted Group 2 studies, are described in Whitfield *et al.* (2006) and Rodriguez-Zas *et al.* (2006). These studies encompassed comparisons of one-day-old and forager *Apis mellifera* (*A. m.*) honey bees from different subspecies. These comparisons included *A. m. mellifera* honey bees raised in an *A. m. mellifera* (MM) host colony, *A. m. mellifera* honey bees raised in an *A. m. ligustica* (ML) host colony, *A. m. ligustica* honey bees raised in an *A. m. ligustica* (LL) host colony, and *A. m. ligustica* honey bees raised in an *A. m. mellifera* (LM) host colony. The honey bee species *A. mellifera* and *A. cerana* are closer to each other than to *A. dorsata* and *A. florea* (Arias and Sheppard, 2005; Raffiudin and Crozier, 2007). Meta-analysis of these studies can improve the detection of genes that are consistently over- (or under-) expressed in forager honey bees relative to one-day-old honey bees, regardless of the genetic make-up of the species or sub-species of the honey bee sampled or colony. Meta-analysis also supports the detection of genes that have unique expression patterns across species or within groups of species.

In all studies, forager honey bees were identified as honey bees that had pollen loads on hind legs. In Group 2 studies, forager honey bees could also have distended abdomens containing nectar or water load, and were 16-day to 17-day-old after adult emergence. The number of observations per study ranged from 12 (studies LL, LM, ML, and MM) to 24 (studies AC, AD, AF), and study AM had 22 observations. Due to

variations in the size of the brain across honey bee species, brain samples were pooled (30 to 60 brain samples per pool) in the Group 1 studies. One-day-old and forager honey bees pertaining to the same colony (three colonies) were hybridized to the same microarray in a direct design with reverse labeling in all Group 1 studies. Direct comparison of one-day-old and forager honey bees was obtained using four microarrays (two microarrays per dye-labeling assignment) per colony for a total of 12 microarrays per study in Group 1. Individual brain samples were used in all Group 2 studies, each study including 20 microarrays in a loop design. Each 20-microarray loop corresponded to one combination of honey bee subspecies and colony (LL, LM, ML, and MM). One of the microarrays in all Group 2 studies included a direct comparison between one-day-old and forager honey bees and an indirect comparison through 4 intermediate maturation stages (Rodriguez-Zas *et al.*, 2006). Brain gene-expression data from all six stages in Group 2 studies was normalized together to improve the adjustment for technical variation (See Section 2.3.2). Only the normalized gene expression observations from one-day-old and forager honey bees were analyzed to make the models and analyses across all studies, regardless of group, comparable.

The expression of genes from individual brains was assessed using the double-spotted *A. mellifera* brain 9K version 3.0 cDNA microarray (Rodriguez-Zas *et al.*, 2006). A total of 5001 cDNA transcripts in the microarray were assigned to 3610 individual genes in the honey bee genome assembly version 2, and approximately 1970 genes have Gene Ontology information. Although an *A. mellifera* microarray was used to study different honey bee species and subspecies (AF, AD, AC, and *A. ligustica*), hybridization efficiency differences were minimal (Sen Sarma *et al.*, 2007; Whitfield *et al.*, 2006).

2.3.2 Data processing and individual-study analysis

Data processing included the removal of spots that were flagged by the scanning software (GenePix Pro 5.0) or did not surpass a minimum threshold of 200, and log2 transformation of the background-subtracted foreground intensities. Log-transformed values were normalized using a global LOWESS transformation (Cui *et al.*, 2003) to remove dye bias within microarray and duplicate spots within gene were averaged. Microarray elements that did not have observations in all samples were removed from analysis to ensure the availability of the maximum possible information to estimate the parameter. A two-stage approach was used to adjust for technical sources of variation (Wolfinger *et al.*, 2001). In the first stage, global dye and microarray effects were removed across all microarray elements or cDNAs transcripts, and in the second-stage individual-study or meta-analyses were implemented.

The individual-study analysis encompassed the second-stage model that described each transcript within study with the fixed effects of dye, maturation stage (one-day-old or forager), and the random effect of honey bee sample. Sample (honey bee) effects were assumed to be identically and independently distributed (iid), and from a Normal distribution with mean zero and a common honey bee variance σ_b^2. This common honey bee variance was due to the fact that the limited number of observations per sample precluded the precise estimation of separate honey bee variances. A microarray effect was not included in the model due to the design of Group 2 studies that included two honey bee samples not present on multiple microarrays. The number of transcripts studied within study ranged from 7734 to 7737. The number of gene expression

measurements available per gene, study, and maturation stage ranged from 6 (Group 2 studies) to 12 (Group 1 studies AC, AF and AM).

Three approaches that combine the information on expression patterns in one-day-old *versus* forager honey bees across studies and within transcript were considered. The approaches were: standard overlap of genes with significant differential expression across studies (detected by the individual-study analyses), study-level meta-analysis of expression contrast estimates (obtained by the individual-study analyses), and sample-level meta-analysis of transcript expression across studies.

2.3.3 Study-level meta-analysis

The study-level meta-analysis approach is pertinent when estimates of expression between conditions of interest (i.e. forager and one-day-old stages), and not the raw or normalized measurements of gene expression intensity, are available. The study-level meta-analysis (or Study) combines summary measurements across studies. The summary measurements are the estimates of the difference in brain gene expression between forager and one-day-old honey bees obtained from the individual-study analyses, standardized by the corresponding standard error. A study-level meta-analysis of non-standardized estimates (N_Study) was implemented to facilitate the comparison of study-level meta-analysis to individual-study and sample-level meta-analysis because the standardized estimates are unit-less, and therefore the results of the N_Study would have the same units as the results from the other analyses. An advantage of the standardized study-level meta-analysis is that the estimates have already been adjusted for technical sources of variation (i.e. dye, microarray) during the individual-study analysis stage.

A hierarchical mixed effects model was used to incorporate all sources of variation associated with the study-level standardized estimates combined in the study-level meta-analysis. For each transcript, the study-specific estimate is described with an overall difference in expression effect (μ), and the random effect of study (s_i):

$$y_i \sim N\left(\mu + s_i, \sigma^2\right) \qquad [2.1]$$

$$\mathbf{s} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

where y_i is the (standardized) estimate of difference in expression (forager *versus* one-day-old) obtained from the analysis of the *ith* study (i = 1 to 8), σ^2 is the error variance, and $\boldsymbol{s}$ is the vector of study effects. Study was assumed to have a Normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$ of dimension k x k, where k is eight, the number of studies. The variance-covariance matrix had a diagonal structure with off-diagonals of zero, and diagonals representing the variance of the estimate of differential expression within study. The study-level meta-analysis model accounted for potential heterogeneity of variance among studies, and this is particularly important when combining information from potentially different subspecies. Meanwhile the standardized estimates of differential expression (y_i) are adjusted for all other sources of variation fitted in the within-study model (i.e. dye and microarray effects), and standardized by the standard error of the estimate. In comparison, Choi *et al.* (2003) considered the meta-analysis of differences between the means of the condition levels standardized by the pooled standard deviation.

2.3.4 Sample-level meta-analysis

Like the individual-study analyses, the sample-level meta-analysis encompassed the second-stage model describing each transcript with the mixed effects model:

$$y_{ijklm} \sim N\left(\mu + d_i + m_j + s_k + b(s)_{kl}, \sigma^2\right)$$

$$\mathbf{s} \sim N(\mathbf{0}, \mathbf{U}) \qquad [2.2]$$

$$b_{kl} \sim iid \;\; N\left(0, \sigma_b^2\right)$$

where y_{ijklm} is the normalized log-transformed brain gene expression corresponding to *mt*h observation, from the *lth* honey bee sample, at the *jth* maturation stage, pertaining to the *kth* study, labeled with the *ith* dye. Study effect was assumed to have a Normal distribution with mean zero and diagonal variance structure ***U*** of dimension k x k with a potentially different variance for each study. Sample effects were assumed to be *iid* and from a Normal distribution with mean zero and a common microarray variance $\sigma^2_{b,}$ because of the limited number of observations per sample. Microarray was not included in the model so the sample-level meta-analysis model would be as similar as possible to the individual-study analysis model. A total of 142 observations were analyzed.

2.3.5 General analysis considerations

Individual-study, study-level, and sample-level meta-analysis estimates were obtained using a restricted maximum-likelihood approach and implemented using the SAS mixed procedure (SAS Institute Inc., NC, USA). An experiment-wise type I error rate $\alpha = 1 \times 10^{-3}$ was used to identify transcripts differentially expressed across maturation stages. This significance threshold is equal to the threshold used by Sen

Sarma *et al.* (2007) and Whitfield *et al.* (2006). This threshold also proved to be a good compromise between the number of differentially expressed genes detected and the type I error rate while allowing to explore the overlap of a substantial number of genes across analyses. The analyses were compared in terms of overall number of transcripts identified as differentially expressed and the direction (sign) of the differential expression (over or under expressed in forager *versus* one-day-old honey bees).

The meta-analyses results of 7734 transcripts allowed the identification and characterization of different gene expression scenarios that benefited from different meta-analytical approaches. These scenarios were further explored using funnel plots that depict the estimate of differential expression between forager and one-day-old honey bees, and the associated confidence interval obtained in the individual-study, study-level, and sample-level meta-analysis. Four transcript cases were of particular statistical and biological relevance. First, transcripts that were not found differentially expressed in any individual study, yet were detected by the study-level or sample-level meta-analyses. Second, transcripts that had differential expression in opposite directions across studies and were not identified in the study-level or sample-level meta-analyses. Third, transcripts that were found differentially expressed in multiple individual studies and the sample-level meta-analysis, but were not detected by the study-level meta-analysis. Fourth, transcripts that were found differentially expressed in multiple individual studies and the study-level meta-analysis, but were not detected by the sample-level meta-analysis. Adequate characterization of these possible scenarios can help design additional targeted experiments to resolve the discrepancies among studies.

Results from the study-level and sample-level meta-analyses were further examined using functional enrichment analysis. Lists of differentially expressed genes were assigned to Gene Ontology (GO) biological processes and molecular functions classes based on fruit fly annotations. Representation of genes in GO classes was evaluated using Fisher´s exact (two-tailed) test and False Discovery Rate multiple test adjustment (Al-Shahrour *et al.*, 2006).

2.4 Results

A total of 7734 transcripts were analyzed and a summary of the number of transcripts with significant (unadjusted raw P-value < $1x10^{-3}$) differential expression obtained from the individual-study, study-level, and sample-level meta-analyses is presented in **Table 2.1**. The significance threshold corresponded to an approximate false discovery rate-adjusted *P*-value < 0.1 in the within study analyses. A breakdown of the results by positive (transcript over-expressed in forager compared to one-day-old honey bees) and negative (transcript over-expressed in one-day-old compared to forager honey bees) significant differential expression is provided in **Table 2.2**. The individual studies correspond to eight microarray datasets that do not share samples or microarrays, and include brain gene-expression measurements from one-day-old and forager honey bees. The eight studies were also divided into Group1 and Group 2, corresponding to the two separate publications (Sen Sarma *et al.*, 2007; Whitfield *et al.*, 2006) in which the studies were first presented. The four studies in Group 1 correspond to four distinct species; *A. mellifera* (AM), *A. cerana* (AC), *A. dorsata* (AD) and, *A. florea* (AF) honey bees raised in colonies of the same species (Sen Sarma *et al.*, 2007). The four studies in Group 2

correspond to *Apis mellifera* (*A. m.*) honey bees from two subspecies raised in two host colonies: *A. m. mellifera* honey bees raised in an *A. m. mellifera* (MM) host colony, *A. m. mellifera* honey bees raised in an *A. m. ligustica* (ML) host colony, *A. m. ligustica* honey bees raised in an *A. m. ligustica* (LL) host colony, and *A. m. ligustica* honey bees raised in an *A. m. mellifera* (LM) host colony. Thus, in addition to individual- and meta-analysis, results from the two sets of studies (Group 1: studies AC, AD, AF, AM, and Group 2: MM, ML, LM, LL) are also described.

2.4.1 Comparison of individual-study and meta-analyses

The number of transcripts with significant differential expression between one-day-old and forager honey bees obtained from the individual study analyses ranged from 65 to 540 (**Table 2.1**). The AD study had the lowest number of differentially expressed transcripts and lowest overlap with any other study. Excluding results from AD, the number of differentially expressed genes present in at least two studies in Group 1 (AC, AF, AM) was on average 55, and slightly lower than the number of genes in at least two studies in Group 2 that ranged between 73 and 86. The percentage of cDNA transcripts differentially expressed and overlapping among the studies ranged from 4.3% to 39.5% (average of 16.1%) relative to the number of cDNAs identified in one of the studies. The highest and lowest overlap of differentially expressed transcripts was found in the comparison of studies AC against AM and AD, respectively (**Table 2.1**). Using the number of differentially expressed transcripts between one-day-old and forager honey bees as an indicator of the similarity between studies, AC, AF, and AM in Group 1 are more similar to each other with an overlap ranging from 53 to 60 transcripts, or

approximately 38% of the significant genes. In Group 2, all studies have a similar overlap in number of significant transcripts, ranging from 57 to 86, or approximately 20% of the significant genes. The number of differentially expressed transcripts found in studies from two different groups ranged from 12 to 38, excluding AD that also had the lowest overlap with studies from Group 2. Across groups of studies, MM, ML, and LM are next in proximity to AM with 38, 37, and 35 genes in common, respectively. The highest overlap of differentially expressed genes among studies from different Groups was found between *A. mellifera* AM and MM studies.

The study-level and sample-level meta-analyses were able to overcome the weak consistency among studies, identifying 32 and 853 transcripts with significant differential expression, respectively (**Table 2.1**). The overlap of transcripts with differential expression between the study-level and subject-level meta-analyses was 68.8% of the minimum number of transcripts identified among both analyses. The overlap of transcripts with significant differential expression between the individual-study analyses and the study-level meta-analysis ranged from 18.8% to 28.1% of the transcripts identified within study, excluding study AD that had an overlap of 9.4%. The overlap of transcripts with significant differential expression between individual studies and the sample-level meta-analysis ranged from 23.3% to 80.9%.

The Venn diagram in **Figure 2.1** depicts the overlap of transcripts identified as differentially expressed by the study-level and sample-level meta-analyses and in at least two individual-study analyses. Of the 853 transcripts detected by the sample-level meta-analysis, 621 transcripts were not detected in at least two individual-study analyses, and 349 transcripts were not detected by any single individual-study analysis. Likewise, of

the 32 transcripts detected by the study-level meta-analysis, 17 were not detected in at least two individual-study analyses. Of the 465 transcripts differentially expressed in the analysis of at least two studies, 236 were detected by meta-analysis, and 229 transcripts could not be confirmed by either the study- or sample-level meta-analyses. The differential expression of 11 transcripts was corroborated in all three cross-study approaches. The number of transcripts differentially expressed in two, three, and more studies was 320, 100, and 45 respectively (**Table 2.2**).

The higher the overlap of transcripts between any individual-study analysis and any meta-analyses, the lower the number of differentially expressed transcripts that was unique to that individual-study analysis. The number and percentage of transcripts detected by any individual-study analysis and sample-level meta-analysis is equal to or higher than the number of transcripts detected by the same individual-study analysis and study-level meta-analysis (**Table 2.1**). In addition, the relative percentage of genes that overlap between individual analyses and sample-level meta-analysis is substantially higher and more variable than the overlap with the study-level meta-analysis in all the Group 1 studies, relative to the Group 2 studies. The analysis of the AC study had the highest relative overlap of differentially expressed transcripts with sample-level meta-analysis among all individual-study analyses (80.9%), but had an intermediate overlap with the study-level meta-analysis results (18.8%) relative to other individual-study analyses. Conversely, the analysis of the ML study had the lowest overlap of transcripts differentially expressed with the sample-level meta-analysis among all individual-study analyses (23.3%), but had an intermediate overlap with the study-level meta-analysis (18.8%) relative to other individual-study analyses. The overlap between the transcripts

detected to be differentially expressed by the individual analyses of Group 2 studies and study-level or sample-level meta-analyses were consistent across studies (**Table 2.1**).

2.4.2 Consideration of gene expression profiles

Consideration of the sign (i.e. up-or down-regulation profile) of the transcripts detected by the individual and meta-analyses provided additional insights into the results (**Table 2.2**). There were no changes in the sign of transcripts detected as differentially expressed in two or more individual-study analyses within group, with the exception of 2 transcripts that had positive signs (over-expression in forager compared to one-day-old honey bees) in the analysis of the AM study, and had negative signs in the analysis of the AF study. In addition, a few transcripts exhibited changes in the expression pattern when studies from both groups were considered simultaneously. Amongst the Group 1 individual-study analyses, the AM study had the highest number of transcripts (14 transcripts) that had a different sign in at least one Group 2 individual-study analysis. The percentage of transcripts differentially expressed with the same sign and overlapping among the studies ranged from 1.8% to 43.2%, relative to the number of transcripts identified in one of the studies. The highest and lowest overlap of differentially expressed transcripts with the same sign was found in the comparison of the analysis of the AC study against the AM and AD studies, respectively (**Table 2.2**). The percentage of transcripts differentially expressed with different signs and overlapping among the studies ranged from 0% to 4.8%, relative to the number of transcripts identified in one of the studies. The analysis of the AD study resulted in the highest percentage of transcripts that had an opposite sign in any other individual-study analysis (4.8%), followed by the

analysis of the AM study. The difference in ranking between absolute and relative counts is due to the fewer number of significant transcripts detected in the analysis of the AD study, compared to the AM study.

All the transcripts that exhibited differential expression in an individual-study analysis *and* study-level meta-analysis had a consistent pattern or sign (**Table 2.2**). Most transcripts that exhibited differential expression in an individual-study analysis *and* sample-level meta-analysis had a consistent pattern or sign. The transcripts that had a different sign in an individual-study analysis compared to the sample-level meta-analysis had the same sign as the meta-analysis in other individual-study analyses. Most analyses, individual-study and meta-study, had similar numbers of differentially expressed transcripts with positive (over-expressed in forager compared to one-day-old honey bees) and negative signs. In addition, no particular sign or pattern of differential expression dominated across analyses. The exception was study AM that had 97 and 213 positively and negatively differentially-expressed transcripts, respectively (**Table 2.2**). Although there were no substantial differences in the number of positive and negative estimates across all analyses considered, the overlap of counts by sign between individual analyses and sample-level meta-analysis did not necessarily reflect the trends observed in the individual analyses. This situation was observed in the LM, ML, and MM studies. For example, the total number of positive and negative results in ML was 263 and 277, respectively, yet the overlap with sample-level meta-analysis was 67 and 49, respectively (**Table 2.2**).

2.4.3 Transcript meta-analysis scenarios

To typify the strength of the different types of approaches to integrate information across studies, results from the individual-study, study-level and subject-level meta-analyses for scenarios of particular statistical and biological relevance were compared. **Figure 2.2** presents the estimates (and 95% confidence intervals) of differential expression between forager and one-day-old honey bees corresponding to four transcript scenarios. **Figure 2.2A** depicts a transcript case where meta-analysis detected differential expression, meanwhile each individual-study analysis failed to detect differential expression. Both the sample-level and study-level meta-analyses detected differential expression for transcript BB170018A20B07, with significant differential expression across maturation stages. However, none of the eight individual-study analyses detected differential expression. This transcript represents the honey bee gene GB10350-PA that is similar to the fruit fly gene (FlyBase ID) FBgn0029997 (GO:0005515, protein binding molecular function), which encodes a protein reported to interact with the *Cyclin K* and muscle *LIM* proteins (Giot *et al.*, 2003). **Figure 2.2B** depicts a transcript (identifier BB170024B10C11) that exhibited differential expression in four individual-study analyses. Neither meta-analysis was able to detect this transcript because there was an inconsistent pattern of differential expression across the studies considered. This transcript represents the honey bee gene GB19001-PA that is similar to the fruit fly gene FBgn0032832 (*sick* or *sickie*), which encodes a protein reported to interact with a *pall (pallbearer)* protein (Giot *et al.*, 2003). **Figure 2.2C** depicts a transcript (identifier BB170024B20H01) that was differentially expressed in all except one individual-study analysis and the sample-level meta-analysis. However, the study-

level meta-analysis was not able to uncover a significant difference. This transcript represents the honey bee gene GB13606-PA, predicted to code for a hypothetical protein. On the other hand, **Figure 2.2D** presents a transcript (identifier BB170004B20H08) that was consistently and differentially expressed in three individual-study analyses and in the study-level meta-analysis, but the sample-level meta-analyses was not able to uncover a significant difference. This transcript represents the honey bee gene GB15917-PA that is similar to the fruit fly gene FBgn0040351 (GO:0005509, calcium ion binding molecular binding), which encodes a protein reported to interact with the *Mephisto/Sickle* and the *KCNQ* potassium channel proteins (Giot *et al.*, 2003).

2.4.4 Genes and Gene Ontology classes

Individual genes and Gene Ontology (Gene Ontology Consortium, 2000) categories detected by the meta-analyses approaches were evaluated. Of the 863 unique transcripts with differential expression detected by either meta-analysis, 351 were new transcripts that would have not been detected in at least one individual-study analysis. Of the 351 transcripts, 347 transcripts were solely detected by the sample-level meta-analysis, two transcripts were solely detected by the study-level meta-analysis (identifiers BB160008A20A01, BB170022B20E11), and two additional transcripts (BB160014A10H01, BB170011B10F10) overlap between the study-level and sample-level meta-analyses. The latter four transcripts have not been assigned to honey bee genes or fruit fly gene orthologs.

The 347 transcripts detected by the sample-level meta-analysis mapped to 125 fruit fly genes with GO information (**Appendix A**). The representation of molecular

functions and biological processes among the 134 genes was analyzed using Fisher's exact test (Al-Shahrour *et al.*, 2006). Genes for two molecular functions, cofactor binding (GO:0048037) and purine nucleotide binding (GO:0017076) had significant over-representation at P-value < 2.14×10^{-3} and P-value < 7.04×10^{-4}, respectively. Biological processes in which transcripts were significantly over-represented were macromolecule metabolic process (GO:0043170), cellular metabolic process (GO:0044237), and primary metabolic process (GO:0044238) with P-values < 2.86×10^{-2}, 4.69×10^{-2} and 6.57×10^{-2}, respectively. Functional analysis of transcripts significant and differentially expressed from the sample-level meta-analysis that overlapped with any other individual-study or study-level meta-analyses identified enrichment of the following GO biological process categories: generation of precursor metabolites and energy (GO:0006091), cellular macromolecule metabolic process (GO:0044260), transport (GO:0006810), cell differentiation (GO:0030154), and system development (GO:0048731).

In terms of individual genes, among the 347 transcripts detected solely by the sample-level meta-analysis, two transcripts merit special attention because they are new to the list of genes associated with one-day-old and forager differences, and because of their known biological role. Transcript BB160024A10A12 similar to the fruit fly gene FBgn0026086 (*Adar*), was over-expressed in forager relative to one-day-old honey bees, and is involved in adult behavior (GO:0030534), adult locomotory behavior (GO:0008344), and response to heat (GO:0001666). Transcript BB170016B10A03, similar to fruit fly gene FBgn0027108 (*Innexin 2*), was also over-expressed in foragers and is involved in olfactory behavior (GO:0042048).

Forty-five genes were identified as differentially expressed in four or more studies, of which 39 appeared in one or both meta-analyses and 12 had GO information (**Table 2.3**). Genes over-expressed in foragers relative to one-day-old honey bees include genes corresponding to FlyBase IDs FBgn0038180, FBgn0051997 and FBgn0003036. Gene FBgn0038180 exhibits *Chitinase* activity and *Chitin* can be found in the exoskeleton of insects. FlyBase ID FBgn0003036, the *Para* gene, functions in male courtship behavior and veined wing generated song production. Conversely, genes under-expressed in foragers relative to one-day-old honey bees include FlyBase IDs FBgn0022355 (*Transferrin 1*, associated with defense response and ion transport activities), FBgn0035423 (associated with translation), FBgn0037146 (oxidoreductase activity), FBgn0038471 (methyltransferase activity), FBgn0038516 (oxidoreductase activity), FBgn0050035 (carbohydrate transmembrane transport activity) and FBgn0030412 (neurotransmitter secretion). Two genes had opposite differential expression patterns (signs) among studies. The *Sick* gene (Flybase ID FBgn0032832), which participates in the defense response to Gram-negative bacterium, was over-expressed in forager relative to one-day-old honey bees in studies LL, ML and MM, yet was under-expressed in forager relative to one-day-old honey bees in the study AD. This gene is represented as a funnel plot in **Figure 2.2B**. The *Oatp26F* gene (Flybase ID FBgn0051634) that participates in organic anion transportation, was over-expressed in forager relative to one-day-old honey bees in studies AC, AF and AM, yet under-expressed in forager relative to one-day-old honey bees in study MM.

2.5 Discussion

2.5.1 Study and sample-level meta-analyses

Figure 2.1 demonstrates the capability of meta-analysis to synergistically integrate consistent information across studies regardless of the significance level within study. Meta-analyses detected 627 differentially expressed transcripts that were not detected by overlap of lists of significant transcripts in at least two individual-study analyses. The failure of meta-analyses to detect 229 transcripts detected in at least two-individual analyses can be linked to contradictory or variability in the information provided by the individual studies. Although the relative overlap between the individual-study and study-level meta-analysis was always lower or equal than the overlap of the individual-study and sample-level meta-analysis, the variability of the overlap across studies further confirms that the advantages of the study- and sample-level meta-analysis depend not only on the transcript but also on the studies being integrated (**Table 2.1**). The overlap in number of differentially expressed transcripts between the individual-study and meta-analyses approximately doubled relative to the values presented in **Table 2.1,** when the statistical significance threshold of the false discovery rate adjusted *P*-values was set to 0.2. The increment of the overlap was higher in the comparisons between individual-study and study-level meta-analysis (on average 2.3 fold) than in the comparisons between individual-study and sample-level meta-analysis (on average 1.9 fold). This slight difference may be due to the low sample size of the individual-study analyses and study-level meta-analyses that were more benefited by a less stringent significant threshold than the sample-level meta-analysis.

The vast majority of the transcripts that exhibited differential expression in an individual-study analysis showed consistent pattern or sign in other studies and meta-analysis (**Table 2.2**). Transcripts with significant yet different expression pattern in one study-level and sample-level meta-analysis were detected in other studies with pattern consistent with that detected in the meta-analysis. For example, transcript BB170030B10G10 was differentially expressed at P-value < $1x10^{-3}$ in the individual-study analyses of the AC and AM studies and was borderline significant in the AD, AF, ML, MM, and sample-level meta-analyses. The sign of the MM result was opposite to all other studies and the sample-level meta-analysis.

Transcripts with an inconsistent direction of differential expression or significance across the different meta-analyses require further study within and across studies. Adequate characterization of these possible scenarios can help design additional targeted experiments aimed at resolving the discrepancies among studies. The sample-level meta-analysis detected more differentially expressed transcripts than the study-level meta-analysis in cases with consistent expression profile across studies that exceeded potential variation across studies. The study-level approach was more appropriate to detect differential expression in transcripts with patterns that were consistent across studies, but less pronounced than the variation across studies. Because study is the experimental unit of the study-level meta-analysis, this approach may not have consistent and precise information to detect differential expression when a limited number of microarray experiments are available.

The objective combination of information implemented in the complementary meta-analytical approaches was able to mine the signal of differential expression of data

in different scenarios. The funnel plots for selected genes illustrated the strength of meta-analyses to accumulate consistent expression patterns across studies and gained power to detect differential expression that otherwise would have been considered non-differentially expressed using simple overlap of gene list within (**Figure 2.2**). The funnel plots also depicted the capability of mixed-effects meta-analysis to accommodate potentially high heterogeneity of variances across studies. The confidence interval associated with each estimate depicted in the funnel plots show that genes with substantial heterogeneity of variances between studies favor the use of mixed-effects models to accommodate for this scenario. A clear demonstration of the complementary advantage of the meta-analytical approaches across gene and study scenarios is provided in **Figure 2.2**. For example, **Figure 2.2A** illustrates the ability of the meta-analyses to combine consistent expression patterns of genes across studies and gain precision of estimates.

2.5.2 Meta- and individual-study analyses

The imperfect overlap of results from the individual- and meta-study analyses corroborates reports that the association between the expression of numerous genes and behavioral maturation is highly sensitive to other genetic or environmental factors (Denison and Raymond-Delpech, 2008). The complementary nature of the individual and meta-study approaches allows the identification of study-independent and study-dependent gene expression patterns. Evaluation of the overlap between individual-study analyses and sample-level meta-analysis provided insights into the multi-study scenarios that benefit the most from meta-analyses. For example, among Group 1 studies, the AM

study had the highest overlap in number of significant transcripts with other studies, followed by AC and AF studies (**Table 2.1**). The same trend was observed in the overlap between these studies and the sample-level meta-analysis. On the other hand, the greater overlap in significant transcripts between Group 2 studies (plus the moderate overlap with Group 1 studies) relative to the overlap among Group 1 studies was not reflected in the overlap between Group 2 studies and the sample-level meta-analysis. An explanation for the seemingly contradictory behavior of sample-level meta-analysis and Group 1 *versus* Group 2 studies is the dimensionality of the overlap across more than two studies. Most of the overlaps between studies within Group 1 and other studies were generally observed only on pairs of studies, whereas most overlaps between studies within Group 2 and other studies were generally observed on more than two studies. Consequently, more transcripts got a "boost" from the consistency on other studies in Group 1 study meta-analyses than in Group 2 study meta-analyses.

Another insight gained from the application of model-based meta-analysis approaches to the eight honey bee studies was the assessment of the variation in brain gene expression across studies, relative to the variation across honey bee genotypes. The reason for the higher overlap of differentially expressed genes among Group 2 studies, relative to the overlap among Group 1 studies, is that the samples pertain to *A. mellifera* subspecies, meanwhile Group 1 studies corresponded to *A. mellifera* and other honey bee species. As expected, the transcripts identified in the analyses of the AM study in Group 1 exhibited the higher overlap with results from the Group 2 studies that used two *A. mellifera* subspecies (**Table 2.1**).

The transcripts detected in the analysis of the AC study had the highest relative overlap with sample-level meta-analysis and average overlap with the study-level meta-analysis results (**Table 2.1**). These results are consistent with the high overlap between the lists of transcripts detected by AC, AF and AM. The superior performance of the sample-level relative to the study-level meta-analysis for this study suggests that the analysis of the observations was able to pool the consistent information across studies that have low study-to-study variation. The combination of eight consistent estimates of differential expression in the study-level meta-analysis was not able to compensate for the number of estimates, and thus the overlap between the lists of transcripts from individual studies was lower. The inability of study-level meta-analysis to overcome partial consistency on a limited number of studies was also observed in the meta-analysis of mouse embryo studies (Rodriguez-Zas *et al.*, 2008).

2.5.3 Consideration of gene expression profiles

Although the list of transcripts detected in the analysis of the AM study had high overlap with the lists of transcripts differentially expressed in the AC and AD studies, the AM list of transcripts had the lowest overlap with the sample-level meta-analysis list of genes. This result was related to the change in the sign or pattern of the differential expression of one-day-old relative to forager honey bees between the individual analysis and sample level meta-analysis (**Table 2.2**). The consideration of the overlap of significance *P*-values and sign of the estimates is critical to understanding this scenario. Although there was a high number of transcripts with significant differential expression

in AM and AC or AD, the sign of some of these transcripts differed between these studies, and consequently, the meta-analysis did not detect these transcripts.

The analyses of AF and AD studies had the highest and lowest overlap of transcripts with study-level meta-analysis, respectively (**Table 2.1**). The former outcome suggests a situation where consistent estimates in AF and other studies, together with the low variation within study, overcomes the limited number of studies (eight estimates) analyzed, enhancing the capability of the study-level meta-analysis to detect differential expression. In this situation, the sample-level meta-analysis was unable to detect differential expression because the consistency of the results across studies was not able to compensate for the variation across studies. The outcome associated with the AD study is due to the low consistency of estimates between AD and other studies together with limited number of studies analyzed.

The sample-level meta-analysis detected more differentially expressed transcripts than the study-level meta-analysis in the presence of consistent patterns of expression in a few studies and small variation between studies relative to the overall signal of differential expression. This is particularly evident on the ability of the sample-level meta-analysis to detect differentially expressed transcripts that overlap with individual studies, even though the signs may differ (**Table 2.2**). Transcripts with significant yet different expression patterns in one study and sample-level meta-analysis were detected in other studies with patterns consistent with the meta-analysis. There was no change in the sign of the profile of differential expression between individual analyses and study-level meta-analysis. Thus, the analysis of eight estimates by the study-level meta-

analysis was not able to detect transcripts that may be significant in multiple studies but have different sign, for the levels of variation within study in the present work.

2.5.4 Comparison to previous work

Results from the analyses of individual studies were consistent with results from the combined analysis of the AC, AD, AF, and AM studies presented by Sen Sarma *et al.* (2007), and with results from the combined analysis of the MM, ML, LM, and LL data sets presented by Whitfield *et al.* (2006) and Rodriguez-Zas *et al.* (2006). Study-level and sample-level meta-analysis assist on the detection of genes with expression patterns that are consistent across studies. Genes that are not detected by the meta-analyses can be further investigated on an individual-study basis to identify gene expression profiles unique to one or few studies or conditions. Similarly, Sen Sarma *et al.* (2007) characterized brain gene-expression profiles that were different between species. Sen Sarma *et al.* (2007) reported a total of 1772 genes with differential expression (*P*-value < $1x10^{-3}$) between one day-old and forager honey bees across all four species (AC, AD, AF, AM), and of these, 218 genes were differentially expressed in two or more species. In the present analysis, 521 transcripts were differentially expressed in at least one species, and 113 transcripts were differentially expressed in two or more species in Group 1, also studied by Sen Sarma *et al.* (2007) (**Table 2.4**). Sen Sarma *et al.* (2007) reported that AF and AM had the highest number of genes with differential expression between one-day-old and forager honey bees in common among all study pair-wise comparisons. In the present study, a simple comparison of lists of differentially expressed genes indicated that AC and AM had the highest number of genes in common (**Table 2.1**). This result is

consistent with the fact that, among all four species, AC and AM are most similar in ecological, physiological, and behavioral characteristics. AD had the lowest number of genes in common with the rest of the Group 1 studies (between 9 and 12) and AC, AF, and AM had the highest number of genes in common (53 to 60). Sen Sarma *et al.* (2007) also reported that AD and AC had the lowest overlap of differentially expressed genes among all pair-wise comparisons. Based on ecological and behavioral considerations, AF was expected to exhibit the least overlap with the rest of the species.

The difference on the total number of differentially expressed genes reported by Sen Sarma *et al.* (2007) and in this study can be attributed to two reasons. First, the size of the data sets in the individual analyses presented in our study is approximately one-fourth of the total data set analyzed in Sen Sarma *et al.* (2007). The combined analysis of all four data sets is likely to offer more precise adjustments for technical sources of variation. Because a purpose of this study was to demonstrate the implementation of meta-analysis to integrate multiple (and typically small) studies, the four species were treated and analyzed as independent studies, thus potentially reducing the capability of each analysis to detect differentially expressed transcripts. Second, the model used to detect differential expression in Sen Sarma *et al.* (2007) (including analysis of ratios using two ANOVAs) differed from the model considered in this study. The normalized gene-expression values were analyzed with a mixed model including dye, sample as fixed factors and microarray as the random factor. The resulting data set was used to calculate ratios of forager to one-day-old expression levels for every gene and each microarray. A second ANOVA including sample was performed. Finally an F-test and a post-hoc Tukey's test were used to detect ratios significantly different among the 4 species and

between pairs of species, respectively. In the present study, the normalized expression levels of one-day-old and foragers were analyzed instead of ratios and the individual analyses model included the effects of dye, maturation stage, and sample. The data structure of the remainder data sets (Group 2) did not allow the inclusion of microarray in the model, and thus was also excluded from the Group 1 analyses.

Results from the individual analyses of Group 2 studies were consistent with results from the joint analysis presented by Whitfield *et al.* (2006). Whitfield *et al.* (2006) reported more genes (3745) with differential expression between one-day-olds and foragers than Sen Sarma *et al.* (2007). Similarly, analyses of individual studies (MM, LM, ML, and LL) in Group 2 resulted in more genes differentially expressed than in Group 1. The lower number of genes with differential expression detected in any one individual analysis in the present study (1752 transcripts) compared to Whitfield *et al.* (2006) (3745 genes) was attributed to different sample sizes and experimental models. Whitfield *et al.* (2006) analyzed brain gene-expression measurements from 108 microarrays including one-day-old, forager, and honey bees at other four additional intermediate stages of development, all arranged in a loop design. In the present study, only data from one-day-old and forager honey bees were compared, thus reducing the size of the data analyzed and information available to adjust for technical sources of variation. This difference in data sets was accompanied by a difference in the model used by Whitfield *et al.* (2006). The model used by Whitfield *et al.* (2006) to analyze the complete data set including all four studies (honey bee - colony subspecies combinations MM, LM, ML, and LL) and all ages included the effects of dye, age, subspecies of honey bee, colony, interactions, and random effects of microarray and sample.

The higher number of genes exhibiting differential expression between one-day-old and forager honey bees found in Group 2 studies relative to Group 1 studies is consistent with the higher number of differentially expressed genes reported by Whitfield *et al.* (2006) relative to the number reported by Sen Sarma *et al.* (2007) (**Table 2.2**). In the present work, the analyses of the LL and LM studies, that shared the subspecies of honey bee and differ on the colony, had the highest number of differentially expressed genes that overlapped (**Table 2.1**). This result is in agreement with Rodriguez-Zas *et al.* (2006) that reported that the vast majority of the genes with differential expression across six time points in *A. mellifera ligustica* honey bees exhibited the same pattern across colonies.

The higher overlap of differentially expressed transcripts observed between pairs of Group 2 studies was expected, as all the samples corresponded to closely related honey bees subspecies (*A. mellifera mellifera* and *A. mellifera ligustica*). Also, the highest overlap of differentially expressed genes among studies from different Groups was found between AM and MM, followed by AM and ML. This result reflects that samples from the same bee species (*A. mellifera*) were used in both Groups, and confirms the stronger association of the bee species relative to the colony species on the gene expression patterns (Whitfield *et al.*, 2006). Results from the present studies also identified similarities between AM and LM that share the same colony subspecies (*A. mellifera*). This similarity may be due to the adjustment or adaptation of *A. mellifera ligustica* bees to the *A. mellifera mellifera* host colony. The reason for the limited overlap between genes differentially expressed across groups of studies, even within honey bee species, may be the use of different populations and environments (Alaux *et al.*, 2009). The four

species of honey bees used by Sen Sarma *et al.* (2007) were all collected from suburban areas of Bangalore, while the honey bees in Group 2 (Whitfield *et al.*, 2006) were collected in France.

The similarity between studies AM and MM, ML, and LM based on the number of differentially expressed transcripts in common is consistent with the evaluation of the impact of honey bee and colony or host subspecies effect on brain gene-expression patterns across age. Results from the present studies identify similarities between AM and MM and ML that share the same honey bee or sample subspecies (*A. mellifera*). These results support the finding of Whitfield *et al.* (2006) that honey bee subspecies had a higher effect on brain gene-expression patterns across maturation stages than colony subspecies. Results from the present studies also identified similarities between AM and LM that share the same colony subspecies (*A. mellifera*). This similarity may be due to the adjustment or adaptation of *A. mellifera ligustica* honey bees to the *A. mellifera mellifera* host colony.

When considering non-age effects, Whitfield *et al.* (2006) reported that the majority of the differential expression was among honey bee subspecies, distantly followed by colony effect and interactions. Likewise, in the present study the highest number of overlapping genes among Group 2 studies was among honey bee - host combinations that shared the same honey bee subspecies (**Table 2.1**). Furthermore, Whitfield *et al.* (2006) reported that one-day-old honey bees exhibited highly consistent brain gene expression regardless of honey bee subspecies meanwhile forager honey bees exhibited more variation within and across subspecies. Thus, the moderate overlap of

genes differentially expressed among Group 2 studies found in this study may be due to fluctuations at the forager stage.

2.5.5 Gene Ontology classes and individual genes

The identification of individual genes and study of the representation of GO biological processes and molecular functions provided further insights into the information gained from the integration of information from multiple studies using meta-analysis. Functional analysis identified enrichment of generation of precursor metabolites and energy, cellular macromolecule metabolic process, transport, cell differentiation, system development, cofactor binding and purine nucleotide binding. Results of GO analyses are consistent with functional categories reported in previous studies. Functional analysis of 145 out of the 218 genes found by Sen Sarma *et al.* (2007) identified enrichment of numerous GO categories including protein binding, ion binding, and nucleic acid binding functions, response to biotic and abiotic factors, metabolism, pigmentation, and regulation of circadian rhythm, among others. Functional analysis of "hive to forager" genes found by Whitfield *et al.* (2006), with hive encompassing 0-, 4-, 8-, 12-, and 17-day old honey bees, identified enrichment of genes associated with energy pathway physiological processes.

Among the transcripts only found differentially expressed in the sample-level meta-analysis, transcripts corresponding to the fruit fly genes *Adar* and *Innexin 2* were over-expressed in forager relative to one-day-old honey bees. The *Adar* gene is associated with adult locomotory behavior and gene *Innexin 2* is associated with olfactory behavior, both behaviors critical for honey bee foraging. Genes *Chitin*, *Para, Transferrin*

1, *Sick*, *Oatp26F* and genes associated with translation, oxidoreductase, methyltransferase, carbohydrate transmembrane transport, and neurotransmitter activity (FBgn0022355, FBgn0035423, FBgn0037146, FBgn0038471, FBgn0038516, FBgn0050035, FBgn0030412, FBgn0032832, and FBgn0051634).

Consistent with Sen Sarma *et al.* (2007) and Whitfield *et al.* (2006), the expression of the honey bee orthologs to the fruit fly genes *Tctp* and *PebIII*, was lower in forager compared to one-day-old honey bees, but the differential expression was only significant for *Tctp*. Whitfield *et al.* (2006) reported a list of candidate genes for honey bee behavioral maturation that span all the six maturation stages (0, 4, 8, 12, 17 day-old nurse and 17 day-old forager honey bees) considered. Out of this list, the percentages of transcripts differentially expressed ($P < 1x10^{-3}$ or borderline) in two or more individual-study analyses in the present work and in at least one species studied by Sen Sarma *et al.* (2007) were 70% and between 58 and 75%, respectively. This result suggests that the results from the meta-analysis are supported by a previous independent study, and reiterates the superior ability of meta-analysis to detect differentially expressed transcripts compared to individual studies (i.e. 70% *versus* 66.5%). Among the genes detected in Whitfield *et al.* (2006) and by meta-analysis in the present study are *Inos*, *Cah1*, *Hsc70cb*, *Mlc-c*, *Bm-40-spa*, *Zormin*, *Smd3*, *Tctp*, *Orc1*, *Ef2b*, *Sh3beta*, *PebIII*, *Rfabp*, *Fax,* and *Mmpp2,* and the FlyBase IDs were FBgn0050387 (receptor signaling protein), FBgn0037146 (glutamate 5-kinase), FBgn0037303 (cysteine protease inhibitor), FBgn0037140 (organic cation porter) .

2.5.6 Model extensions

The effect of behavior on gene expression may vary with other (secondary) factors or covariates. The interaction between the main (behavior) and secondary factors may originate at the study or sample level. Study-level factors influence all samples within a study in a similar fashion; meanwhile sample-level factors can have variable effects among honey bee samples. Study-level secondary variables can be included in the meta-analysis model. In this study, adjustment for secondary study-level factors, like study group, can be incorporated into the individual or traditional estimate-based meta-analyses because there are multiple studies per level of secondary factor. Group represents a study-level covariate, and thus, unbiased estimates can be obtained even in the presence of maturation stage heterogeneity across studies (Schmid *et al.*, 2004). Colony is an example of a within-study source of variation that can also be included in the meta-analysis model. Group was not included in the present meta-analysis because the main goal was to investigate the benefits of meta-analysis expected in most situations, and because of the limited number of studies (four) per group. The availability of multiple studies per group that would allow an accurate adjustment for group effects is rare. In addition, the honey bee sample effect included in the individual-study and sample-level meta-analysis models also accounted for the previous sources of variation.

2.6 Conclusions

The objective combination of information implemented in the complementary meta-analytical approaches was able to mine the signal of differential expression of data

in different scenarios. Model-based meta-analysis approaches can rise above seemingly weak consistency among studies based on simple comparison of lists of genes. The sample-level meta-analysis detected more differentially expressed transcripts than the study-level meta-analysis among transcripts with consistent patterns of expression in a few studies, transcripts with expression well-described by similar model parameter estimates across studies, and transcripts with low variation between studies relative to the overall signal of differential expression. Study-level meta-analysis is the appropriate approach when only estimates of differences in expression among conditions of interest are available. Genes that do not conform to the assumptions of the sample-level meta-analysis, and have consistent expression patterns but substantial variation across studies, are benefited by the study-level meta-analysis.

Meta-analytical approaches uncovered genes associated with differences between one-day-old and forager honey bees across studies, regardless of the species or sub-species of honey bee sampled or in the colony. Among these, genes *Adar*, *Innexin 2*, *Chitin*, *Para, Transferrin 1*, *Sick*, *Oatp26F*, FBgn0022355, FBgn0035423, FBgn0037146, FBgn0038471, FBgn0038516, FBgn0050035, FBgn0030412, FBgn0032832, and FBgn0051634 are strong candidates for additional studies.

CHAPTER 3

DEVELOPMENT OF META-CLASSIFICATION TECHNIQUES TO IDENTIFY GENE BIOMARKERS THAT ACCURATELY CLASSIFY SAMPLES

3.1 Abstract

Classification of samples into groups (i.e. behaviors, treatments, stages) based on gene expression profiles using data mining and model-based strategies is an area of active research. However, most approaches use the gene expression and not the classification group as dependent variable to detect candidate biomarker gene. Also, the performance to classify samples decreases when the biomarker genes uncovered on one experiment are applied to another experiment. Lastly, most approaches provide one set of biomarkers and alternative biomarkers are proposed using less suitable, tagged-on strategies. To address these limitations, a three-stage meta-classification strategy to identify classifier biomarker genes that provide accurate classification of samples across experiments is put forward. In the first stage, non-linear mixed effects models were used to identify promissory classification genes across studies. Three complementary approaches that integrate information across experiments on a by-gene basis were compared: a) overlap of genes that have expression profiles significantly associated with the classes, b) joint analysis of estimates of association between gene expression and classes obtained from each experiment, and c) joint analysis of the association between gene expression and classes across experiments. In the second stage, complementary multi-gene classification approaches based on promissory classification genes identified in the first stage were

evaluated. Different criteria to select promissory classifiers and to combine the classifiers were evaluated using stepwise, forward and backward methods based on the correct classification rate. In the third stage, alternative biomarker sets with similar classification performance to the set developed in the second stage were identified using a score selection method, where the top 10 models that included combinations of the best classifiers were outputted. A leave-one-study-out cross-validation analysis assessed the performance of the newly identified classifier genes on samples other than those used to develop the classifiers on both the second and third stages. The strategy was evaluated on eight microarray experiments that compared the gene expression of honey bees representing two behavior classes (one-day-olds and foragers) on different species and sub-species. Evaluation of the classification resulting from the first stage confirmed the capability of the non-linear meta-analysis of estimates or observations across experiments to detect separate candidate gene classifiers that would have been missed by the individual analysis of each experiment. The subsequent identification of a biomarker set and cross-validation on independent experiments in the second stage allowed the detection of a gene set with outstanding correct classification rate. The identification of additional gene sets with comparable classification performance in the third stage provided insights into the biological processes that differ between classes. Results were further examined using functional enrichment analysis, finding over-representation of genes with the molecular function of ferric iron binding, and biological processes of cell homeostasis, iron ion transport, and cation homeostasis. Genes identified as possible biomarkers for behavior among all stages were *Mrjp5*, *Tsf1*, *Lys-2*, *PNUTS,* CG16791-PA, LOC410038, BI516304, BI514790, BI509882, GB14166, GB14157, and BI513123.

The proposed three-stage meta-classification strategy is a practical and informative alternative to identify complementary biomarker genes, molecular functions and biological processes that offer consistent and reliable classification across experiments.

3.2 Background and Motivation

Classification of samples into groups using gene expression information is an area of active research (Zhu and Hastie, 2004). Genes that have expression patterns associated with the probability that a sample pertains to a class or group can be used as biomarkers to classify samples with unknown classification. This paradigm can have a critical impact on the prognosis and diagnosis of diseases and disorders based on the expression measurements of classifier genes. In addition, the identification of classifier genes can be used for the effective design of experiments targeting specific genes or pathways. Numerous computationally- and statistically-based methods have been used to classify samples based on gene expression measurements, including support vector machines (SVM) (Lee and Lee, 2003; O'Donnell *et al.*, 2005; Ramaswamy *et al.*, 2001), penalized logistic regression (Zhu and Hastie, 2004) and quantile regression (Wright *et al.*, 2003).

The methods to identify genes that can be used to classify samples have three major challenges. First, the high-dimensionality of microarray studies, typically including thousands of gene expression profiles per sample, necessitates the streamlining of hundreds or thousands of potential classifiers. Second, the limited number of samples per class further exacerbates the need to condense the number of classifier genes into scores. In some methods, the number of predictor variables must be lower than the

number of observations. In general the more predictor variables, the higher the risk to develop a classifier function that has outstanding performance on the training samples used to develop the classifier function but poor performance on non-training samples. This situation is also known as over-fitting or over-parameterization. Third, there are multiple criteria to assess the performance of a classifier or combination of classifiers. Furthermore, different study objectives and data characteristics may favor one criterion over others. The total number of classifier genes, statistical significance of the classifier genes, and correct classification rate are some of the criteria applied in the identification of classifier genes. Fourth, the results from the application of the computational or statistical methods to classify samples must be easy to interpret. Several classification methods can accommodate a large number of predictor variables, and have good classification performance, but rely on complex or non-linear combinations of the original gene expression predictor variables (e.g. partial least squares, penalized logistic regression). In these cases, the relationship between the classifier genes and sample class is difficult to interpret in biological terms, thus hampering the understanding of the biological processes and planning of additional experimental work.

Microarray gene expression technology is becoming more widely used, and multiple microarray experiments that include samples from the same classes or groups are becoming available. Fixed and random meta-analysis approaches have been extensively used to combine information from multiple experiments and identify genes that are differentially expressed across groups (Adams *et al.*, 2008; Rodriguez-Zas *et al.*, 2008). However, these meta-analysis techniques have not been applied to identify

classifier genes. The integration of information across experiments is expected to increase the ability to detect genes that can classify samples.

Although many approaches to identify biomarker genes based on expression profiles have been proposed, several limitations can be noted. In most cases, biomarkers are identified from models in which the dependent variable is the gene expression, not the classification group of interest. Also, the performance of the approaches to classify samples may decrease substantially when the biomarker genes uncovered on one experiment are applied to another experiment. For example, Rhodes et al. (2004) used a leave-one-out approach to test genes that were identified as having significant association with cancer class, however classifiers were trained and tested on the same dataset in which they were identified. Zhang and Fenstermacher (2005) proposed the identification of biomarker reporter genes using a rank-sum test statistic and the combination of expression levels across studies using a linear index that is trained and tested across datasets, however testing on an independent dataset was not conducted, and therefore the results may be biased as their application to other datasets may not work. Lastly, most approaches provide one set of biomarkers and alternative biomarkers are proposed using less suitable, tagged-on strategies. For example, a multi-factorial model is used to identify a set of biomarker genes and additional genes are identified solely based on their correlation to the original set of biomarker genes. Then, all biomarkers resulting from the different analyses are combined on a single list that can be explored for common functional categories.

Tan *et al.* (2006) uses partial least squares, and Fort and Lambert-Lacroix (2005) combine partial least squares and penalized logistic regression to classify samples from

microarray data. While these approaches do take into account the highly-dimensional nature of microarray studies through the use of a dimensionality reduction method, there is no consideration of the actual sample classes in the studies. The value of formal meta-analysis approaches that combine more than simple lists of genes (Ma and Huang, 2009; Rhodes *et al.*, 2004; Wren, 2009) to integrate information across studies has not been exploited in the area of biomarker identification and sample classification.

Most studies that aim at identifying biomarkers use either individual gene profiles or a large number of gene profiles relative to the number of samples available for classification. Individual genes are unlikely to have a high correct classification performance, especially in studies with more than 10 samples or multiple groups or single-handedly controlled complex processes. On the other hand, the use of large numbers of genes to classify a number of samples that is similar or small leads to artificial perfect or nearly-perfect classification due to model over-specification. Thus, when applied to other datasets, these genes do not perform well. A compromise between classification accuracy and generalization in the context of microarray experiments and meta-analysis has not been explored.

The overarching goal of this study is to present a novel strategy to classify samples and identify robust classifier genes that overcomes the limitations of previous approaches. The proposed strategy is based on the meta-analysis of classes modeled using gene expression measurements from multiple experiments. A multi-stage approach that begins with the detection of individual gene classifiers, followed by the identification of biomarker sets with reliable performance across experiments is put forward. The proposed strategy was applied to the identification of biomarkers of behavior class using

gene expression profiles from 8 honey bee microarray experiments. Complementary approaches that address the challenges encountered in the application of classification approaches to individual microarray experiments, and the challenges to combine the information from multiple experiments, were evaluated. Approaches that provide easy-to-interpret results were considered, and multiple indicators of classifier suitability were assessed. Further analysis of these biomarkers provided insights into the molecular functions, and biological processes that distinguish the classes. The results from the proposed strategy can be used in the effective design of subsequent confirmatory experiments.

3.3 Materials and Methods

3.3.1 Data sets

Eight microarray gene expression studies from two experiments were available for the meta-analysis. All studies compared the expression of genes at two maturation stages, or behaviors, newly emerged one-day-old adult honey bees ("one-day-olds") and 17-day-old adult forager honey bees. The goal of these experiments was to profile the transcriptome of honey bees that occurs in parallel to profound physiological, anatomical, and behavioral differences observed between one-day-old and forager honey bees. A newly emerged honey bee performs within-hive tasks associated with the maintenance and survival of the colony for two to three weeks and then performs foraging activities outside the hive (Rodriguez-Zas *et al.*, 2006). Eight studies were divided into 2 groups based on species: Group 1 - AC, AD, AF, AM; Group 2 - LL, LM, ML, MM. For in-

depth information about the experiments and data preprocessing, see Data Sets and Data Processing sections of Chapter 2, also in Adams *et al.* (2008).

3.3.2 Models and methods

A three-stage strategy was used to identify genes that could be use to reliably classify honey bee samples into maturation stage or behavioral groups across experiments. In the first stage, three approaches that integrate information across studies within gene were implemented. In the second stage, results from the first stage were incorporated into multiple classification approaches to develop classifier functions and identify classifier genes. Results were evaluated using a cross-validation approach. In the third stage, additional biomarker gene sets with comparable classification performance were uncovered and sets were evaluated using a cross-validation approach. To gain insights into the molecular functions, biological processes and pathways that exhibit distinct expression profiles between classes, a functional analysis of the resulting biomarker gene list was undertaken.

3.3.2.1 First stage: multi-study analyses

Three meta-analysis approaches were used to combine data from multiple studies and identify promising gene classifiers of samples. The three approaches applied within gene were: a) individual-study analysis, b) study-level meta-analysis, and c) sample-level meta-analysis.

3.3.2.1.1 Individual-study logistic meta-analysis. All eight studies were analyzed separately using a logistic regression model (Agresti, 2007). Behavior was the binary response variable, with a 50:50 ratio of forager and one-day-old honey bee observations

in genes with no missing observations due to data preprocessing. The logistic model considered for each study was:

$$\log\left(\frac{\pi_{ijl}}{1-\pi_{ijl}}\right) = \mu + D_i + \beta x_j + C_l \qquad [3.1]$$

where $\log\left(\frac{\pi_{ijl}}{1-\pi_{ijl}}\right)$ is the logit link function, or the natural logarithm transformation, of the ratio between the probability that an observation pertains to a forager and the complementary probability that the observation pertains to a one-day-old honey bee. This log of odds (odds of forager *versus* one-day-old) was described with a linear model that included an intercept (μ), the effect of the i-*th* dye (D_i , i = red, green), β or the change in log odds per unit change in the expression of a gene (x_j), and the effect of the l-*th* colony (l = 1, 2, 3). A t-test was used to assess the statistical significance of the association between the odds of behavior and gene expression. The maximum number of observations analyzed per gene in the individual-study analysis was 24 in the Group 1 studies and 16 in the Group 2 studies.

3.3.2.1.2 Study-level logistic meta-analysis. Estimates of the change in odds of behavior per unit change in gene expression obtained in the individual-study logistic analysis were jointly analyzed. The estimates were standardized using their corresponding standard error and were analyzed using a linear mixed effects model:

$$\begin{aligned} \hat{\beta}_i &= \mu + s_i + e_i \\ \mathbf{s} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \mathbf{e} &\sim N(\mathbf{0}, \mathbf{I}\sigma^2) \end{aligned} \qquad [3.2]$$

where $\hat{\beta}_i$ is the standardized estimate of the change in behavioral odds per unit change in gene expression obtained from the individual-study logistic analysis of the i-*th* study (*i* = 1 to 8), ***s*** denotes the vector of study effects**,** and ***e*** denotes the vector of residuals. The study effects were assumed to be independent, identical, and normally distributed with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$ of dimension *k* x *k* where *k* is eight, the number of studies. The variance-covariance matrix had a diagonal structure with off-diagonals of zero, and variance of the estimate of differential expression within study in the diagonals. This assumption accommodated potential heterogeneity of variance among studies. Likewise, residuals were assumed to be independent, identical, and normally distributed with mean zero and common variance σ^2. The maximum number of estimates analyzed per gene in the study-level meta-analysis analysis was eight.

3.3.2.1.3 Sample-level logistic meta-analysis. This analysis used the same model as the individual-study analyses to describe the log odds of behavior across all studies and includes a study effect:

$$\log\left(\frac{\pi_{ijkl}}{1-\pi_{ijkl}}\right) = \mu + D_i + \beta x_j + S_k + C(S)_{kl} \qquad [3.3]$$

The model terms are defined as in the individual-study analysis with the exception of S_k that denotes the effect of the k-*th* study (k = 1 to 8). Like in the individual-study analysis, honey bee was not included in the model because a honey bee can only pertain to a single behavior class. Microarray was not included in the model because two microarrays found in the Group 2 studies only had one behavior class. The maximum number of observations analyzed per gene in the sample-level meta-analysis analysis was 158.

Estimates were obtained using a restricted maximum-likelihood approach and a multiple test adjustment was implemented to control for false positive results. A false

discovery rate threshold P-value < 0.1 corresponded approximately to an unadjusted P-value < 0.001. A search for genes that were associated with significant differential odds (DO) of behavior in the individual-study, study-level, and sample-level meta-analyses was conducted. The analyses were compared in terms of the genes identified as associated with differential odds of behavior, and the direction (sign) of the log odds. A positive log odds estimate indicates that a unit increase in the expression of a gene is associated with an increase in the chances of the sample to be a forager honey bee. A negative log odds estimate indicates that a unit increase in the expression of a gene is associated with an increase in the chances of the sample to be a one-day-old honey bee. Promising classifier genes to use in the second stage of the classification strategy were identified based on the statistical significance and consistency of the association with the odds of behavior across analyses.

3.3.2.2 Second stage: multi-gene classification

In the first stage of the classification strategy, 7,734 genes were analyzed individually, and genes that had significant and consistent association with the odds of behavior across multiple analyses were identified. The second stage of the classification strategy complements the first stage in two ways. First, four classification techniques based on the simultaneous consideration of multiple genes were developed. These techniques have complementary benefits. Second, the resulting classification functions were evaluated in terms of the accuracy to assign samples into the known classes. The multiple classification techniques were trained and tested on the complete data set, which included all eight studies.

3.3.2.2.1 First classification technique. A logistic model identical to the sample-level logistic meta-analysis, but including as covariates the expression of the top five most significant genes, was used. The five genes were associated with significant DO in both the study-level and sample-level meta-analyses. Only five genes were considered because the minimum number of samples within study was eight.

The next three classification techniques used expression information from the top 25 genes from stage one results. The use of the top 25 genes allowed the consideration of additional patterns of association between the odds of behavior and gene expression not represented among the top five genes. However, the inclusion of the top 25 genes as covariates in the same way as in the first classification technique resulted in large standard errors and often extreme regression coefficients. These results are symptoms of nearly complete separation (or perfect classification of the samples) due to a relatively large number of classifier genes relative to observations, or high collinearity between the classifier genes (Hosmer and Lemeshow, 1989; Menard, 2001). Collinearity and complete separation do not result in biased parameters or inaccurate inference, but result in inefficient estimation of the parameters. Furthermore, this scenario is associated with the identification of a gene set that is an ideal or nearly ideal classifier for the particular data set under consideration but with poor generalizable properties in other data sets. The goal of the study is to evaluate classification techniques that provide unbiased, generalizable, and precise estimators and accurate classification of samples for the sample size (slightly less than 80 honey bees), number of classes (two classes), and classification techniques considered. Thus, the following three classification techniques summarized the information from the top 25 genes into fewer classifier variables.

3.3.2.2.2 Second classification technique. A logistic model identical to the first classification technique, but including the top 25 most significant genes, was considered. Principal component analysis (PCA) was used to explore the capability of the top 25 genes to separate the observations into the known one-day-old and forager classes. Principal component analysis is a multidimensional reduction technique used to identify combinations of the original gene expression variables that account for much of the variation in the data (Stekel, 2003). In PCA, a high dimensional space is projected onto a lower dimensional space, while retaining as much of the variability from the original data as possible. A gene expression experiment is composed of p components, or genes, which can be represented graphically with p axes. Using PCA, the number of components can be reduced to q. Eigenvalues are calculated for each of the q components, or axes, which are related to the amount of variation explained by the axis (Johnson and Wichern, 2007). The sum of the eigenvalues is equal to the number of components, and the eigenvalues are expressed as a percentage of the total variation.

To reduce the number of classifier genes and overcome the problem of over-parameterization, three variable selection analyses were used. Multiple variable selection analyses were used because of their complementary advantages. A logistic analysis was used to fit a logistic regression model to response data, and provided three possible methods of variable selection: forward, backward and stepwise. Selection methods are extremely useful when numerous genes, in the case of microarray gene expression data, are fitted to a model. It allows for only a few key genes to be selected from a larger group of genes, and fitted to the model for analysis. In forward selection, the first n explanatory variables are added to the model. Chi-square statistics are computed for each

variable not in the model, and if the largest of these is significant, it is added to the model. This step of analyzing the remaining variables is repeated until no variables meet the significance threshold for entry into the model. When a variable is added to the model using forward selection, it is never removed. In backward selection, again the first n explanatory variables are added to the model. In this case, a Wald test is used to examine the remaining variables. The Wald statistic is calculated using the formula,

$$\frac{\left(\hat{\theta}-\theta_0\right)^2}{\text{var}(\hat{\theta})} \qquad [3.4]$$

where $\hat{\theta}$ is the maximum likelihood estimate of the parameter of interest, θ_0 is the proposed value, and $\text{var}(\hat{\theta})$ is the variance of the maximum likelihood estimate (Wald, 1943). This statistic can then be compared to a chi-squared distribution. The least significant variable that does not meet the specified threshold is removed from the model, and is never re-added. The process is repeated until no remaining variables meet the criteria. Stepwise selection can be considered a combination of both forward and backward methods. Variables previously entered into the model can be removed in another step, likewise, those removed previously have a chance to be re-entered into the model in subsequent steps. The process continues until no variables can be added, or if the last variable added is also the last one removed in the next step. A P-value < 0.2 was used as a threshold to enter a gene into the model using forward and stepwise selection, and to drop a gene from the model using backward and stepwise selection. Genes selected by all three methods were combined and a final stepwise logistic model was conducted to obtain the final set of genes to be used as classifiers.

3.3.2.2.3 Third classification technique. A discriminant analysis including the top 25 most significant genes as in the second classification technique was considered. To reduce the number of classifier genes and again overcome the problem of over-parameterization, stepwise selection with a variable entry and removal threshold P-value < 0.1 was used. Discriminant analysis and logistic regression are techniques applied to data in which categorical dependent variables are being analyzed. Discriminant analysis is considered the reverse of a one-way multivariate analysis of variance (MANOVA). The independent variables from MANOVA become the categories of the dependent variable for discriminant analysis. Likewise, the dependent variables from MANOVA become the predictors for discriminant analysis (Press and Wilson, 1978). Discriminant analysis is commonly used when only two categories exist in the dependent variable, but it can be used for multi-level categories as well. Unlike discriminant analysis, logistic regression always refers to the use of a dichotomous dependent variable. This may prove to be a great advantage of discriminant analysis over logistic regression. On the other hand, logistic regression has been used more often, as the approach is considered to be more statistically robust, and easier to understand than discriminant analysis. However, both methods lead to the same conclusions on a qualitative level (Press and Wilson, 1978). Gene significant P-values from the discriminant analysis were equal to those from a logistic regression model.

3.3.2.2.4 Fourth classification technique. A logistic model identical to the first and second classification techniques, but including a coefficient corresponding to the expression of the top 25 most significant genes (instead of the expression of five genes), was used. The predictive coefficient was obtained by applying partial least squares (PLS)

and encompassed the first three PLS factors. Similar to PCA, PLS is capable of projecting a high dimensional space onto a lower dimensional space, while still capturing a majority of the variability from the original data. The difference between PCA and PLS, in general, is that PCA produces a weight matrix reflecting the covariance structure between the predictor variables ("Partial Least Squares", 2008), whereas PLS produces a weight matrix reflecting the covariance structure between the predictor and response variables (Abdi, 2007). The PLS analysis directly considers information involving group separation when redefining the space (Liu and Rayens, 2007). Since PCA develops principal component functions without consideration of the affiliation of samples to groups, the resulting functions are not necessarily highly correlated with the conditions under experiment (Liu and Rayens, 2007).

3.3.2.3 Third stage: identification of additional biomarkers

The preceding two stages allow for the identification of a set of genes that accurately classified the samples into groups. However, other sets of genes are expected to also offer accurate classifications. Our goal was to identify all the genes in the microarray platform that can be used to classify samples, not just the best set. In addition, the identification of multiple gene classifiers permitted the investigation of the representation of Gene Ontology molecular functions, biological processes and pathways among the complete biomarker set. This approach allowed to uncover not just individual genes or sets of genes, but also to apply a systems biology approach to uncover the functions and processes that exhibit distinct profiles between the classes. These functions and processes may be responsible or associated with other functions that elicit the differences between classes and can be explored in follow-up experiments.

Correlations between genes were investigated as a preliminary and quick means to find sets of genes with profiles similar to the biomarker genes identified in the previous stage. However, selection of additional genes solely based on high correlations resulted in high error rates. A likely reason for the poor performance of the correlation approach is that the model factors in the first two stages cannot be accounted for by the correlation.

A novel strategy was used to identify additional sets of biomarker genes with good classification performance. As before, a logistic regression model was used to describe the sample class (one-day-old or forager), but here the score method was used to identify classifier genes with strong statistical association with the class, and overcome the problem of over-parameterization. In score selection, a branch-and-bound algorithm from Furnival and Wilson (1974) is used to identify a specified number of "best" models, or those models with the highest likelihood scores, measured as chi-square statistics. The branch-and-bound algorithm is completed in two steps, branching and bounding. In branching, a given set of S candidate genes will return two or more smaller subsets (S_1, S_2, …) with the largest chi-square values. In bounding, the upper and lower bounds of the chi-square values are computed, where if the lower bound for a subset of candidates, A, is greater than the upper bound for a separate subset of candidates, B, then B is discarded. A minimum upper bound threshold is typically set for all subsets. The process is continued until the set of candidates, S, is reduced to the specified target models, or when the upper bound matches the lower bound (Furnival and Wilson, 1974).

The top 10 classifier sets of sizes one to five genes were identified for analysis through implementation of the start and stop options to impose the minimum and

maximum subset sizes, respectively. As in previous techniques, a maximum of five genes were used, as the minimum number of samples within the studies was eight and based on the number of classifiers identified in the second stage. Outputting the best 10 models allowed for enough flexibility to arrive at models that produced great classifiers without generating an excessive number of models with low performance. Principal component analysis (PCA) was used to complement the results found using the score technique through chi-square and eigenvalue outputs showing the amount of variability the model is accounting for in regards to separating the observations into the known one-day-old and forager classes.

3.3.3 Validation

A leave-one-study-out cross-validation analysis was conducted to assess the performance of the classifier genes identified in the second and third stages of the analysis on samples other than the ones used to develop the classifiers. The training data set consisted of seven of the eight studies, and the testing data set contains the remaining study. A leave-one-study-out cross-validation was used, instead of a leave-one-sample-out cross-validation, to address dependencies between observations from different samples measured in the same microarray or study. A leave-one-group-of-studies-out analysis was not investigated because of the limited number of observations to develop classifiers that would have been available in the training data set.

Multiple indicators of the accuracy to classify samples were computed. In this study, correct classification rate was the fraction of observations that were correctly classified, sensitivity was the fraction of forager observations (positives) that were

correctly classified, and specificity was the fraction of one-day-old observations (negatives) that were correctly classified. All training and cross-validation analyses were implemented using SAS (SAS Institute Inc., NC, USA).

3.3.4 Gene Ontology

Once the gene biomarkers were identified, the representation of Gene Ontology molecular functions, biological processes and pathways among these genes was investigated (Gene Ontology Consortium, 2000). This allowed to uncover molecular functions and biological processes that are candidates to understand the differences between the classification groups and aided in the interpretation of the classifier genes. Representation was assessed using functional enrichment analysis. Lists of genes were assigned to Gene Ontology (GO) biological process and molecular function classes based on fruit fly annotations. Representation of genes in GO classes was evaluated using Fisher´s exact (two-tailed) test and False Discovery Rate multiple test adjustment P-value < 0.1 (Al-Shahrour *et al.*, 2006).

3.4 Results

3.4.1 First-stage: Multi-study analyses

The study-level and sample-level logistic meta-analyses identified 63 and 870 genes associated with significant (FDR P-value < 0.1) differential odds of behavior class, respectively **(Table 3.1)**. No genes associated with significant differential odds were identified in any individual-study analysis.

The analysis of gene expression as a response variable and behavior as an explanatory variable using linear mixed effects models (including also the fixed effect of dye and the random effect of honey bee sample) reported by Adams *et al.* (2008) uncovered a higher number of significant associations in the individual-study, study-level, and sample-level meta-analyses than the present analysis of the behavior binary response variable (included the fixed effects of dye and colony). The difference between the results of Adams *et al.* (2008) and the present study is due to the difference in the model terms considered. This was confirmed by the similar results obtained from the analyses of the associations between behavior and single gene expression when one variable (behavior) was the response variable and the other (expression) was the explanatory variable (and vice versa), and excluding all other variables. This outcome may be due to lower information available when the response variable is binary relative to continuous. These results support the hypothesis that integration of consistent information across studies using meta-analysis increases the likelihood to detect genes associated with differential odds, especially when each study has limited information.

The signs of the β estimates (change in odds of behavior per unit change in gene expression) from the study-level and sample-level meta-analyses were consistent. The majority of the genes identified as significantly associated with differential odds in the study-level meta-analysis were also identified by the sample-level meta-analysis (**Table 3.1**). The partial overlap between the study-level and sample-level meta-analysis gene lists indicated the complementary nature of both meta-analysis approaches also observed in the analysis of gene expression as a response variable reported by Adams *et al.* (2008). Sample-level meta-analysis can detect more differential odds than the study-level meta-

analysis when the odds of behavior are well described by similar gene expression changes across studies, and have small variation across studies. Study-level meta-analysis was well-suited for behavior odds that exhibit consistent associations with gene expression across studies but also have substantial variation across studies, or behavior odds that do not conform to the overarching assumptions of the sample-level meta-analysis.

3.4.2 Second stage: Multi-gene classification

Biomarker genes are identified using UniGene identifiers (Pontius *et al.*, 2003) and/or FlyBase annotation symbols (Tweedie *et al.*, 2009) when available. Otherwise, GenBank accession identifier and NCBI dbEST identifiers are provided. The top twenty-five genes selected for consideration as biomarkers based on the significance of the association between odds of behavior and expression in the study-level and sample-level logistic meta-analyses of individual genes and corresponding gene annotation is provided in **Table 3.2**. The associations of the top genes on an individual basis, and the odds of behavior using a sample-level logistic meta-analysis approach, are presented in **Table 3.3**.

The advantages of the integration of meta-analysis and logistic model include an extensive theoretical framework and easy to interpret results. Case in point, the estimates in **Table 3.3** correspond to the logarithmic odds of forager *versus* one-day-old. Back-transformation of the estimate in the underlying scale to the estimate in the observed scale provides information on the odds of behavior ($\pi/1-\pi$) and probability of either behavior (π, $1-\pi$). For example, BI503685 [BB170016B20C12] had the most significant (and negative) association with odds of behavior. The estimate β was -2.04, and thus the

negative sign indicates increases in the expression of BI503685 are associated with a decrease in the probability of being a forager. Back-transforming the estimate leads to the interpretation that the change in odds of forager *versus* one-day-old is 0.13 ($\exp^{-2.04}$) per unit increase in the expression of BI503685. This value, though positive, indicates a decrease in the probability of the sample being a forager. The reason is because odds have positive (> 0) values, with values between 0 and 1 indicating that the probability of the event in the denominator ($1 - \pi$, or one-day-old) is more likely than the event in the numerator (π, or forager). Thus, the estimate 0.13 indicates a decrease in the odds that the observation pertains to a forager honey bee with increase of expression level. Further transformation of the odds into the change in probability of forager (π) gives the value 0.115 (0.13/(1+0.13)) per unit increase in the expression of BI503685. Like with the odds, probabilities are positive numbers between 0 and 1 with values < 0.5 indicating a decrease in the probability of an observation to pertain to a forager relative to a one-day-old honey bee.

The gene lysozyme 2 (*Lys-2*) *CG11159-PA* [GB15106, BI503571] had the most significant and positive association with odds of behavior. This gene is involved in the biological process of translation (GO:0006412) and the molecular function of triplet codon-amino acid adaptor activity (GO:0030533). The estimate, β, was 0.60, and thus the positive sign indicates increases in the expression of *Lys-2* are associated with an increase in the probability of being a forager. Back-transforming the estimate leads to the interpretation that the change in odds of forager *versus* one-day-old is 1.82 ($\exp^{0.60}$) per unit increase in the expression of *Lys-2*. This value indicates an increase in the odds that the observation pertains to a forager honey bee with an increase in gene expression, and

the estimate 1.82 indicates that the odds of forager increases with increase of expression level. Further transformation of the odds into the change in probability of forager (π) gives the value 0.65 (1.82/(1+1.82)) per unit increase in the expression of *Lys-2*. Again as mentioned above, probabilities are positive numbers between 0 and 1 with values > 0.5 indicating an increase in the probability of an observation to pertain to a forager relative to a one-day-old honey bee.

The results presented in **Table 3.3** demonstrate that the association between significance P-value and classification error rate is not linear, even among the most significant genes. The significance P-value for the top 10 genes (analyzed individually) ranged from $1x10^{-7}$ to 0.0003, and the classification error rate ranged from 25% to 41%, which translates to a range from 40 to 62 misclassified observations among the 158 observations analyzed. For example, the expression of GB10622 [BI509691, BB170018B10A09] had a more significant (P-value < 0.00014) association with the odds of behavior than the gene major royal jelly protein 5 (*Mrjp5*) [GB10622, BI509691, BB170018B20G10] (P-value < 0.0003). However, the classification error rate of GB10622 was higher (62 misclassified observations) than that of *Mrjp5* (57 misclassified observations).

Another interesting result summarized in **Table 3.3** is the association between estimates, statistical significance and type of misclassification. Genes BI513447 [BB160012B20G04] and BI515017 [BB160017A10D11] had nearly identical differential odds estimates and standard errors (approximately -0.46 and 0.12, respectively), similar P-values (0.00009, and 0.0002, respectively), and slightly different correct classification rates (0.64 and 0.60, respectively). However, the correct classification rates of

observations into forager and one-day-old groups for BI513447 were 0.61 and 0.72, respectively, while the equivalent rates for BI515017 were 0.67 and 0.54, respectively. Meanwhile, BI513447 had a better performance classifying observations into the one-day-old group relative to forager group, and BI515017 had a better performance classifying observations into the forager group relative to the one-day-old group. These results show that for similar estimates and P-values, the accuracy to assign observations to each class can vary drastically. Furthermore, the identification of classifier genes must consider not only P-values and estimates but also overall correct classification rate and within-group correct classification rate. The forager, one-day-old and overall correct classification rates presented in **Table 3.3** were computed on the same observations used to develop the logistic predictive equation. On the other hand, the P-value refers to the population of observations, of which the data set under consideration is only one realization. Another find was that in **Table 3.3**, even though both BI513447 and BI514790 [BB160016A20H02] had similar P-values, the genes gave complementary information. Gene BI513447 had a higher specificity (more true negatives), suggesting that this gene is more suited in predicting one-day-old honey bees. This is also noted by the negative estimate value. Gene BI514790 has a higher sensitivity (more true positives), which suggests it is more suited in predicting forager honey bees. This suggestion is also noted by the positive estimate value.

An exploratory PCA of the top 25 genes identified as candidate classifiers of samples from different behaviors was conducted. **Figure 3.1** depicts the location of the average of the observations per study (to facilitate the visualization) using the first two principal components. The first two principal components explained 74.5% of the

variation of the gene expression, and the first three principal components explained 82.6% of the variation.

The expression of the top 25 genes allowed for linear separation (solid line) between one-day-old (denoted with red markers and labeled with "O") and forager (denoted with black markers and labeled with "F") groups. This result was expected because the 25 genes were selected based on their significant association with behavior. However, significance alone does not ensure good separation between classes as demonstrated in **Table 3.3**. In addition, the top 25 genes allowed for linear separation (dashed line) between Group 1 (markers labeled AM, AD, AF, and AC) and Group 2 (markers labeled LL, LM, ML, and MM) studies. The PCA demonstrated that the samples within behavior class in the Group 2 studies are more consistent amongst themselves (narrowly spread) than the samples within behavior class in the Group 1 studies. The distance between study groups is consistent with the sample collection within study across groups. These results are consistent with the report by Adams *et al.* (2008) that the overlap of genes found differentially expressed between behaviors using linear meta-analysis was higher between studies within Group than across Groups. Although the distance between behaviors is more evident among studies in Group 1 than among studies in Group 2 in the principal component plot, Adams *et al.* (2008) reported more genes differentially expressed between behaviors in the Group 2 studies. This can be explained by the higher consistency (less noise) among observations from the same behavior in Group 2 studies than in Group 1 studies portrayed in **Figure 3.1**.

3.4.2.1 First classification technique

Results from sample-level logistic meta-analysis, including a linear combination of the expression of the top five most significant genes (CG16791-PA (GB18624, BI508922), collagen alpha-1(II) chain precursor LOC410038 (GB15988, BI508870), lysozyme 2, *Lys-2* (GB15106, BI503571), *Mrjp5*, and transferrin 1, *Tsf1* (GB19745, BI511993)), are presented in **Table 3.4.** As expected, the consideration of the expression of five genes improved the correct classification rate (only 8/158 samples misclassified or 0.05 misclassification rate) relative to each individual gene. The significance P-value of these genes decreased when analyzed together, suggesting collinearity (correlation) among some of these genes. Genes *Tsf1* and LOC411154 had the highest drop in P-values (from $7.5x10^{-3}$ and $3.4x10^{-4}$ to 0.46 and 0.27, respectively). **Table 3.5** presents the pair-wise correlation between the top 25 genes. Gene *Tsf1* had a strong negative correlation (-0.7) with LOC410038. Likewise, LOC411154 had a moderate correlation (0.46) with CG16791-PA. Removal of *Tsf1* from the model did not deteriorate the correct classification rate, further confirming the collinearity of this gene with LOC410038. **Table 3.4** also includes the estimates and standard errors of the change in behavior odds per unit change in the gene expression. The estimates for genes LOC410038 and *Tsf1* have opposite signs, consistent with the correlation between these genes reported in **Table 3.5**. The known function of these genes may provide insights into the differences between the one-day-old and forager groups. *Tsf1* is involved in numerous biological processes [cellular iron ion homeostasis (GO:0006879); defense response (GO:0006952); iron ion transport (GO:0006826)] and molecular functions [ferric ion binding (GO:0008199); iron ion transmembrane transporter activity

(GO:0005381)], and increases in expression of the gene have been linked to activated ovary development in worker bees (Koywiwattrakul and Sittipraneed, 2009). Sen Sarma *et al.* (2007) found *Tsf1* in brain samples to show significant differential expression between species for behavior, confirming the current study results for the gene's association with behavior. Gene LOC411154, similar to fruit fly gene *Nat1*, is also involved in numerous biological processes [autophagic cell death (GO:0048102); salivary gland cell autophagic cell death (GO:0035071); RNA metabolic process (GO:0016070)] and molecular functions [protein binding (GO:0005515); translation initiation factor activity (GO:0003743)], and is found to play a crucial role in translational regulation, as *Nat1* knockout individuals possessed numerous abnormalities, or showed embryonic lethality (Yoshikane *et al.*, 2007). Gene CG16791 is involved in the process of phagocytosis and engulfment (GO:0006911). Gene LOC410038 had no known associations to any biological processes or molecular functions (Gene Ontology Consortium, 2000). Sen Sarma *et al.* (2009) identified *Lys-2* as having a role in three species of honey bee, AC, AD, AM, during dance communication behavior. The current study found *Lys-2* to be significantly associated with the classification of foraging versus one-day-old behaviors, and as foragers engage in dance behaviors upon returning to the hive, this helps to confirm *Lys-2*'s association with honey bee behavior.

3.4.2.2 Second classification technique: Logistic

A logistic model similar to the model used in the first classification technique, but including the top 25 most significant genes, was considered. Results from this model are presented in **Table 3.6**. All observations were correctly classified. As can be expected, the more genes used to classify samples, the lower the error rate, as additional genes help

to explain the behavior differently compared to previously considered genes. However, many genes had similar association with behavior, and therefore would not give extra aid in classification when another gene previously accounted for the same information in the classification model. The over-parameterization of the model, including highly collinear variables and complete separation or perfect classification of the observations obtained from this model, resulted in high standard errors of the estimates or weak model efficiency. Thus, gene variables were removed from the model using a combination of backward, forward, and stepwise selection methods. The stepwise selection method using a P-value < 0.2 for variable entry into the model and a P-value < 0.1 for variable stay alone identified only one classification gene, LOC410038. However, this gene had a high classification error rate of 29.47% (47/158 samples misclassified). A model with four genes (CG16791-PA, LOC410038, BI516304, and BI514790) offered perfect classification of the observations into behavioral groups. Standard errors were lower than what was seen in the 25 gene model, as fewer variables corrected the over-parameterization issue that arose previously with the inclusion of too many genes in the model. The correlation between genes CG16791-PA and BI516304 and between LOC410038 and BI514790 was moderate (absolute value of approximately 0.50), and all genes had a P-value > 0.05. Interestingly, even though BI516304 had the highest classification error rate among all top 25 genes when considered alone (56% of the observations were incorrectly classified, **Table 3.5**), and was moderately correlated with CG16791-PA, this gene was selected for in the final model and contributed to the perfect classification achieved by the final logistic model. Removal of BI516304 from the logistic model resulted in the misclassification of 7 of the 158 observations. This result

confirms that the expression of BI516304 contributed to the description of the odds of behavior in a manner complementary to the other genes. As anticipated, the simple correlation between two genes is not a sufficient indicator of the collinearity between genes with respect to the response variable in models that include other terms and non-linear models.

3.4.2.3 Third classification technique

A linear discriminant analysis including the top 25 genes was conducted. This discriminant analysis was over-parameterized as in the logistic analysis, and included 25 genes. **Table 3.7** provides the coefficients for the linear discriminant function corresponding to each behavior. The extreme coefficient values for some genes are evidence of the over-parameterization or extensive number of discriminant variables relative to the number of observations to classify, and the number of classes. The genes with more divergent coefficients between the forager and one-day-old discriminant functions (e.g. BI513447 and BI514790) are expected to be the ones that remain in the discriminant function after the stepwise variable selection approach is implemented. However, due to the over-parameterization of the model, the absolute differences between behavior discriminant coefficients are not a reliable indicator of gene classifying strength. The relative difference of coefficients between behaviors obtained from a less parameterized discriminant model is a more suitable indicator of classifier strength.

The final discriminant function, after stepwise variable selection on the top 25 genes, included 10 genes. The genes were LOC410038, CG16791-PA, *Lys-2*, BI503685, BI514790, BI513447, LOC411154, phosphatase nuclear targeting subunit CG33526-PD isoform D (*PNUTS*, GB12917, BI507137), *Tsf1*, and BI504564 (GB17444), and

produced perfect classification. Similar to *Lys-2*'s association with dance communication behavior, Sen Sarma *et al.* (2009) found GB17444 to also be significantly associated with the behavior in species AD, AF and AM, supporting the current finding of its role in the foraging behavior. Also within these genes is *PNUTS,* which is involved in numerous known biological processes and molecular functions, including phagocytosis, protein targeting, regulation of catalytic activity, nucleic acid binding, and zinc ion binding (Gene Ontology Consortium, 2000). The large standard errors of a few genes indicated that the discriminant model was inefficient and suggested that this model may be over-parameterized. For the data set considered, the variable selection methods applied to logistic and discriminant analyses identified the same predictor genes and had perfect classification. Stepwise, forward and backward methods of selection were applied to the top 25 genes and 15 biomarker genes remained in the model (**Table 3.8**). A stepwise logistic analysis was further used to analyze these 15 genes. Results from this analysis revealed the same four genes as when this method was applied using only logistic analysis and not stepwise discriminant analysis. The final genes identified by the variable selection approaches applied to the discriminant and logistic analyses were different. This may be due to the use of a pooled residual variance common to both classification groups in the discriminant analysis. When a discriminant analysis was applied to the four genes identified by the logistic analysis (summarized in **Table 3.6)**, using individual within-group covariances for each group, the results were consistent between the discriminant and logistic approaches.

3.4.2.4 Fourth classification technique

Partial least squares (PLS) was used to combine the information from the top five or top 25 genes into the first three PLS factors. A logistic model similar to that in the first and second classification techniques, but including the PLS coefficient as an explanatory variable instead of the top five or 25 genes, was used to classify observations into behavior groups. The goal of PLS was to reduce the number of logistic model terms with limited loss of information. The predictive coefficient was obtained by applying partial least squares (PLS) and encompassed the first three PLS factors that accounted for 84.45% of the variation of the 25 genes. The first PLS component explained 57.22% and 64.95% of the variation of the top five and 25 genes, respectively. The strongest association (P-value) between the PLS coefficient and behavior was found when the PLS analysis represented the top five genes. However, the lowest classification error rate was found when the PLS analysis represented the top twenty-five genes. The statistical significance and classification error rate from PLS with five and 25 genes were P-value <0.0001 and 6%, and P-value < 0.13 and 0%, respectively. These results confirm that P-value and classification error rate provide complementary information on the ability of multiple genes to classify observations, and that additional genes can provide complementary information and aid in the classification of the samples.

3.4.3 Third stage: Identification of additional biomarkers

A logistic model using the score method of selection on all 7734 genes was considered. Results from this approach are presented in **Table 3.9**. Thirty-seven biomarker genes were identified through score selection from the five models, combined.

No models including one to three genes had perfect classification rates. For models with one, two and three gene combinations, error rate ranges were 1.3 to 6.3%, 6.3% to 14%, and 24.7% to 35%, respectively. For models with four and five gene combinations, error rates ranged from 0% to 2.5%, and 0% to 1.9%, respectively. Out of the top 10 specified models based on the Chi-square statistic, both four- and five-gene sets produced three models each with perfect classification rates (**Table 3.9**). As previously confirmed, the more genes used to classify samples, the lower the error rate, as additional genes aid in the explanation of behavior different than any previously considered genes.

Score selection of the top 10 models with combinations of four genes identified numerous genes among models with perfect classification rates that were not within the top 25 genes previously identified in the first stage of the study, including genes tropomyosin 1 CG4898-PD, isoform D (LOC408583, BI513123), BI514288 (BB160014B20F05), LOC726976 (GB18177, BI514401), LOC724644 (GB10220, BI515411), BI515625 (BB160020A10B04), LOC408687 (GB17707, BI517000), CG5630-PA (LOC725773, GB16694, BI509685) and WAP four-disulfide core domain 2 (LOC408864, GB10310, BI510230] (**Table 3.9**).

Genes CG16791-PA and BI513447 were jointly found in the top 25 genes from stage one as well as in score selection of four gene combinations with perfect classification rates, and are considered top genes to classify samples based on behavior. Neither gene was identified previously in the literature as having significant associations with behavior. Although a match to the BI513447 sequence has not been found in the fruit fly genome, the sequence has a very significant match (E-value = 10^{-35}) in a *Nasonia giraulti* (wasp) BAC library (BAC NG_Bb-47D01). Gene CG16791-PA is involved in

phagocytosis and engulfment. Phagocytosis and encapsulation are a mechanisms of the insect innate immune system as phagocytosis involves the engulfment of relatively small objects such as bacteria, yeast or apoptotic cells by plasmatocytes (Feldhaar and Gross, 2008). The CG16791-PA biomarker of behavior group found in this study suggests that the genes and pathways of the immune system could hold additional clues as to the differences between behavior groups. Due to differences in their tasks, one-day-old and forager honey bees are exposed to different immune challenges with distinctive intensity and range.

The score selection of the top 10 models with combinations of five genes identified 9 genes among models with perfect classification rates that were not within the top 25 genes previously identified in the first stage of the study, including genes coronin, actin binding protein, 2B (LOC409316, GB20076, BI510708), CG1112-PA (LOC725953, GB10854), LOC726976 (BI14401, GB18177), LOC724644 (BI515411, GB10220), LOC408583, BI515625, LOC408687 (GB17707), BI509910, BI508683, and BI509882 (**Table 3.9**). Gene CG1112-PA was found to show significant differences in dance communication behavior for honey bees in species AD, AF and AM by Sen Sarma *et al.* (2009). Genes with known molecular functions include CG1112-PA and LOC408583, which are involved in carboxylesterase activity (GO:0004091) and actin binding (GO:0003779), respectively (Tweedie *et al.*, 2009). Gene BI513447 was jointly found in the top 25 genes from stage one as well as in score selection of four gene combinations with perfect classification rates. Genes BI513447, LOC724644, BI515625 and BI517000 were commonly found in models of both four and five gene combinations, suggesting their importance in classification of behavior. Gene BI513447 has a match to the

Nasonia girualti BAC library, LOC724644 is an hypothetical protein conserved across many insect species (*Anopheles gambiae, Aedes aegypti, Tribolium castaneum, Ixodes scapularis, Acrythosiphon pisum*) and the rest of the EST have not been annotated. Results from the application of this technique suggest the importance of including all genes for analysis regardless of previously identified significance, as an insignificant gene in combination with others may better classify unknown samples than a significant gene.

3.4.4 Validation

The multiple-gene classification techniques were trained and validated on the same data set, including observations from the eight studies. A leave-one-study-out validation approach was used to assess the performance of logistic models to classify observations in data sets not used in the training of the logistic classification function. The logistic model was trained using seven out of the eight studies and validated on the remaining data set. **Tables 3.10**, **3.11 and 3.12** can be interpreted using study AC from **Table 3.10** as an example: genes listed for study AC were those identified when all studies except AC were used in the training data set; those genes were then tested on AC, of which 23/24 samples were correctly classified.

Table 3.10 presents the results from the application of the final logistic model (after variable selection, section 3.4.2.2) to each training data set. The classification was perfect in five of the independent testing data sets, and in the remaining data sets the classification error rate was small, ranging from 0.04 to 0.19. Two of the data sets with imperfect classification had one gene less (BI516304 and BI514790 in validation data

sets AC and ML) and one additional gene (LOC411154 in validation data set ML). The validation data set with the poorest classification was the AF study. The limited capability of the genes selected from other studies to classify the honey bees in the AF study is consistent with findings by Sen Sarma *et al.* (2007) and Adams *et al.* (2008). The rationale for this could be that phylogenetic analyses indicate that the honey bee species in the AM and AC studies (*Apis mellifera* and *Apis cerana*, respectively) are more closely related to one another than to the AD and AF (*Apis dorsata* and *Apis florea*, respectively) studies. Consideration of the correct classification of forager and one-day-old observations (i.e. sensitivity and specificity) indicated that all incorrectly classified observations in the AF testing data set were from forager honey bees that were classified as one-day-old observations. **Table 3.11** presents the results from the application of the final logistic model to each training data set, where genes were selected after stepwise, forward and backward selection methods were applied to the training data sets. Application of the selection methods aided in perfect classification rates for studies AC and AF, where previously error rates existed (**Table 3.10**). **Table 3.12** presents the results from the application of the final logistic model to each training data set, where genes identified from the score selection classification technique were selected after the backward selection method was applied to the training data sets. Three of the eight studies presented error rates, ranging from 0.04 to 0.21. The validation data set with the poorest classification was the AF study, similar to what was seen in **Table 3.10**.

3.4.5 Gene Ontology Analysis

The biomarker genes detected in the different scenarios were analyzed for representation of molecular functions and biological processes using Fisher's exact test (Al-Shahrour *et al.*, 2006). Among genes identified as significantly differentially-expressed from the subject-level meta-analysis in the first stage, the biological process glutamine family amino acid metabolic process (GO: 0009064) and the molecular function monooxygenase activity (GO:0004497) were over-represented with P-values < 4.56×10^{-3} and 6.32×10^{-3}, respectively. Biomarker genes identified in the second stage using the top 25 genes (nine had fruit fly annotations), genes for one biological process, microtubule nucleation (GO:0007020), had significant over-representation at P < 2.87×10^{-4}. Biomarker genes identified in the third stage using the score method of selection in logistic (10 with fruit fly annotations), genes for one molecular function, ferric iron binding (GO:0008199), had significant over-representation at P-value < 4.44×10^{-5}. Biological processes in which genes were over-represented were cell homeostasis (GO:0019725), chemical homeostasis (GO:048878), transition metal ion transport (GO:0000041), cation homeostasis (GO:0030003), iron ion transport (GO:0006826), metal ion homeostasis (GO:0006875), and di-, tri-valent inorganic cation homeostasis (GO:0030005) with P-values < 3.67×10^{-4}, 3.12×10^{-4}, 6.36×10^{-5}, 4.22×10^{-4}, 5.22×10^{-5}, 7.67×10^{-4} and 6.31×10^{-4}, respectively. No pathway was over-represented among the biomarker genes identified. The current study identified GO categories that were not found in Adams *et al.* (2008) or Sen Sarma *et al.* (2007; 2009), suggesting additional functions and processes that influence honey bee behavior.

3.5 Conclusions

Meta-analysis has already been identified as an ideal tool in analyzing high-dimensional data like microarray gene expression experiments. Only meta-analyses of the expression of individual microarray genes have been reported. Likewise, the value of classification approaches to identify biomarker genes and pathways associated with biological processes has been widely recognized. Due to the complexity of the biological systems considered, typically multiple microarray elements are necessary to identify biomarker genes that provide an accurate classification of samples. The extraordinary capability of combining meta-analysis and multi-gene classification approaches to identify reliable biomarker genes for behavior is demonstrated in this study.

Results from the present study show that the integration of information from multiple studies, and expression profiles from multiple genes, improved the classification of samples relative to the analysis of individual studies or use of individual gene expression profiles. Study-level and sample-level logistic meta-analyses have complementary strengths to detect genes significantly associated with the odds of the class under consideration (i.e. behavior). The combination of statistical significance associations between gene expression and odds of behavior, and classification error rate, is a powerful indicator of the ability of genes to classify observations accurately. Among the top twenty-five most significant genes from the meta-analysis results were numerous genes identified as key classifiers of behavior, including *Mrjp5*, *Tsf1*, *Lys-2*, and *PNUTS*. The logistic regression model provided estimates and P-value indicators of the association between genes and the odds of behavior that were easy to interpret. The

global and behavior-specific classification error rates provided information complementary to the estimates and P-values.

Inclusion in the logistic model of a high number of genes (relative to the number of samples available), and collinearity between genes, can result in inefficient estimates. The combination of backward, forward and stepwise variable selection methods resulted in a parsimonious classification model without compromising classification accuracy. A logistic model using the score method of selection on all 7734 genes identified thirty-seven biomarker genes as potential classifiers of behavior. The leave-one-study-out cross-validation approach for the first strategy confirmed genes *CG16791-PA*, LOC410038, BI516304, and BI514790, and the second strategy confirmed genes BI509882, GB14166, GB14157, and BI513123, which are considered reliable classifiers of observations between one-day-old and forager honey bee samples. Genes identified as reliable classifiers of behavior were identified with involvement in numerous biological processes and molecular functions, including cell homeostasis (GO:0019725), iron ion transport (GO:0006826), cation homeostasis (GO:0030003) and ferric iron binding (GO:0008199). Through the three-stage strategy presented, reliable classifier genes for behavior were identified based on their gene expression levels. The results from the three-stage approach can be used to design experiment aimed at corroborating the association between the biomarker genes identified and the difference between the groups compared.

CHAPTER 4

META-ANALYSIS OF GENE EXPRESSION STUDIES PROVIDES INSIGHTS INTO CATTLE EMBRYO DEVELOPMENT ACROSS REPRODUCTIVE TECHNIQUES, DEVELOPMENTAL STAGES AND SOURCES

Adams HA, Southey BR, Everts RE, Marjani SL, Tian XC, Lewin HA, Rodriguez-Zas SL
(Submitted to Functional and Integrative Genomics)

4.1 Abstract

The complex process of embryo development is affected by numerous genetic factors. Previous work has characterized genes at individual pre- and post-implantation stages, or considered either embryonic or maternal tissue sources, or specific reproductive technologies. The goal of this study was to integrate information from four gene expression studies encompassing various developmental stages (7 d, 25 d, 280 d), sample sources (embryonic or extra-embryonic) and reproductive technologies (artificial insemination, AI; somatic cell nuclear transfer, NT) to gain a more comprehensive understanding of factors influencing embryo development in the cattle model through individual-study and meta-analytical techniques. Meta-analysis uncovered 434 additional genes, across technologies, not detected by individual study analyses alone (36 to 846 significant genes per study) or overlap between studies, illustrating its advantages over simple across-study comparisons. Further analyses of genes within and between technologies, sources and stages identified suites of genes particular to the variable of interest (e.g. identified *OXT* from all extra-embryonic tissue sources, and *DKK3* from all

25 d samples). Gene Set Enrichment Analysis uncovered functional categories associated with differences between technologies (e.g. transferase activity [genes *AKT1*, *CSNK2B*, *POLG2*, *STK24*, *ACVR2A*, *WDR33*, *PRIM2A*, and *CRLS1*]). Conclusions: 1) stage has more influence on expression than technology or source; 2) embryonic and extra-embryonic sources have a similar response to changes in stage or technology; and 3) the effect of technologies on efficiency and large offspring syndrome (LOS) may rely on large expression modifications of only a few genes.

4.2 Background and Motivation

Meta-analysis is used to integrate information from studies that considered the same experimental factors or conditions (e.g. treatments) and levels within a factor. However, it is seldom the case that multiple microarray experiments are undertaken to study the same conditions. In most cases, studies share some of the levels of a factor and the standard meta-analytical approaches. This is because the differences between studies prevent the direct comparison of gene lists across studies or combination of P-values or fold-change estimates of particular levels that are not represented in all studies. The proposed research addressed the situation of integrating information from studies that do not share conditions.

The goals of this study were: 1) to develop a meta-analysis approach that allowed for the combination of information across partially overlapping studies, 2) to implement the approach to characterize gene expression profiles in cattle embryos spanning a variety of conditions, 3) to compare the results of this novel approach against other standard meta-analysis approaches.

The availability of both analysis results and raw data permitted the comparison of three complementary approaches to combine studies (comparison of gene lists, model-based study-level meta-analysis and sample-level meta-analyses) to gain a more comprehensive understanding of the genes, biological processes and pathways associated with embryo development. Two sets of meta-analyses were undertaken to answer different questions. The first set of meta-analyses identified differentially expressed genes between AI and NT technologies, regardless of developmental stage or sample source. The second meta-analysis identified differentially expressed genes between stages, sources or technologies for particular additional condition. These two sets of analyses enabled the identification of genes with common profiles across stages and sources, and of genes that exhibit differential expression between assisted reproductive technologies at particular stages or sources and between stages or sources within technology. This study revealed insights into the transcriptomic regulation of reproductive technologies within and across stages and sources that could help in the understanding of the efficiency of reproductive technologies, and ultimately improve the reproductive performance of livestock through genomic-based selection and breeding practices.

4.3 Materials and Methods

4.3.1 Data sets

Four cattle gene expression studies that compared samples from AI and NT technologies that differ in developmental stage and sample source were selected. The partial overlap of conditions between these studies, the shared platform and the

availability of the cattle genome made them the ideal study set to meta-analyze and uncover genes and pathways critical to embryo development. The studies encompassed two different sample sources (embryonic or E, and extra-embryonic or X) and three developmental stages (day 7 of gestation or d7, day 25 of gestation or d25, and at term or d280). The four microarray studies are denoted d280X (Everts *et al.*, 2008), d7E (Marjani *et al.*, 2009; Smith *et al.*, 2007), d25X (Everts *et al.*, 2007a) and d25E (Everts *et al.*, 2007b). The number of AI (and NT) samples were 9 (20), 5 (6), 6 (6), and 9 (20) for the studies d7E, d25E, d25X, and d280X, respectively. In the d7E study, NT embryos from two cell lines with 21.8% and 6.4% term development percentage were analyzed as one NT group based on preliminary comparisons that indicated consistent profiles between lines. The four studies used a 13,257 oligonucleotide spotted cattle microarray platform created from cattle placenta and spleen cDNA libraries (Gene Expression Omnibus or GEO platform identifier GPL2853, (Loor *et al.*, 2007)). Of the probes, 9,655 have a match on Human RefSeq (March 2007), 1,460 have identification on cattle UniGene, 54 have no hits, and the rest have matches to other Mouse, Human or cattle databases (Loor *et al.*, 2007). All studies used a reference design with dye swap such that each sample appeared in two microarrays with different dyes.

4.3.2 Data normalization

All available information was used to normalize the data within each study to remove or minimize any global study-specific effects. Basic data processing included log2 transformation of the background-subtracted foreground intensities, and the removal of spots that were flagged by the scanning software (GenePix Pro 5.0) or did not surpass

the median intensity of the negative control spots for both dyes. Log-transformed values were normalized using a global LOWESS transformation (Cui *et al.*, 2003) to remove dye bias within microarray, and duplicate spots within gene were averaged following procedures that are described by Rodriguez-Zas *et al.* (2006) and Adams *et al.* (2008), and implemented in Beehive ("Beehive", 2006). Gene expression data from duplicated spots in the platform were averaged within dye. Global dye and microarray effects were removed across all microarray elements or genes (Wolfinger *et al.*, 2001). All data processing and meta-analyses were conducted using the SAS mixed procedure (SAS Institute Inc., NC, USA).

4.3.3 Meta-analysis of reproductive technologies

The study of gene expression patterns between reproductive technologies (AI and NT) spanning developmental stages and sample sources is complex. The complexity stems from the fact that knowledge of the genes and pathways influencing the efficiency of different reproductive technologies is advanced not only from uncovering consistent differential gene expression across studies (i.e. stages or sources), but also from detecting differential gene expression unique to a stage or source. To address this situation and effectively uncover complementary types of gene expression patterns that could provide insights into the differential efficiency of reproductive technologies, three different meta-analysis strategies to integrate differential gene expression information between AI and NT samples across stages and sources were considered. The three strategies were: 1) standard overlap of genes with significant differential expression across studies identified by the individual-study analyses; 2) study-level meta-analysis of the differential

expression estimates (obtained from the individual-study analyses); and 3) sample-level meta-analysis of gene expression across studies.

4.3.3.1 Individual-study analysis

The individual-study model described each gene within study with a linear mixed-effects model that included the fixed effects of dye, reproductive technique (AI or NT), and the random effects of sample nested within reproductive technique, and microarray. The gene expression values, y_{ijkl}, are represented as follows:

$$y_{ijkl} \sim N\left(\mu + t_i + d_j + c_k + a_l, \sigma^2\right) \qquad [4.1]$$

$$c \sim N(\mathbf{0}, T)$$

$$a \sim N(\mathbf{0}, \mathbf{A})$$

where t_j is the effect of the i-*th* treatment (AI, NT, reference sample), d_j is the effect of the j-*th* dye, c_k is the effect of the k-*th* cattle sample, a_l is the effect of the l-*th* microarray, and σ^2 is the error variance. The sample and microarray effects were assumed to be independent and have a Normal distribution (~ N) with mean zero and common sample and array variances σ^2_c ($\boldsymbol{T} = \boldsymbol{I}\sigma^2_c$) and σ^2_a ($\boldsymbol{A} = \boldsymbol{I}\sigma^2_a$), respectively. For study d280X, the fixed effect of gestation length (whether the sample was term or preterm) was also added to the individual-study model, as it was found to have a significant effect on gene expression in the original analysis (Everts *et al.*, 2007a). The total number of observations analyzed per study was 29, 11, 12 and 29 for d7E, d25E, d25X, and d280X, respectively.

4.3.3.2 Study-level meta-analysis

This approach (termed Study) combined estimates of differential expression between the reproductive technologies resulting from the individual-study analyses. These estimates are sometimes the only information available from studies with

published results but no publicly available raw data (Adams *et al.*, 2008). The differential expression estimates combined were first standardized by their corresponding standard errors to account for differences in the precision of the estimates resulting from studies with different sizes and levels of technical noise that cannot be corrected for by the model terms. The linear mixed-effects model used to combine the standardized estimates of differential expression included an overall mean (μ), and the random effect of study (s_i):

$$y_i \sim N\left(\mu + s_i, \sigma^2\right) \qquad [4.2]$$

$$\mathbf{s} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

where y_i is the standardized estimate of difference in expression between AI and NT samples obtained from the analysis of the i-*th* study (i = 1 to 4), μ is an indicator of the consistency of differential expression across studies that used different sample sources and were done at different developmental stages, σ^2 is the error variance, and s is the vector of study effects. The random effect study was assumed to have a Normal distribution with mean zero and variance-covariance matrix S of dimension 4 x 4. The diagonals of the variance-covariance matrix were the variances of the estimates from each study, and the off-diagonals were assumed to be zero.

4.3.3.3 Sample-level meta-analysis

A linear mixed-effects model was used to describe the normalized gene expression intensities and test for reproductive technology effect across all four studies. The model (termed Sample) was:

$$y_{ijklm} \sim N\left(\mu + d_i + t_j + s_k + c(s)_{kl} + a_m, \sigma^2\right) \qquad [4.3]$$

$$\mathbf{c} \sim N\left(0, \mathbf{I}\sigma_c^2\right)$$

$$\mathbf{a} \sim N\left(0, \mathbf{I}\sigma_a^2\right)$$

where y_{ijklm} is the normalized log-transformed gene expression, μ is the overall mean, d_i is the effect of the i-*th* dye, t_j is the effect of the j-*th* treatment (AI, NT, and reference sample), s_k is the effect of the k-*th* study (k=1 to 4), c_l is the effect of the l-*th* cattle sample, a_m is the effect of the m-*th* microarray, and σ^2 is the error variance. Random effects of cattle sample and array were assumed to be independent and normally distributed with a mean of zero and variances σ^2_c and σ^2_a, respectively. The potential effect of amplification of the d7E samples was tested by including the effect of amplification, nesting the study effect. The effect of amplification was significant in a very small percentage of the more than 10,000 genes analyzed and was removed from the analysis. This result is consistent with the report of Smith *et al.* (2005). Both the study and sample models search for consistent patterns of expression across studies.

The set of studies considered allows the investigation of an additional set of comparisons, in addition to that between reproductive technologies. For this purpose, an alternative specification of the sample-level meta-analysis model that combines the sample source, developmental stage and reproductive technology into a single factor was used in the sample-level meta-analysis. A total of nine contrasts that were divided into two sets of contrasts were used to extract a more comprehensive understanding of the gene expression profiles across sample sources, developmental stages and assisted reproductive technologies. The first set of contrast was specified within reproductive technology: embryonic at 7 d *versus* embryonic at 25 d, extra-embryonic at 25 d *versus* extra-embryonic at 280 d, and embryonic at 25 d *versus* extra-embryonic at 25 d. These contrasts permit the profiling of embryonic and extra-embryonic samples across

developmental stages and the differences in gene expression between embryonic and extra-embryonic sources at the same time point, separately for AI and NT. The second set of contrasts allowed the study of differential expression between reproductive technologies within sample source or stage, and included AI *versus* NT in embryonic samples regardless of developmental stage, AI *versus* NT in extra-embryonic samples regardless of developmental stage, and AI *versus* NT at 25 d regardless of sample source.

4.3.4 Post-analysis mining of results

Genes were considered differentially expressed within analysis if the test statistic was significant at a minimum P-value $< 1 \times 10^{-3}$ and the gene exhibited a minimum fold change threshold ≥ 1.23. In this study the significance threshold was comparable to a False Discovery Rate adjusted P-value < 0.05 (Benjamini and Hochberg, 1995). The comparison of results across analyses also considered the sign or direction of the differential expression. The results from the analyses were also compared using funnel plots. Four scenarios with biological and statistical implications were investigated; 1) genes that were not differentially expressed in any individual study, yet were detected by the study-level and sample-level meta-analyses; 2) genes that had differential expression between AI and NT samples in two studies, but were not detected by meta-analysis; 3) genes that were found differentially expressed in the sample-level meta-analysis but not in the study-level meta-analysis, and; 4) genes that were differentially expressed in the study-level meta-analysis but were not in the sample-level meta-analysis.

Gene Set Enrichment Analysis (GSEA) of all gene expression profiles was undertaken to uncover over-represented biological processes, gene pathways and

molecular functions within expression groups (Al-Shahrour *et al.*, 2006; Subramanian *et al.*, 2005). Due to the higher number of genes with Gene Ontology and pathway assignments in human than in cattle, cattle genes that have human annotations for molecular function, biological processes and KEGG (Kanehisa and Goto, 2000) gene pathways were considered. Among all analyzed probes, 31.6% had corresponding human functional annotations. The magnitude of the standardized estimates of the difference between conditions (e.g. AI *vs* NT, 7 d *vs* 25 d) was used to indicate differential gene expression.

4.4 Results and Discussion

4.4.1 Meta-analysis of reproductive technologies

4.4.1.1 Differential gene expression between AI and NT samples within and across studies

The number of genes with significant differential expression between AI and NT samples obtained from the analysis of individual studies ranged from 36 to 846 (**Table 4.1**) and the trend in the results are consistent with previous reports from each study (Everts *et al.*, 2007a; Everts *et al.*, 2007b; Everts *et al.*, 2008; Marjani *et al.*, 2009; Smith *et al.*, 2007). **Appendices B** and **C** present the gene identifiers, corresponding cDNA identifiers, gene symbols, estimates of difference in expression between conditions, and P-values resulting from all of the analyses. No particular sign or pattern of differential expression dominated across analyses, with the exception of study d280X that had 189 and 657 positively and negatively differentially expressed genes, respectively. This finding suggests that the expression of genes in the placenta is down-regulated with NT

relative to AI during the final stages of embryo development. The percentage of genes differentially expressed and overlapping among studies ranged from 0.0% to 13.1% (average of 7.65%) relative to the number of genes identified in one of the studies. The low overlap among the list of genes from individual study analyses may be due to the threshold used to identify significant differentially expressed genes relative to the power of each study and demonstrates the potential limitation of these methods to combine studies (**Table 4.1**).

The study-level and sample-level meta-analyses were able to uncover 168 and 380 differentially expressed genes between AI and NT across studies, respectively (**Table 4.1**). The Venn diagram in **Figure 4.1** depicts the overlap of genes identified as differentially expressed by the individual studies, study-level and sample-level meta-analyses. Of the 380 genes detected by the sample-level meta-analysis, 302 genes were not detected by any single individual- study analysis. Likewise, of the 168 genes detected by the study-level meta-analysis, 149 were not detected by any single individual-study analysis. These genes did not reach the significance level in any single study and the pooling of consistent information by the meta-analyses allowed these genes to be uncovered. For example, study-level meta-analysis alone found OLIGO_01033 [AF489575] to be differentially over-expressed in NT *versus* AI samples. This gene corresponds to the cattle gene endothelin converting enzyme 2 (*ECE2*), which is known to act in human in brain and heart development, along with other processes crucial to human embryonic development (Yanagisawa *et al.*, 2000). Two genes were significant in an individual study, the study-level, and the sample-level meta-analyses. The OLIGO_12625 [CK772645] was over-expressed in NT *versus* AI samples in d25E and in

the study-level and sample-level meta-analyses, and corresponds to human gene solute carrier family 22 (organic cation transporter) member 1 (*SLC22A1*). A previous study states that expression of *SLC22A1* does not exist during embryo development (Schweifer *et al.*, 1997), however another study shows the gene encodes for membrane-spanning transporter proteins that show tissue-specific expression patterns in embryos (Verhaagh *et al.*, 1999). The OLIGO_04736 [Hs.610896] was over-expressed in NT *versus* AI samples in d280X, study-level and sample-level meta-analyses. The gene is similar to calcium/calmodulin-dependent protein kinase II delta (*CAMK2D*), which functions in ATP, calmodulin and nucleotide binding, as well as transferase activity and calmodulin-dependent protein kinase activity (Maglott *et al.*, 2005).

Meta-analysis techniques offer the capability to discriminate between lack of consistency among the studies on differentially expressed genes due to lack of precision in multiple studies or due to true inconsistency of the gene expression profiles between reproductive technologies across studies. The number of genes differentially expressed in one and two studies was 1028, and 14, respectively (**Table 4.2**). Eight of the fourteen genes differentially expressed in 2 studies had opposite signs. No genes were found differentially expressed between AI and NT in more than two individual studies. Of the 380 genes detected by the sample-level meta-analysis, 376 genes were not detected in the overlap of two individual-study analyses. Likewise, of the 168 genes detected by the study-level meta-analysis, 168 were not detected in at least two individual-study analyses. The higher number of differentially expressed genes from meta-analysis augments the number of genes that can be explored in subsequent experiments.

Of the differentially expressed genes detected in at least one individual study, 95 were detected by meta-analysis, and 947 genes were not detected by either the study- or sample-level meta-analyses. The low overlap between individual studies and meta-analysis is likely due to differential expression being limited to a single developmental stage or tissue such that the gene expression was inconsistent across all studies. Furthermore, this overlap of differentially expressed genes is likely not an indication of lack of statistical precision of the sample-level meta-analysis (based on the number of observations) because both the study- and sample-level meta-analyses were able to uncover additional differentially expressed genes that were not detected in the overlap of gene lists from two studies.

4.4.1.2 Pattern of differential gene expression between AI and NT samples within and across studies

Consideration of the gene expression profile between the individual-study analyses and meta-analyses provided additional insights into the results (Table 1). There were 8 genes detected as differentially expressed in two individual-study analyses that had positive signs (over-expression in AI compared to NT) in the analysis of one study, and had negative signs in the analysis of another study (**Table 4.3**). The OLIGO_03653 [DV784696], associated with the human gene heart and neural crest derivatives expressed 2 (*HAND2*), was significantly over-expressed in AI *versus* NT samples in d280X, and significantly under-expressed in AI *versus* NT samples in d7E. This may suggest differences in gene expression in extra-embryonic *versus* embryonic sources in genes associated with development, or changes in expression patterns throughout gestation. The *HAND2* gene is involved in the development of cardiovascular structures (Morikawa and

Cserjesi, 2008), and has been found to be crucial to the survival of mouse embryos past 10.5 days post-conception (Srivastava *et al.*, 1997). The OLIGO_08124 [CN440810], associated with the human gene polymerase, gamma 2, accessory subunit (*POLG2*), was significantly over-expressed in AI *versus* NT samples in d7E, and was significantly under-expressed in AI *versus* NT samples in d25E. This result is consistent with reports that *POLG2* is involved mitochondrial DNA replication (which is important for embryo survival), and during normal development, the expression of *POLG2* increases dramatically during the blastocyst stage (Facucho-Oliveira *et al.* 2007). In 4 of 8 genes (**Table 4.3**), the opposition in expression signs occurred between studies d7E and d25E. This result suggests that the expression of some genes may be adjusted early in development to compensate for expression levels that were altered as a consequence of NT, allowing the embryo to survive longer in the pregnancy.

The percentage of genes differentially expressed with the same sign and overlapping among studies ranged from 0% to 11.6%, relative to the number of genes identified in one of the studies. The highest overlap of differentially expressed genes with the same sign was found in the comparison of the analysis of d280X against d25X. This result is in agreement with previous work that demonstrated aberrant gene expression in the extra-embryonic tissue of NT fetuses show aberrant gene expression from 25 d throughout gestation to term (Hashizume *et al.*, 2002; Hill *et al.*, 2002). This supports our finding that genes differentially expressed in d25X will remain differentially expressed in d280X. The percentage of genes differentially expressed with different signs and overlapping among the studies ranged from 0% to 10.5%, relative to the number of genes identified in one of the studies. Following the trends observed with absolute numbers, the

highest percentage of genes that had opposite signs was found between d7E and d25E (10.5%), whereas d25X had the lowest percentage of genes with opposite signs when compared to d280X. These findings suggest that the embryo may be more susceptible to NT than the extra-embryonic tissue and that the embryonic system is trying to compensate the effects of NT on gene expression at an earlier stage of development.

Comparison of results from the simple overlap of individual-studies and meta-analyses offered insights into the relative advantages of each approach. All 17 genes that were identified as differentially expressed in one or more individual studies, and in the study-level meta-analysis, had consistent signs across analyses. Study-level and sample-level meta-analyses were able to uncover genes involved in cell growth and proliferation (OLIGO_02307 [CN437097; polo-like kinase 4 (*PLK4*)]), DNA replication, recombination and repair (OLIGO_02116 [CV981642; structural maintenance of chromosomes 4-like-1 (*SMC4L1*)], OLIGO_07973 [BC102356; EBNA1 binding protein 2 (*EBNA1BP2*)], and OLIGO_03810 [CK977601; topoisomerase II binding protein 1 (*TOPBP1*)]), cellular assembly and organization (OLIGO_02116 [CV981642; *SMC4L1*], and OLIGO_10402 [BC105399; vesicle-associated membrane protein 3 (*VAMP3*)]), and cell cycle (OLIGO_02116 [CV981642; *SMC4L1*], OLIGO_03810 [CK977601; *TOPBP1*], and OLIGO_07973 [BC102356; *EBNA1BP2*]), that were identified by previous work (Everts *et al.*, 2008).

The loss in the total number of differentially expressed genes found from the simple overlap of studies and study-level meta-analysis could be expected because of the low number of studies (and thus experimental units) available to the study-level meta-analysis. An advantage of the study-level meta-analysis is that only summary statistics

(i.e. differential expression), and not the raw data, are necessary. In addition, the study-level meta-analysis incorporates the uncertainty of the differential expression from each study into the model (separately from differential expression). The consideration of the estimate and uncertainty of the differential expression separately supports a more revealing integration of information across studies than the combination of the estimate and uncertainty into one single measure (P-value). The additional insights result from the fact that a significant P-value can arise from different cases, including a high differential expression and moderate uncertainty and moderate differential expression and low uncertainty. Of the 78 genes that exhibited differential expression in an individual-study analysis and sample-level meta-analysis, the vast majority (71 genes) had a consistent pattern or sign across analyses. The few inconsistencies could be due to the fact that the meta-analysis considers information from multiple individual studies, and for example, a combination of 3 studies with insignificant but positive expression may outweigh one study with a significant but negative expression pattern.

4.4.1.3 Stage- and source- specific differential gene expression between AI and NT samples

The investigation of differential expression between AI and NT samples across developmental stages and sample sources considered in this study brings additional challenges to the typical meta-analysis. The challenges arise from the insights that can be gained not only from the consistent differential expression across studies but also from differential expression observed only at a specific stage or in a specific resource. The list of genes that were differentially expressed in only one study offers insights into differences in gene expression that are associated with the reproductive technologies that

are only evident at particular developmental stages or tissue resources (**Table 4.1**). The appearance (or disappearance) of differential expression between AI and NT samples across developmental stages may suggest that the effect of the reproductive technology compounds (or diffuses) with stage. The OLIGO_08251 [CB170385] was under-expressed in NT *versus* AI samples in d280X, however was not differentially expressed in d25X. This gene corresponds to the human gene cytochrome P450, family 11, subfamily A, polypeptide 1 (*CYP11A1*) and is associated with reproductive system diseases. Our finding is consistent with reports of the importance of this gene in placental progesterone synthesis (Hu *et al.*, 2004). The OLIGO_04004 [DT817577] was under-expressed in NT *versus* AI samples in d7E, but not differentially expressed in d25E. The gene is similar to the human gene additional sex combs like 1 (*ASXL1*), and is vital for the normal determination of segment identity in the developing embryo (Maglott *et al.*, 2005). Mutations in this gene are associated with myelodysplastic syndromes (Gelsi-Boyer *et al.*, 2009) that encompass ineffective blood production in the embryo and with reduced vascularization in placental tissues. These phenotypes may be associated with the frequency of large offspring syndrome (LOS) in fetuses produced by NT or with failed pregnancies.

Changes in the differential expression between AI and NT in embryo and extra-embryonic samples may suggest that one source is more susceptible to reproductive technologies than the other. The OLIGO_05509 [BM364003; similar to protein phosphatase 1, regulatory subunit 12C (*PPP1R12C)]* was over-expressed in NT *versus* AI samples in d7E and d25E, and was not differentially expressed in either d25X or d280X. BLAST (Altschul *et al.*, 1990) searches of the corresponding EST (GenBank

Acc: BM364003; GenBank gi: 18107372) against the human and cow databases produced no known gene match as of December 2009. This gives OLIGO_05509 a potential role in embryo development not previously known in the literature. In humans, protein phosphatase 1 (PP1) is involved in glycogen metabolism, calcium transport, smooth and skeletal muscle contraction, protein synthesis, and cell division (Bollen and Stalmans, 1992). Conversely, two other genes were identified as differentially expressed in extra-embryonic sources (d25X, d280X) and not in embryonic sources (d7E, d25E). The OLIGO_08173 [CV798714] was over-expressed in AI *versus* NT samples in both d25X and d280X, and corresponds to the human gene lymphocyte antigen 6 complex, locus G6C (*LY6G6C*). This gene is part of the Major Histocompatibility Complex III (MHC III) and is associated with susceptibility to autoimmune diseases (Mallya *et al.*, 2006). Under-expression of this gene among NT samples in the extra-embryonic tissue may suggest an importance to placental development, but its expression is not as crucial to the embryo, at least up to 25 d. The OLIGO_06132 [BC110250] was under-expressed in AI *versus* NT samples in both d25X and d280X, and corresponds to the human gene thyroid hormone receptor associated protein 6 (*THRAP6*). The gene has known functions in vitamin D receptor binding, transcription activator and coactivator activity, RNA polymerase II transcription mediator activity, and thyroid hormone receptor binding, and is involved in the androgen receptor signaling pathway and regulation of cellular transcription, among others (Maglott *et al.*, 2005).

4.4.1.4 Scenarios of meta-analysis of reproductive technologies

The comparison of results from the simple overlap of differentially expressed genes between reproductive technologies in individual-studies and meta-analyses also

offered insights into the relative advantages of each approach. All 17 genes that were identified as differentially expressed in one or more individual studies, and in the study-level meta-analysis, had consistent signs across analyses. Study-level and sample-level meta-analyses were able to uncover genes previously identified by Everts *et al.* (2008) that are involved in cell growth and proliferation (OLIGO_02307 [CN437097; polo-like kinase 4 (*PLK4*)]), DNA replication, recombination and repair (OLIGO_02116 [CV981642; structural maintenance of chromosomes 4-like-1 (*SMC4L1*)], OLIGO_07973 [BC102356; EBNA1 binding protein 2 (*EBNA1BP2*)], and OLIGO_03810 [CK977601; topoisomerase II binding protein 1 (*TOPBP1*)]), cellular assembly and organization (OLIGO_02116 [CV981642; *SMC4L1*], and OLIGO_10402 [BC105399; vesicle-associated membrane protein 3 (*VAMP3*)]), and cell cycle (OLIGO_02116 [CV981642; *SMC4L1*], OLIGO_03810 [CK977601; *TOPBP1*], and OLIGO_07973 [BC102356; *EBNA1BP2*]).

The loss in the total number of differentially expressed genes found from the simple overlap of studies and study-level meta-analysis could be expected because of the low number of studies (and thus experimental units) available to the study-level meta-analysis. An advantage of the study-level meta-analysis is that only summary statistics (i.e. differential expression), and not the raw data, are necessary. In addition, the study-level meta-analysis incorporates the uncertainty of the differential expression from each study into the model (separately from differential expression). The consideration of the estimate and uncertainty of the differential expression separately supports a more revealing integration of information across studies than the combination of the estimate and uncertainty into one single measure (P-value). The additional insights result from the

fact that a significant P-value can arise from different cases, including a high differential expression and moderate uncertainty and moderate differential expression and low uncertainty. Of the 78 genes that exhibited differential expression in an individual-study analysis and sample-level meta-analysis, the vast majority (71 genes) had a consistent pattern or sign across analyses. The few inconsistent signs between individual study and meta-analysis are due to the fact that the meta-analysis considers information from multiple individual studies, and for example, a combination of three studies with non-significant but positive expression can outweigh one study with a significant but negative expression pattern.

Funnel plots were used to typify the strength of the different types of approaches to integrate information across studies (individual-study, study-level and sample-level meta-analyses) for scenarios of particular statistical and biological relevance. **Figure 4.2** presents the estimates (and 95% confidence intervals) of differential expression between AI and NT samples corresponding to four gene expression scenarios. The estimate from the study-level meta-analysis appears distant from the rest because it is standardized and thus has no units as opposed to the other estimates that are not standardized. **Figure 4.2A** depicts a gene case where both the sample-level and study-level meta-analyses detected differential expression, with significant differential expression across reproductive techniques. However, none of the four individual-study analyses detected differential expression. This result suggests that individual studies lacked sufficient statistical power to detect differential expression but the model-based integration of information overcame this limitation. This gene (OLIGO_00155 [CR550800]) is similar to the human gene Deoxycytidylate deaminase (*DCTD*) (GO:0006220, pyrimidine nucleotide metabolic

process), which is an important factor that catalyzes the deamination of dCMP to dUMP (Weiner *et al.*, 1993). Inhibition of *DCTD* results in the inhibition of growth in chicken embryos (Roth *et al.*, 1963). **Figure 4.2B** depicts a gene (OLIGO_06762 [CX948597]) that exhibited differential expression in two individual-study analyses (d280X, d25E). Neither meta-analysis was able to detect an overall difference between AI and NT across studies for this gene because there was an inconsistent pattern of differential expression across the studies considered. This gene corresponds to the human mitochondrial fission regulator 1 (*MTFR1*) that protects cells from oxidative stress by separating the cell surroundings from the external environment (GO:0005886) (Westermann, 2002). This protein is associated with the mitochondrial inner membrane and induces mitochondrial fission; these events play a role in embryo development (Chan, 2006; Monticone *et al.*, 2007). **Figure 4.2C** depicts a gene (OLIGO_10402, [BC105399]) that only exhibited differential expression in the sample-level meta-analysis. In this case, neither the individual studies nor the study-level meta-analysis had the power to detect an overall differential expression because of study-level meta-analysis is only based on four observations. This gene represents the human gene, vesicle-associated membrane protein 3 (*VAMP3*) that is involved in cell migration and integrin trafficking (Luftman *et al.*, 2009). Cell migration is an essential process for embryonic development (Luftman *et al.*, 2009), as cells are responding to chemotaxis. Conversely, **Figure 4.2D** depicts a gene (OLIGO_02116, [CV981642]) that exhibits differential expression in the study-level meta-analysis, and not in the sample-level meta-analysis or any individual-study analysis. In this case, the modeling of study effects as fixed effects may have removed variation between reproductive technologies that was not affected by the incorporation of within-

study variation used in the study-level meta-analysis. This gene represents the human gene, structural maintenance of chromosomes 4-like 1 (*SMC4L1*) that is involved in protein binding (GO:0005515), nucleotide binding (GO:0000166) and ATP binding (GO:0046982). This gene was found to be over-expressed in fetal wounds relative to postnatal wounds (Colwell *et al.*, 2008).

4.4.1.5 Biological processes, molecular functions and gene pathways associated with differential expression

Gene Pathways (Kanehisa and Goto, 2000), and Gene Ontology (Gene Ontology Consortium, 2000) molecular functions, biological processes and cellular components that were enriched among genes over- or under-expressed in AI *versus* NT from the sample-level, study-level meta-analysis and individual-studies were uncovered using GSEA. The sample-level meta-analysis revealed that the molecular function transferase activity (GO:0016772, P-value < 0.0008) was overrepresented among genes over-expressed in AI *vs* NT. Genes that have been assigned to this function are involved in the transfer of phosphorus-containing groups between compounds (The Gene Ontology Consortium) and include OLIGO_07021 [AY781100; v-akt murine thymoma viral oncogene homolog 1 (*AKT1*)], OLIGO_08080 [BC110170; *CSNK2B*], OLIGO_08124 [CN440810; polymerase (DNA directed), gamma 2, accessory subunit (*POLG2*)], OLIGO_03960 [DV924388; STK24], OLIGO_04559 [BF039418; activin A receptor, type IIA (*ACVR2A*)], OLIGO_03891 [AW465883; *WDR33*], OLIGO_12731 [CK770512; *PRIM2A*], and OLIGO_02324 [CN437126; *CRLS1*]. A number of these genes were found in the literature to be vital to the development of the embryo. Gene *AKT1* was over-expressed in AI *versus* NT samples. Previous work has found the gene to be

associated with placental development and fetal growth, where deficiencies in expression can lead to adverse effects on development and growth (Yang *et al.*, 2003). Gene *ACVR2A* was over-expressed in AI *versus* NT samples, and is suggested to have multiple roles in murine organogenesis, including inner ear, tongue and lung development (Feijen *et al.*, 1994). Gene *POLG2* has associated function during the blastocyst stage, specifically for mitochondrial DNA transcription and replication (Lloyd *et al.*, 2009).

Several Gene Ontology processes critical in embryo development were over-represented among the genes differentially expressed between AI and NT samples across or at specific stages and sources. Cell division (P-value < 0.0013, over-represented among genes with expression in NT > AI), mitotic cell cycle (P-value < 0.0005, over-represented among genes with expression in NT > AI), regulation of cell shape (P-value < 0.0006, over-represented among genes with expression in AI > NT), M phase of mitotic cell cycle (P-value < 0.0006, over-represented among genes with expression in NT > AI) and regulation of anatomical structure morphogenesis (P-value < 0.0006, over-represented among genes with expression in AI > NT) were over-represented among the differentially expressed genes in the d280X study. The cellular components differentially represented in d280X were integral to plasma membrane (P-value < 0.0006, over-represented among genes with expression in AI > NT), intracellular organelle part (P-value < 0.0013, over-represented among genes with expression in NT > AI), intracellular membrane-bound organelle (P-value < 0.0038, over-represented among genes with expression in NT > AI) and ruffle (P-value < 0.0009, over-represented among genes with expression in AI > NT); in d25E membrane (P-value < 0.0001, over-represented among genes with expression in AI > NT); in study-level meta-analysis ribosome (P-value <

0.0065, over-represented among genes with expression in AI > NT); and in sample-level meta-analysis membrane-bound vesicle (P-value < 0.0069, over-represented among genes with expression in AI > NT) and organelle lumen (P-value < 0.0075, over-represented among genes with expression in AI > NT).

4.4.2 Meta-analysis of stages, sources and technologies within conditions

4.4.2.1 Differential expression between stages and sources within reproductive technology

The second specification of the sample-level meta-analysis enabled the identification of differentially expressed genes between d7E and d25E samples, d25X and d280X samples, and d25E and d25X samples within AI and NT samples (**Table 4.4**). **Appendices D to G** present the gene identifiers, estimates of difference in expression between conditions and P-values resulting from all contrasts. The highest number of differentially expressed genes within reproductive technology was in the comparison of 7 d and 25 d NT embryos (1,149 genes). The number of differentially expressed genes within technology between 25 d and 280 d extra-embryonic tissue were the same (AI = 340, NT = 334). The fewest differentially expressed genes were found when comparing tissue sources within reproductive technology and developmental stage and comparing reproductive technologies within tissue source and developmental stage, relative to reproductive technology and developmental stage comparisons. These findings have multiple potential implications. First, changes in development stage may be associated with higher or more wide-spread changes in gene expression than changes in reproductive technology or sample source. Second, similar changes in gene expression

associated with developmental stage or reproductive technology are observed in embryonic and extra-embryonic sources. Third, the difference between reproductive technologies on reproductive efficiency and the occurrence of LOS may be associated with significant differences in the expression of a few genes and less significant differences in the expression of the reminder genes.

The overlap of genes differentially expressed between 7 d and 25 d samples, separate for AI (483 genes) and NT (1149 genes) resulted in 373 bovine genes corresponding to 372 human annotations). Gene set enrichment analysis uncovered numerous biological processes relevant to embryo development over-represented among these genes, including anatomical structure development (P-value < 7×10^{-5}, over-represented in genes with expression at 7 d > 25 d), multicellular organismal development (P-value < 3×10^{-4}, over- represented in genes with expression at 7 d > 25 d), system development (P-value < 7×10^{-5}, over- represented in genes with expression at 7 d > 25 d), and organ development (P-value < 2×10^{-4}, over- represented in genes with expression at 7 d > 25 d).

Among the genes differentially expressed across developmental stages within reproductive technology, OLIGO_07962 [BT021073], homologous to the human gene peroxiredoxin 1 (*PRDX1*), has been associated with developmental competence in bovine embryos, and is involved in antioxidant defenses (Mourot *et al.* 2006). Gene *PRDX1* was over-expressed in 7 d versus 25 d in AI embryos, however was under-expressed in 7 d versus 25 d in NT embryos. This suggests that under-expression of the gene at 7 d NT embryos may be associated with developmental problems in NT relative to AI embryos. Similarly, OLIGO_12806 [BC104598], homologous to human gene *SLC16A1*, was over-

expressed in 7 d versus 25 d AI embryos, and under-expressed in 7 d versus 25 d NT embryos. Somers *et al.* (2006) also found *SLC16A1* to be under-expressed in NT relative to non-NT embryos, and suggested that a reduced expression of the gene, specifically during the early blastocyst stage, may affect the embryo's viability.

When considering the overlap of genes differentially expressed between 25 d and 280 d extra-embryonic sources in AI and NT samples, OLIGO_10954 [AJ580584] homologous to human gene major histocompatibility complex, class II, DQ beta 1 (*HLA-DQB1*), was significantly under-expressed in 25 d versus 280 d, regardless of the reproductive technology. This result is consistent with reports that this gene is expressed in the trophoblast (outer layer cells of the blastocyst that will form the placenta) of cloned pregnancies after the fifth month of pregnancy (Hill *et al.* 2002). Expression of *HLA-DQB1* within cattle trophoblast tissue has not been found during the first trimester pregnancies, as is supported by the lower expression of *HLA-DQB1* in early extra-embryonic tissue (25 d) compared to *HLA-DQB1* expression in tissue later in pregnancy (280 d) regardless of the reproductive technology used.

4.4.2.2 Differential expression between reproductive technologies

Additional contrasts that considered the expression of AI compared to NT samples regardless of developmental stage and tissue source were analyzed. Results from the analysis of the contrasts within source and stage revealed 16, 4, 1 and 21 differentially expressed genes for AI *versus* NT samples within embryonic 7 d, embryonic 25 d, extra-embryonic 25 d, and extra-embryonic 280 d samples, respectively (**Table 4.4**). For AI *versus* NT in embryonic samples regardless of developmental stage (d7E, d25E), there were 20 differentially expressed genes, of which 7 were over-

expressed in AI *versus* NT samples, and 13 were over-expressed in NT *versus* AI samples. The OLIGO_05064 [Hs.517517] represents human gene E1A binding protein p300 (*EP300*), and was over-expressed in NT *versus* AI samples. This gene is involved in nervous system development, homeostasis, and apoptosis (Bundy *et al.*, 2006; Byun *et al.*, 2009).

For AI *versus* NT in extra-embryonic samples regardless of developmental stage (d25X, d280X), 22 differentially expressed genes were identified, of which 2 were over-expressed in AI *versus* NT samples, and the remaining 20 were over-expressed in NT *versus* AI samples. The OLIGO_12229 [M25648] is a gene that attracts interest in these contrasts, representing human gene oxytocin (*OXT*). Oxytocin is involved in smooth muscle contractions during parturition and lactation, and in maternal behavior and cardiovascular function (Gimpl and Fahrenholz, 2001). The gene was over-expressed in NT *versus* AI extra-embryonic samples at 280 d. Previous reports have found that large amounts of *OXT* are produced by the corpus luteum when too much prostaglandin (*PGF2a*) is present, decreasing embryonic survival (Seals *et al.*, 1998). Prostaglandin increases during times of stress usually encountered through embryos created using assisted reproductive techniques (Schallenberger *et al.*, 1989). As expression of *OXT* was higher in NT samples, this gene becomes a good candidate for further analysis in the identification of genes affecting embryo survival in terms of NT.

For the contrast of AI *versus* NT at 25 d regardless of the sample source, 39 differentially expressed genes were identified, of which 34 were over-expressed in AI *versus* NT samples. The OLIGO_08092 [DV907192] represents human gene dickkopf homolog 3 (*DKK3*), and was over-expressed in AI *versus* NT samples. The gene is key

in the *Wnt* signal transduction pathway that participates in early embryonic development (Kleber and Sommer, 2004). Under-expression of the gene in NT samples may suggest a delay in development of embryos produced via nuclear transfer methods. Analysis of other genes with similar functions would be beneficial in understanding expressional differences in embryos produced using differing reproductive techniques.

Variable combinations can be compared in terms of three separate scenarios, outlined in **Figure 4.3**, where venn diagrams illustrate genes overlapping between combinations within the scenarios. **Figures 4.3 (a)** and **(b)** illustrate large overlap between AI and NT embryonic contrasts between 7 d and 25 d, and between AI and NT extra-embryonic contrasts between 25 d and 280 d, respectively. Numerous genes overlap between all three comparisons for each, with 211 and 231 DEG in common between all three comparisons in **(a)** and **(b)**, respectively. For example, 271 genes are differentially expressed between 7 d (blastocysts) and 25 d (embryo), regardless of reproductive technology. However, the majority of these genes (211) have a difference in the pattern of differential expression across ages in AI *versus* NT. For the remaining 60 genes, the pattern of across developmental stages does not seem to be affected by the type of reproductive technology used. Thus, at early developmental stages the difference in efficiency between the two technologies is likely not to be due to these 60 genes. However, at 25 d in **Figure 4.3 (c)**, a majority of the genes (28) were differentially expressed in AI and not in NT samples, and no genes overlapped between the contrast of NT embryonic *versus* extra-embryonic samples at 25 d and the other two comparisons (AI embryonic *versus* NT extra-embryonic samples at 25 d, and AI embryonic *versus* AI extra-embryonic samples at 25 d). This would suggest that there are genes sensitive to

reproductive technology, and in this case, expression is limited to samples created using the AI technique.

Study of the DEG within and across sources, developmental stages, and reproductive technologies helps uncover genes vital to embryo development. DEG across developmental stage regardless of technique, but the sign of differential expression would add additional insight into whether it is over- or under-expressed in 7 d relative to 25 d, or 25 d relative to 280 d of gestation. The OLIGO_06930 [AF031698], corresponding to human gene Prostoglandin-endoperoxide synthase 2 (*PTGS2*), was over-expressed (positive estimate values) in the contrasts AI extra-embryonic samples at 25 d *versus* 280 d and NT extra-embryonic samples at 25 d *versus* 280 d, which suggests the gene is associated with expression during 25 d gestation, and expression decreases towards term. Without the direction of differential expression, knowing the importance of the gene at a specific developmental stage is unknown. Similarly, study of DEG found solely within a single contrast (e.g. AI *versus* NT at 25 d or AI *versus* NT at 280 d) may provide insight into genes associated with a certain technique or stage. **Figure 4.3 (c)** has a large overlap between the three comparisons between extra-embryonic and embryonic samples at 25 d (AI, NT, and AI *versus* NT) relative to the total number of genes significant for each comparison. The contrast between NT embryonic and extra-embryonic at 25 d identified only 11 DEG, of which the majority (8) were also DEG in AI and in AI *versus* NT samples. Although the 8 genes were differentially expressed in NT and AI, these genes exhibit differential patterns (i.e. different signs) in NT and AI and could be associated with the different impact of the reproductive technologies on efficiency.

Of the DEG between embryonic and extra-embryonic at day 25 in AI relative to NT (39), the vast majority were differentially expressed in AI. This result suggests that most of the 39 genes stem from being differentially expressed in embryonic *versus* extra-embryonic samples in AI, but are not differentially expressed in NT. The lack of differential expression between embryonic and extra-embryonic samples at 25 d of the 28 genes in NT, although differential in AI, may be partially responsible for the differences in efficiency between AI and NT. Likewise, the contrast of AI embryonic *versus* NT extra-embryonic samples at 25 d identified 39 DEG, of which 28 were also significant in the contrast of AI embryonic *versus* AI extra-embryonic samples at 25 d, and 8 were significant in both remaining comparisons. Similarly to the previous discussion, this suggests that reproductive efficiency is affected by the technology used to produce the embryos, as no genes overlapped between NT and AI or NT and AI *versus* NT comparisons.

To exemplify the additional insights obtained from the hierarchical meta-analysis of partially-overlapping experiments particular examples are discussed. **Figure 4.4** depicts the changes in expression of the probe corresponding to human gene prostaglandin D2 synthase 21kDa (*PTGDS*, probe OLIGO_08018) for all tissue and stage combinations within reproductive technique. This gene is vital for prostaglandin synthesis (Begemann *et al.*, 2008) and has functions in metabolic homeostasis and sleep regulation (O'Hara *et al.*, 1999). Prostaglandin D2 has been found to be an inhibitor of platelet aggregation in vitro in human pregnancies, and over- and under-expression of the enzyme in amniotic fluids of fetuses has been linked to several abnormalities (Melegos *et al.*, 1996). During normal pregnancy, *PTGDS* concentration increases dramatically

during weeks 12-25 of gestation, and then gradually declines until term (Melegos *et al.*, 1996). **Figure 4.4** shows the changes in *PTGDS* expression between 7 d and 25 d in embryos as a period of potential interest to biologists, as expression levels vary greatly for AI (blue) *versus* NT (green) embryos. Further analysis of genes identified through the meta-analysis may identify key components in the ambiguous problem of reproductive inefficiency.

4.5 Conclusions

Insights into the genes and functional categories that may be responsible for embryo development and differences in efficiency between reproductive technologies in cattle were gained by meta-analyzing four microarray gene expression studies. The complementary meta-analytical approaches revealed genes consistently differentially expressed between reproductive technologies across stages and sources, genes exhibiting differential expression at selected stages and sources and genes exhibiting differential expression across stages or sources within reproductive technology. Results from this novel extension to previous meta-analyses (Adams *et al.*, 2008) allowed the characterization of comprehensive expression profiles that extend conditions not studied within a single experiment.

The identification of differentially expressed genes between AI and NT across all stages and sources by model-based meta-analyses that were not detected by the overlap of gene lists from individual study analyses highlights the advantage of meta-analytical approaches. Furthermore, these results offer additional understanding of the genes with differential expression and functional categories associated with the reproductive

technology used. The highest percentage of genes that had opposite signs (reverse fold change between AI and NT) was found between embryos at 7 d and 25 d whereas extra-embryonic samples at 25 d had the lowest percentage of genes with opposite signs when compared to d280X. These findings suggest that the embryo may be more susceptible to reproductive technologies than the extra-embryonic tissue and that the embryonic system is trying to compensate the effects of the technology on gene expression from early stages. The number of differentially expressed genes within each technology and between the extra-embryonic tissue source on 25 d and 280 d were the same (AI = 340, NT = 334), indicating that most of the gene expression compensation happens early in development, and that only embryos with more "normal" extra-embryonic tissue gene expression patterns develop to 25 d. The change in magnitude of differential expression between AI and NT samples across developmental stages may suggest that the effect of the reproductive technology varies with developmental stage. There was no particular sign or pattern of differential expression that stood out across all analyses, except for study d280X, that had 189 and 657 positively and negatively differentially expressed genes, respectively. This finding suggests that the expression of genes in the placenta is down-regulated with NT relative to AI towards the end stages of fetal development.

Genes that are consistently differentially expressed between technologies across stages and sources may not be critical to explain the differences in efficiency between technologies because these differences did not affected the survival of the embryo at early or late stages of the pregnancy. Genes that are only differentially expressed between technologies at particular stages (*CYP11A1*, *ASXL1*) or sources (*LY6G6C*, *THRAP6*) may be associated with pathways or processes that cannot achieve equilibrium,

or become unstable during development and are responsible for the differences in efficiency. Similarly, genes that are differentially expressed across stages or sources regardless of technology may not be responsible for the low efficiency of NT relative to AI. Those genes that are differentially expressed across stages or sources within technology (*PRDX1*, *SLC16A1*, *HLA-DQB1*) offer additional insights into embryo development.

Gene enrichment functional and pathway analyses of the meta-analysis results were undertaken to uncover biological processes and gene pathways that are key to the differences between the assisted reproductive technologies. The sample-level meta-analysis revealed that transferase activity was overrepresented among differentially expressed genes between AI and NT across all studies, and numerous genes (*AKT1*, *CSNK2B*, *POLG2*, *STK24*, *ACVR2A*, *WDR33*, *PRIM2A*, and *CRLS1*) having this function have been associated with embryo development. We hypothesize that at early stages of development, the embryo system may make corrections to the changes in gene expression imposed by the NT reproductive technology. These corrections allow the embryo to survive until at least day 25, based on the highest number of differentially expressed genes found within reproductive technology was in the comparison of 7 d and 25 d NT samples. Also, the lowest number of differentially expressed genes was found when comparing sample sources within reproductive technology (82 genes in AI *vs* 11 genes in NT) and developmental stage (16 genes in 7d embryos, 4 in 25d embryos, 1 in 25 d extra-embryonic tissue, and 21 at 280d extra-embryonic tissue), relative to comparisons of reproductive technology and stages (530 genes in embryos, 551 genes within extra-embryonic tissue).

The meta-analysis identified novel associations between gene expression and conditions and confirmed others such as *HLA-DQB1* that had similar under expression in 25 d relative to 280 d for extra-embryonic sources in AI and NT samples, *SLC16A1* that was over-expressed in 7 d *versus* 25 d AI embryos, and under-expressed in 7 d *versus* 25 d NT embryos and *PRDX1* that was over-expressed in 7 d *versus* 25 d in AI embryos and under-expressed in 7 d *versus* 25 d in NT embryos. The biological processes identified enhance the understanding of embryo development and efficiency of reproductive technologies and the genes identified are candidates for empirical confirmation.

Conclusions stemming from our meta-analysis include: 1) developmental stage has more influence on gene expression than reproductive technology or sample source; 2) the embryonic and extra-embryonic sources have a similar response to changes in developmental stage or reproductive technology; and 3) the effect of reproductive technologies on efficiency and LOS may rely on a large impact (modification in expression) on a few genes and lower (less significant) impact on other genes.

References

Abdi H: **Partial least squares regression**. In *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage; 2007: 740-744

Adams HA, Southey BR, Robinson GE, Rodriguez-Zas SL: **Meta-analysis of genome-wide expression patterns associated with behavioral maturation in honey bees**. *BMC Genomics* 2008, **9**: 503-517

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**: 2185-2195

Agresti A: *An introduction to categorical data analysis, 2nd ed*. Wiley, New York, 2007

Alaux C, Sinha S, Hasadsri L, Hunt GJ, Guzman-Novoa E, DeGrandi-Hoffman G, Uribe-Rubio JL, Southey BR, Rodriguez-Zas S, Robinson GE: **Honey bee aggression supports a link between gene regulation and behavioral evolution**. *PNAS* 2009, **106(36)**: 15400-15405

Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nat Rev Genet* 2006, **7**: 55-65

Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JMM, Conde L, Blaschke C, Vera J Dopazo J: **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments**. *Nucl Acids Res* 2006, **34:** W472-W476

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**: 403-410

Amdam GV, Norberg K, Page JRE, Erber J, Scheiner R: **Downregulation of vitellogenin gene activity increases the gustatory responsiveness of honey bee workers (*Apis mellifera*)**. *Behav Brain Res* 2006, **169**: 201-205

Anderson GB, Seide GE: **Cloning for profit**. *Science* 1998, **280**: 1400-1401

Arias MC, Sheppard WS: **Phylogenetic relationships of honey bees (Hymenoptera:Apinae:Apini) inferred from nuclear and mitochondrial DNA sequence data.** *Mol Phylogenet Evol* 2005, **37:** 25-35

Assou S, Le Carrour T, Tondeur S, Strom S, Gabelle A, Marty S, Nadal L, Pantesco V, Reme T, Hugnot JP, Gasca S, Hovatta O, Hamamah S, Klein B, De Vos J: **A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas.** *Stem Cells* 2004, **25:** 961-973

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles – database and tools update**. *Nucl Acids Res* 2007, **35:** D760-D765

Beehive. 2006. [http://stagbeetle.animal.uiuc.edu/Beehive/]

Begemann M, Sargin D, Rossner MJ, Bartels C, Theis F, Wichert SP, Stender N, Fischer B, Sperling S, Stawicki S, Wiedl A, Falkai P, Nave KA, Ehrenreich H: **Episode-specific differential gene expression of peripheral blood mononuclear cells in rapid cycling supports novel treatment approaches**. *Mol Med* 2008, **14**: 546-552

Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing**. *J R Stat Soc Ser B* 1995, **57**: 289-300

Beyhan Z, Ross PJ, Iager AE, Kocabas AM, Cunniff K, Rosa GJ, Cibelli JB: **Transcriptional reprogramming of somatic cell nuclei during preimplantation development of cloned bovine embryos**. *Dev Biol* 2007*,* **305**: 637-649

Bollen M, Stalmans W: **The structure, role, and regulation of type 1 protein phosphatases**. *Crit Rev Biochem Mol Biol* 1992, **27(3)**: 227-281

Bovine Genome Sequencing Consortium (BGSC) at Baylor College of Medicine: **Bovine genome project**. 2007. Accessed: August 2008. [http://www.hgsc.bcm.tmc.edu/projects/bovine/]

Bing Y, Ouellette RJ: **Fertilization in vitro**. *Methods Mol Biol* 2009, **550**: 251-266

Brazma A, Parkinson H, Schlitt T, Shojatalab M: **Microarray Databases**. *European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI)*. Accessed: 2008 August. [http://www.ebi.ac.uk/2can/databases/microarray.html]

Bundy JG, Iyer G, Gentile MS, Hu DE, Kettunen M, Maia AT, Thorne NP, Brenton JD, Caldas C, Brindle KM: **Metabolic consequences of p300 gene deletion in human colon cancer cells**. *Cancer Res* 2006, **66**: 7606-7614

Byun JS, Wong MM, Cui W, Idelman G, Li Q, De Siervi A, Bilke S, Haggerty CM, Player A, Wang YH, Thirman MJ, Kaberlein JJ, Petrovas C, Koup RA, Longo D, Ozato K, Gardner K: **Dynamic bookmarking of primary response genes by p300 and RNA polymerase II complexes**. *PNAS* 2009, **106**: 19286-19291

Cash AC, Whitfield CW, Ismail N, Robinson GE: **Behavior and the limits of genomic plasticity: power and replicability in microarray analysis of honeybee brains.** *Genes Brain Behav* 2005, **4:** 267-271

Chan DC: **Mitochondria: dynamic organelles in disease, aging, and development**. *Cell* 2006, **125**: 1241-1252

Chavatte-Palmer P, Heyman Y, Richard C, Monget P, Lebourhis D, Kann G, Chilliard Y, Vignon X, Renard JP: **Clinical, hormonal, and hematologic characteristics of bovine calves derived from nuclei from somatic cells**. *Biol Reprod* 2002, **66**: 1596-1603

Check E: **From hive minds to humans**. *Nature* 2006, **443**: 893

Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling inter-study variation.** *Bioinformatics* 2003, **19:** i84–i90

Colwell AS, Longaker MT, Peter Lorenz H: **Identification of differentially regulated genes in fetal wounds during regenerative repair**. *Wound Repair Regen* 2008, **16**: 450-459

Conlon EM, Song JJ, Liu A: **Bayesian meta-analysis models for microarray data: a comparative study.** *BMC Bioinformatics* 2007, **8:** 80-100

Cui X, Kerr MK, Churchill GA: **Transformations of cDNA microarray data.** *Stat Appl Genet Mol Biol* 2003, **2:** Article 4

DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R: **Combining results of microarray experiments: a rank aggregation approach.** *Stat Appl Genet Mol Biol* 2006, **5:** Article 15

Denison R, Raymond-Delpech V: **Insights into the molecular basis of social behavior from studies on the honeybee, *Apis mellifera*.** *Invert Neurosci* 2008, **8:** 1-9

Dieleman SJ, Hendriksen PJM, Viuff D, Thomsen PD, Hyttel P, Knijn HM, Wrenzycki C, Kruip TA, Niemann H, Gadella BM, Bevers MM, Vos PL: **Effects of *in vivo* prematuration and *in vivo* final maturation on developmental capacity and quality of pre-implantation embryos**. *Theriogenology* 2002, **57**: 5-20

Dinnyes A, Tian XC, Yang X: **Epigenetic regulation of foetal development in nuclear transfer animal models**. *Reprod Domest Anim* 2008, **43**: 302-309

Edwards JL, Schrick FN, McCracken MD, van Amstel SR, Hopkins FM, Welborn MG, Davies CJ: **Cloning adult farm animals: a review of the possibilities and problems associated with somatic cell nuclear transfer**. *Am J Reprod Immunol 2003,* **50**: 113–123

Egger M, Smith GD, Phillips AN: **Meta-analysis: principles and procedures**. *BMJ* 1997, **315**: 1533-1537

Everts RE, Band MR, Liu ZL, Kumar CG, Liu L, Loor JJ, Oliveira R, Lewin HA: **A 7872 cDNA microarray and its use in bovine functional genomics**. *Vet Immunol Immunopathol* 2005, **105:** 235-245

Everts RE, Sommers A, Green CA, Oliveira R, Rodriguez-Zas SL, Sung L-Y, Du F, Evans A, Boland M, Fair T, Lonergan P, Renard J-P, Yang X, Tian X, Lewin HA: **Major differences in gene expression profiles revealed in day-25 placental tissues collected from cows carrying cloned fetuses**. *Plant & Animal Genomes XV Conference San Diego, CA* 2007a [http://www.intl-pag.org/pag/15/abstracts/PAG15_P05k_520.html]

Everts RE, Sommers A, Green CA, Oliveira R, Rodriguez-Zas SL, Sung LY, Du F, Evans ACO, Boland M, Fair T, Lonergan P, Renard JP, Yang X, Tian X, Lewin HA: **Major differences in gene expression profiles revealed in day-25 trophoblast but not embryonic discs collected from cattle clones**. *2nd International Meeting on Mammalian Embryogenomics Conference Abstract* 2007b, Paris, France

Everts RE, Chavatte-Palmer P, Razzak A, Hue I, Green CA, Oliveira R, Vignon X, Rodriguez-Zas SL, Tian C, Yang X, Renard J-P, Lewin HA: **Aberrant gene expression patterns in placentomes are associated with phenotypically normal and abnormal cattle cloned by somatic cell nuclear transfer**. *Physiol Genomics* 2008, **33**: 65-77

Facucho-Oliveira JM, Alderson J, Spikings EC, Egginton S, St. John JC: **Mitochondrial DNA replication during differentiation of murine embryonic stem cells**. *J Cell Sci* 2007, **120**: 4025-4034

Farin PW, Farin CE: **Transfer of bovine embryos produced in vivo or in vitro: survival and fetal development**. *Biol Reprod* 1995, **52**: 676-682

Feijen A, Goumans MJ, van den Eijnden-van Raaij AJM: **Expression of activin subunits, activin receptors and follistatin in postimplantation mouse embryos suggests specific developmental functions for different activins**. *Development* 1994, **120**: 3621-3637

Feldhaar H, Gross R: **Immune reactions of insects on bacterial pathogens and mutualists**. *Microbes Infect* 2008, **10**: 1082-1088

Fishel I, Kaufman A, Ruppin E: **Meta-Analysis of Gene Expression Data: A Predictor-Based Approach.** *Bioinformatics* 2007, **23:** 1599-1606

Fort G, Lambert-Lacroix S: **Classification using partial least squares with penalized logistic regression**. *Bioinformatics* 2005, **21**: 1104-1111

Fry SN, Wehner R: **Honeybees store landmarks in an egocentric frame of reference**. *J Comp Physiol A* 2002, **187**: 1009-1016

Furnival GM, Wilson RW: **Regression by Leaps and Bounds**. *Technometrics* 1974, **16**: 499–511

Gelsi-Boyer V, Trouplin V, Adelaide J, Bonansea J, Cervera N, Carbuccia N, Lagarde A, Prebet T, Nezri M, Sainty D, Olschwang S, Xerri L, Chaffanet M, Mozziconacci MJ, Vey N, Birnbaum D: **Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia**. *Br J Haematol* 2009, **145**: 788-800

The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology.** *Nature Genet* 2000, **25:** 25-29

GenePix Pro 5.0 [http://www.moleculardevices.com]

Ghosh D Barette TR, Rhodes D, Chinnaiyan AM: **Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer**. *Funct Integr Genomics* 2003, **3**: 180-188

Gill C: **Honey bee die-off alarms beekeepers, crop growers and researchers.** *Penn State College of Agricultural Sciences News Releases* 2007, [http://aginfo.psu.edu/news/2007/1/HoneyBees.htm]

Gimpl G, Fahrenholz F: **The oxytocin receptor system: structure, function, and regulation**. *Physiol Rev* 2001, **81**: 629-683

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bicklhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley Jr RL, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of *Drosophila melanogaster*.** *Science* 2003, **302:** 1727-1736

Grozinger CM, Fan Y, Hoover SE, Winston ML: **Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*)**. *Mol Ecol* 2007, **16**: 4837-4848

Hashizume K, Ishiwata H, Kizaki K, Yamada O, Takahashi T, Imai K, Patel OV, Akagi S, Shimizu M, Takahashi S, Katsuma S, Shiojima S, Hirasawa A, Tsujimoto G, Todoroki J, Izaike Y: **Implantation and placental development in somatic cell clone recipient cows**. *Cloning Stem Cells* 2002, **4**: 197-209

Hastie T, Tibshirani R, Friedman J: **Model Assessment and Selection**. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. USA: Springer; 2001: 193-221

Hedges LV: **Meta-analysis**. *Journal of Educational and Behavioral Statistics* 1992, **17**: 279-296

Heyman Y, Chavatte-Palmer P, Lebourhis D, Camous S, Vignon X, Renard JP: **Frequency and occurrence of late-gestation losses from cattle cloned embryos**. *Biol Reprod* 2002, **66**: 6-13

Hill JR, Schlafer DH, Fisher PJ, Davies CJ: **Abnormal expression of trophoblast major histocompatibility complex class I antigens in cloned bovine pregnancies is associated with a pronounced endometrial lymphocytic response**. *Biol Reprod* 2002, **67**: 55-63

Hosmer DW, Lemeshow S: *Applied regression analysis*. Wiley, New York, 1989

Honey Bee Genome Sequencing Consortium: **Insights into social insects from the genome of the honeybee Apis mellifera**. *Nature* 2006, **443**: 931-949

Hong F, Breitling R: **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments**. *Bioinformatics* 2008, **24**: 374-382

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics* 2006, **22:** 2825-2827

Hoy M: **DNA Amplification by the Polymerase Chain Reaction: Molecular Biology Made Accessible**. In *Insect Molecular Genetics: An Introduction to Principles and Applications*. Second Edition. Edited by Hoy M. USA: Elsevier Science; 2003: 206

Hu MC, Jsu HJ, Guo IC, Chung B: **Function of Cyp11a1 in animal models**. *Mol Cell Endocrinol* 2004, **215**: 95-100

Hu P, Greenwood CM, Beyene J: **Statistical methods for meta-analysis of microarray data: a comparative study**. *Inf Syst Front* 2006, **8**: 9-20

Huang Z: **Beetography**. 2007. [http://www.beetography.com/]

Human Genome Sequencing Center (HGSC) at Baylor College of Medicine: **Honey bee genome project**. 2005. Accessed: August 2008. [http://www.hgsc.bcm.tmc.edu/projects/honeybee/]

Humpherys D, Eggan K, Akutsu H, Friedman A, Hochedlinger K, Yanagimachi R, Lander ES, Golub TR, Jaenisch R: **Abnormal gene expression in cloned mice derived from embryonic stem cell and cumulus cell nuclei**. *PNAS* 2002, **99**: 12889-12894

Hunter JE, Schmidt FL: **Methods of Meta-analysis: Correcting error and bias in research findings**. 1990a, *Beverly Hills, CA: Sage*

Hunter JE, Schmidt FL: **Dichotomization of continuous variables: the implications for meta-analysis**. *J Appl Psychol* 1990b, **75**: 334-349

Hunter JE, Schmidt FL: **Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge**. *Int J Select Assess* 2000, **8**: 275-292

International Human Genome Sequencing Consortium (IHGSC): **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**: 931–945

Johnson RA, Wichern DW: **Principal Components**. *Applied Multivariate Statistical Analysis*. Sixth Edition. USA: Pearson Education, Inc.; 2007: 430-480

Kaplan K: **Colony collapse disorder: a complex buzz**. *Agricultural Research Magazine* 2008, [http://www.ars.usda.gov/is.AR/archive/may08/colony0508.htm]

Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucl Acids Res* 2000, **28**: 27-30

Kerr MK, Afshari CA, Bennett L, Bushel B, Martinez J, Walker NJ, Churchill GA: **Statistical analysis of a gene expression microarray experiment with replication**. *Stat Sinica* 2002, **12:** 203-217

Kimball JW: **Genome sizes**. Online *Biology* textbook, 2008. Accessed: August 2008. [http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html]

Kleber M, Sommer L: **Wnt signaling and the regulation of stem cell function**. *Cell Biol* 2004, **16**: 681-687

Koywiwattrakul P, Sittipraneed S: **Expression of vitellogenin and transferrin in activated ovaries of worker honey bees, *Apis mellifera***. *Biochem Genet* 2009, **47**: 19-26

Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonak J, Lind K, Sindelka R, Sjoback R, Sjogreen J, Strombom L, Stahlberg A, Zoric N: **The real-time polymerase chain reaction**. *Mol Aspects Med* 2006, **27**: 95-125

Kues WA, Sudheer S, Herrmann D, Carnwath JW, Havlicek V, Besenfelder U, Lehrach H, Adjaye J, Niemann H: **Genome-wide expression profiling reveals distinct clusters of transcriptional regulation during bovine preimplantation development in vivo**. *PNAS* 2008, **105**: 19768-19773

Lee Y, Lee C: **Classification of multiple cancer types by multicategory support vector machines using gene expression data**. *Bioinformatics* 2003, **19**: 1132-1139

Li W, Yang Y: **How many genes are needed for a discriminant microarray data analysis**. In *Methods of Microarray Data Analysis: Papers from CAMDA'00* 2002. Accessed online: [http://www.nslij-genetics.org/wli/pub/camda00.pdf]

Liao JG, Chin K-V: **Logistic regression for disease classification using microarray data: model selection in a large p and small n case**. *Bioinformatics* 2007, **23**: 1945-1951

Lipsey MW, Wilson DB: **Practical meta-analysis**. *Applied Social Research Methods Series*. Volume 49. USA: SAGE Publications, Inc., 2001

Liu Y, Rayens W: **PLS and dimension reduction for classification**. *Comput Stats* 2007, **22**: 189-208

Lloyd RE, Romar R, Matas C, Gutierrez-Adan A, Holt WV, Coy P: **Effects of oviductal fluid on the development, quality, and gene expression of porcine blastocysts produced in vitro**. *Reprod* 2009, **137**: 679-687

Loor JJ, Dann HM, Guretzky NA, Everts RE, Oliveira R, Green CA, Litherland NB, Rodriguez-Zas SL, Lewin HA, Drackley JK: **Plane of nutrition prepartum alters hepatic gene expression and function in dairy cows as assessed by longitudinal transcript and metabolic profiling**. *Physiol Genomics* 2006, **27**: 29-41

Loor JJ, Everts RE, Bionaz M, Dann HM, Morin DE, Oliveira R, Rodriguez-Zas SL, Drackley JK, Lewin HA: **Nutrition-induced ketosis alters metabolic and signaling gene networks in liver of periparturient dairy cows**. *Physiol Genomics* 2007, **32**: 105-116

Luftman K, Hasan N, Day P, Hardee D, Hu C: **Silencing of VAMP3 inhibits cell migration and integrin-mediated adhesion**. *Biochem Biophys Res Commun* 2009, **380**: 65-70

Ma S, Huang J: **Regularized gene selection in cancer microarray meta-analysis**. *BMC Bioinformatics* 2009, **10**: 1-12

Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: a gene-centered information at NCBI**. *Nucl Acids Res* 2005, **33**: D54-D58

Mallya M, Campbell RD, Aguado B: **Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands**. *Protein Sci* 2006, **15**: 2244-2256

Marjani SL, Le Bourhis D, Vignon X, Heyman Y, Everts RE, Rodriguez-Zas SL, Lewin HA, Renard JP, Yang X, Tian XC: **Embryonic gene expression profiling using microarray analysis**. *Reprod Fertil Dev* 2009, **21**: 22-30

Melegos DN, Yu H, Diamandis EP: **Prostaglandin D2 sythase: a component of human amniotic fluid and its association with fetal abnormalities**. *Clin Chem* 1996, **42**: 1042-1050

Menard S: *Applied logistic regression analysis*. SAGE University Papers Series 07-106. SAGE, Thousand Oaks, California, 2001

Monticone M, Tonachini L, Tavella S, Degan P, Biticchi R, Palombi F, Puglisi R, Boitani C, Cancedda R, Castagnola P: **Impaired expression of genes coding for reactive oxygen species scavenging enzymes in testes of Mtfr1/Chppr-deficient mice**. *Reprod* 2007, **134**: 483-492

Morikawa Y, Cserjesi P: **Cardiac neural crest expression of Hand2 regulates outflow and second heart field development**. *Circulation Res* 2008, **103**: 1422-1429

Mourot M, Dufort I, Gravel C, Algriany O, Dieleman S, Sirard MA: **The influence of follicle size, fsh-enriched maturation medium, and early cleavage on bovine oocyte maternal mRNA levels**. *Mol Reprod Dev* 2006, **73**: 1367-1379

National Honey Board. Accessed: August 2008. [http://www.honey.com/]

National Human Genome Research Institute, National Institute of Health (NHGRI, NIH): **The human genome project completion: frequently asked questions**. 2008. Accessed: August 2008. [http://www.genome.gov/11006943]

Nelson CM, Ihle KE, Fondrk MK, Page RE, Amdam GV: **The gene vitellogenin has multiple coordinating effects on social organization**. *PLoS Biol* 2007, **5**: 673-677

Nguyen DV, Rocke DM: **Tumor classification by partial least squares using microarray gene expression data**. *Bioinformatics* 2002, **18**: 39-50

O'Donnell RK, Kupferman M, Wei SJ, Singhal S, Weber R, O'Malley, Jr. B, Cheng Y: **Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity**. *Oncogene* 2005, **24**: 1244-1251

O'Hara BF, Watson FL, Srere K, Kumar H, Wiler SW, Welch SK, Bitting L, Heller HC, Kilduff TS: **Gene expression in the brain across the hibernation cycle**. *J Neurosci* 1999, **19**: 3781-3790

Ott RL, Longnecker M: *An Introduction to Statistical Methods and Data Analysis*. Fifth Edition. Volume 49. USA: Thomson Learning Inc., 2001

Palmieri C, Loi P, Ptak G, Salda LD: **Review paper: a review of the pathology of abnormal placentae of somatic cell nuclear transfer clone pregnancies in cattle, sheep, and mice**. *Vet Pathol* 2008, **45**: 865-880

Pancaldi V, Schubert F, Bahler J: **Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast**. *Mol BioSyst* 2010, **6**: 543-552

Pankiw T, Page RE: **The effect of genotype, age, sex and caste on response thresholds to sucrose and foraging behavior of honeybees (*Apis mellifera L.*)**. *J Comp Physiol* 1999, **185**: 207-213

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A: **ArrayExpress – a public database of microarray experiments and gene expression profiles**. *Nucl Acids Res* 2007, **35**: D747-D750

Partial Least Squares (PLS). *Statsoft Electronic Textbook*. StatSoft, Inc. 2008, Accessed: August 2008. [http://www.statsoft.com/textbook/stpls.html]

Pearson K: **Report on certain enteric fever inoculation statistics**. *BMJ* 1904, **3**: 1243-1246

Pfister-Genskow M, Myers C, Childs LA, Lacson JC, Patterson T, Betthauser JM, Goueleke PJ, Koppang RW, Lange G, Fisher P, Watt SR, Forsberg EJ, Zheng Y, Leno GH, Schultz RM, Liu B, Chetia C, Yang X, Hoeschele I, Eilertsen KJ: **Identification of differentially expressed genes in individual bovine preimplantation embryos produced by nuclear transfer: improper reprogramming of genes required for development**. *Biol Reprod* 2005, **72**: 546-555

Pontius JU, Wagner L, Schuler GD: **UniGene: a unified view of the transcriptome**. In: *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information, 2003

Popovici V, Goldstein DR, Antonov J, Jaggi R, Delorenzi M, Wirapati P: **Selecting control genes for RT-QPCR using public microarray data**. *BMC Bioinformatics* 2009, **10**: 42-51

Press SJ, Wilson S: **Choosing between logistic regression and discriminant analysis**. *J Amer Stat Assoc* 1978, **73**: 699-705

Pyne S, Futcher B, Skiena S: **Meta-analysis based on control of false discovery rate: combining yeast ChIP-chip datasets**. *Bioinformatics* 2006, **22**: 2516-2522

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures**. *PNAS* 2001, **98:** 15149-15154

Raffiudin R, Crozier RH: **Phylogenetic analysis of honey bee behavioral evolution.** *Mol Phylogenet Evol* 2007, **43:** 543-552

Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer**. *Cancer Res* 2002, **62:** 4427–4433

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression**. *PNAS* 2004, **101**: 9309-9314

Robinson GE: **Genomics and integrative analyses of division of labor in honeybee colonies.** *Am Nat* 2002, **160:** S160-S172

Robinson GE, Grozinger CM, Whitfield CW: **Sociogenomics: Social life in molecular terms**. *Nat Rev Genet* 2005, **6**: 257-270

Rodrigues V, Cheah PY, Ray K, Chia W: **Malvolio, the *Drosophila* homolog of mouse NRAMP-1 (Bcg), is expressed in macrophages and in the nervous system and is required for normal taste behavior**. *Embo J* 1995, **14**: 3007-3020

Rodriguez-Zas SL: **Chapter 7: Analysis of Differentially Expressed Genes**. ANSC 545 Statistical Genomics [class notes] 2005a

Rodriguez-Zas SL: **Chapter 10: Experimental Designs**. ANSC 545 Statistical Genomics [class notes] 2005b

Rodriguez-Zas SL, Southey BR, Whitfield CW, Robinson GE: **Semiparametric approach to characterize unique gene expression trajectories across time.** *BMC Genomics* 2006, **7:** 233-246

Rodriguez-Zas SL, Ko Y, Adams HA, Southey BR: **Advancing the understanding of the embryo transcriptome co-regulation using meta-, functional, and gene network analysis tools.** *Reprod* 2008, **135:** 213–224

Romualdi C, De Pittà C, Tombolan L, Bortoluzzi S, Sartori F, Rosolen A, Lanfranchi G: **Defining the gene expression signature of rhabdomyosarcoma by meta-analysis.** *BMC Genomics* 2006, **7:** 287-302

Roth JS, Buccino G, Klein NW: **Inhibition of growth of chick embryo by inhibition of deoxycytidylate deaminase**. *Science* 1963, **13**: 1473-1474

SAS Institute Inc., **SAS 9.1.3 Help and Documentation**, Cary, NC: SAS Institute Inc., 2000-2004

Schallenberger E, Schams D, Meyer HHD: **Sequences of pituitary, ovarian and uterine hormone secretion during the first 5 weeks of pregnancy in dairy cattle**. *J Reprod Fertil* 1989, **37(Suppl)**: 277-286

Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**: 467-470

Schmid CH, Stark PC, Berlin JA, Landais P, Lau J: **Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors.** *Clin Epidemiol* 2004, **57:** 683-697

Schneider J, Ruschhaupt M, Buness A, Asslaber M, Regitnig P, Zatloukal K, Schippinger W, Ploner F, Poustka A, Sultmann H: **Identification and meta-analysis of a small gene expression signature for the diagnosis of estrogen receptor status in invasive ductal breast cancer.** *Int J Cancer* 2006, **119:** 2974-2979

Schweifer N, Valk PJ, Delwel R, Cox R, Francis F, Meier-Ewert S, Lehrach H, Barlow DP: **Characterization of the C3 YAC contig from proximal mouse chromosome 17 and analysis of allelic expression of genes flanking the imprinted Igf2r gene**. *Genomics* 1997, **43**: 285-297

Seals RC, Lemaster JW, Hopkins FM, Schrick FN: **Effects of elevated concentrations of prostaglandin F2 alpha on pregnancy rates in progestogen supplemented cattle**. *Prostaglandins Other Lipid Mediat* 1998, **56**: 377-389

Sen Sarma M, Whitfield CW, Robinson GE: **Species differences in brain gene expression profiles associated with adult behavioral differences in honey bees.** *BMC Genomics* 2007*,* **8**: 202-215

Sen Sarma M, Rodriguez-Zas SL, Hong F, Zhong S, Robinson GE: **Transcriptomic profiling of central nervous system regions in three species of honey bee during dance communication behavior**. *PLoS One* 2009, **4**: e6408-e6416

Shanks RD, Popp RG, McCoy GC, Nelson DR, Robinson JL: **Identification of the homozygous recessive genotype for the deficiency of uridine monophosphate synthase in 35-day bovine embryos**. *J Reprod Fertil* 1992, **94**: 5-10

Shanks RD, Robinson JL: **Embryonic mortality attributed to inherited deficiency of uridine monophosphate synthase**. *J Dairy Sci* 1989, **72**: 3035-3039

Shrier I, Platt RW, Steele RJ: **Mega-trials vs. meta-analysis: Precision vs. heterogeneity?** *Contemp Clin Trials* 2007, **28**: 324-328

Sinclair KD, Broadbent PJ, Dolman DF: **In vitro produced embryos as a means of achieving pregnancy and improving productivity in beef cows**. *Anim Sci* 1995, **60**: 55-64

Smith DD, Saetrom P, Snove O, Lundberg C, Rivas GE, Glackin C, Larson GP: **Meta-analysis of breast cancer microarray studies in conjunction with conserved *cis*-elements suggest patterns for coordinate regulation**. *BMC Bioinformatics* 2008, **9**: 63-77

Smith SL, Everts RE, Tian XC, Du F, Sung L-Y, Rodriguez-Zas SL, Jeong B-S, Renard J-P, Lewin HA, Yang X: **Global gene expression profiles reveal significant nuclear reprogramming by the blastocyst stage after cloning**. *PNAS* 2005, **102**: 17582-17587

Smith S, Le Bourhis D, Vignon X, Heyman Y, Lewin H, Tian XC, Renard J-P, Yang X: **Global gene expression profiling of single, cloned bovine embryos with different developmental competencies: the good, the bad, and the ugly**. *SSR, Society of Reproduction* 2007, Abstract

Smith SL, Everts RE, Sung L-Y, Du F, Page RL, Henderson B, Rodriguez-Zas SL, Nedambale TL, Renard J-P, Lewin HA, Yang X, Tian XC: **Gene expression profiling of single bovine embryos uncovers significant effects of in vitro maturation, fertilization and culture**. *Mol Reprod Dev* 2009, **76**: 38-47

Somers J, Smith C, Donnison M, Wells DN, Henderson H, McLeay L, Pfeffer PL: **Gene expression profiling of individual bovine nuclear transfer blastocysts**. *Reprod* 2006, **131**: 1073-1084

Southwick EE, Moritz RFA: **Social-control of air ventilation in colonies of honeybees, *Apis mellifera***. *J Insect Physiol* 1987, **33**: 623-626

Srivastava D, Thomas T, Lin Q, Kirby ML, Brown D, Olson EN: **Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND**. *Nat Genet* 1997, **16**: 154-160

Stekel D: **Microarray Bioinformatics**. USA: Cambridge University Press, 2003

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS* 2005, **102**: 15545-15550

Sun H, Fang H, Chen T, Perkins R, Tong W: **GOFFA: Gene Ontology For Functional Analysis - A FDA Gene Ontology tool for analysis of genomic and proteomic data**. *BMC Bioinformatics* 2006, **7:** S23-S33

Tan CS, Ploner A, Quandt A, Lehtio J, Pawitan Y: **Finding regions of significance in SELDI measurements for identifying protein biomarkers**. *Bioinformatics* 2006, **22**: 1515-1523

Tian XC, Kubota C, Enright B, Yang X: **Cloning animals by somatic cell nuclear transfer – biological factors**. *Reprod Biol Endocrinol* 2003, **1**: 98-104

Tweedie S, Ashurner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H, and The FlyBase Consortium: **FlyBase: enhancing Drosophila Gene Ontology annotations**. *Nucl Acids Res* 2009, **37**: D555-D559; doi:10.1093/nar/gkn788

Van Soom A, de Kruif A: **Bovine embryonic development after in vivo and in vitro fertilization**. *Reprod Dom Anim* 1998, **33**: 261-265

Verhaagh S, Schweifer N, Barlow DP, Zwart R: **Cloning of the mouse and human solute carrier 22a3 (Slc22a3/SLC22A3) identifies a conserved cluster of three organic cation transporters on mouse chromosome 17 and human 6q26-q27**. *Genomics* 1999, **55**: 209-218

Wald A: **Tests of statistical hypotheses concerning several parameters when the number of observations is large**. *Amer Math Soc* 1943, **54**: 426-482

Weiner KXB, Weiner RS, Maley F, Maley GF: **Primary structure of human deoxycytidylate deaminase and overexpression of its function protein in Escherichia coli**. *J Biol Chem* 1993, **17**: 12983-12989

Westermann B: **Merging mitochondria matters: Cellular role and molecular machinery of mitochondrial fusion**. *EMBO Rep* 2002, **3**: 527-531

Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE: **Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee.** *Genome Res* 2002, **12:** 555-566

Whitfield CW, Cziko AM, Robinson GE: **Gene expression profiles in the brain predict behavior in individual honey bees.** *Science* 2003, **302:** 296-299

Whitfield CW, Ben-Shahar Y, Brillet C, Leoncini I, Crauser D, LeConte Y, Rodriguez-Zas SL, Robinson GE: **Genomic dissection of behavioral maturation in the honey bee.** *PNAS* 2006, **103:** 16068-16075

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models.** *J Comput Biol* 2001, **8:** 625-637

Wren JD: **A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide**. *Bioinformatics* 2009, **25**: 1694-1701

Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM: **A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma**. *PNAS* 2003, **100**: 9991-9996

Wu CH, McLarty JW: *Neural Networks and Genome Informatics*. New York, NY: Elsevier Science, Inc., 2000

Yanagisawa H, Hammer RE, Richardson JA, Emoto N, Williams SC, Takeda S, Clouthier DE, Yanagisawa M: **Disruption of ECE-1 and ECE-2 reveals a role for endothelin-converting enzyme-2 in murine cardiac development**. *J Clin Invest* 2000, **105**: 1373-1382

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucl Acids Res* 2002, **30**: e15-e24

Yang X, Sun X: **Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers.** *BMC Bioinformatics* 2007, **8:** 118-134

Yang ZZ, Tschopp O, Hemmings-Mieszczak M, Feng J, Brodbeck D, Perentes E, Hemmings BA: **Protein Kinase B alpha/ Akt1 regulates placental development and fetal growth**. *J Bio Chem* 2003, **278**: 32124-32131

Yoshikane N, Nakamura N, Ueda R, Ueno N, Yamanaka S, Nakamura M: **Drosophila *NAT1*, a homolog of the vertebrate translational regulator *NAT1/DAP5/p97*, is required for embryonic germband extension and metamorphosis**. *Dev Growth Differ* 2007, **49**: 623-634

Zhang Z, Fenstermacher D: **An Introduction to MAMA (Meta-Analysis of MicroArray data) System.** *Conf Proc IEEE Eng Med Biol Soc* 2005, **7:** 7730-7733

Zhu J, Hastie T: **Classification of gene microarrays by penalized logistic regression**. *Biostatistics* 2004, **5**: 427-443

Tables and Figures

Figures

Figure 2.1
Venn diagram of the number of differentially expressed transcripts detected by at least two individual-study analyses (Ind), study-level (Study), and sample-level (Sample) meta-analyses.

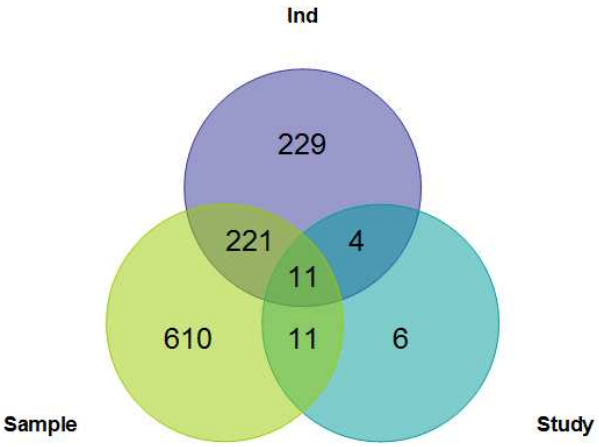
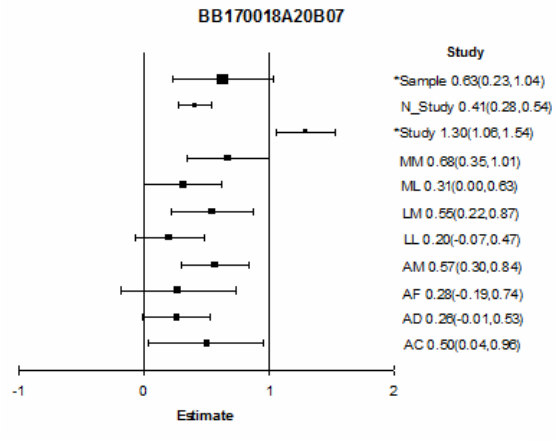


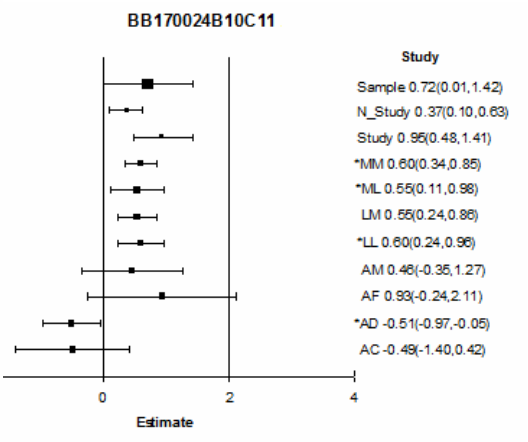
Figure 2.2

Funnel plots of differential expression estimates +/- 1 standard error (S.E.) for *Apis mellifera* transcripts BB170018A20B07 (2A), BB170024B10C11 (2B), BB170024B20H01 (2C), and BB170004B20H08 (2D), by individual-study, study-level, and sample-level meta-analyses.

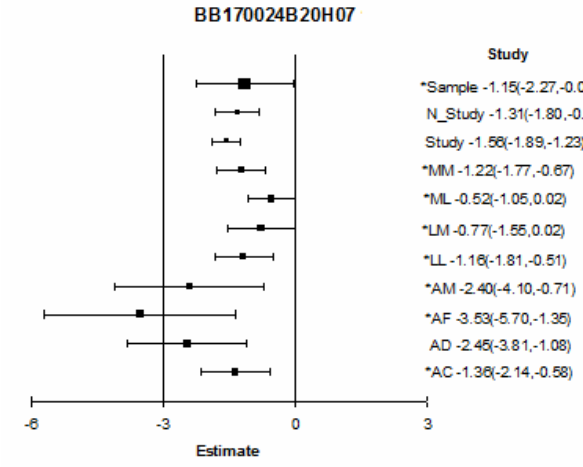
2A



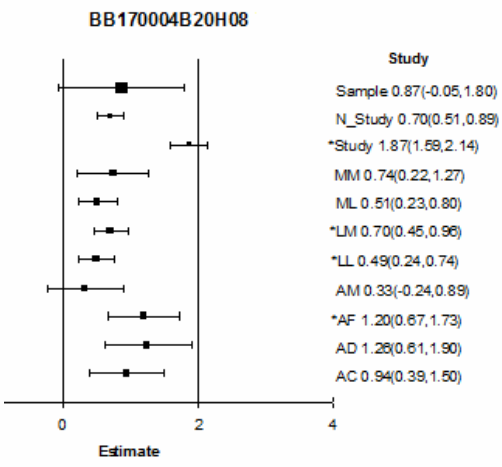
2B



2C



2D



Footnote: Estimates +/- 1 S.E. for each analysis are represented by a square and a horizontal line, respectively. Study denotes study-level meta-analysis of standardized estimates, N_Study denotes study-level meta-analysis of non-standardized estimates, Sample denotes sample-level meta-analysis. AC: *Apis cerana* bees raised on *Apis cerana* colonies; AD: *Apis dorsata* bees raised on *Apis dorsata* colonies; AF: *Apis florea* bees raised on *Apis florea* colonies; AM: *Apis mellifera* bees raised on an *Apis* mellifera colony, LL: *Apis mellifera ligustica* bees raised on an *Apis mellifera ligustica* colony; LM: *Apis mellifera ligustica* bees raised on an *Apis mellifera mellifera colony*; ML: *Apis mellifera mellifera* bees raised on an *Apis mellifera ligustica* colony; MM: *Apis mellifera mellifera* bees raised on an *Apis mellifera mellifera* colony. The size of the square denoting the estimate corresponds to the number of observations in the study (AC, AD, AF n = 24; AM n = 22; LL, LM, ML, MM n = 12; study-level (Study) meta-analysis n = 8; sample-level (Sample) meta-analysis n = 142). Analyses detecting significant (P-value $<1x10^{-3}$) differential expression between forager and one-day-old honey bees are denoted by an asterisk.

Figure 3.1
Principal component analysis of the top 25 genes by sample. Red circles denote one-day-old samples for each study (e.g. O_AC are one-day-olds from study AC), and black circles denote forager samples for each study (e.g. F_AC are foragers from study AC). The solid line shows separation between forager and one-day-old honey bees. The dashed line shows separation between Group 1 studies (AC, AD, AF, AM) and Group 2 studies (LL, LM, ML, MM).

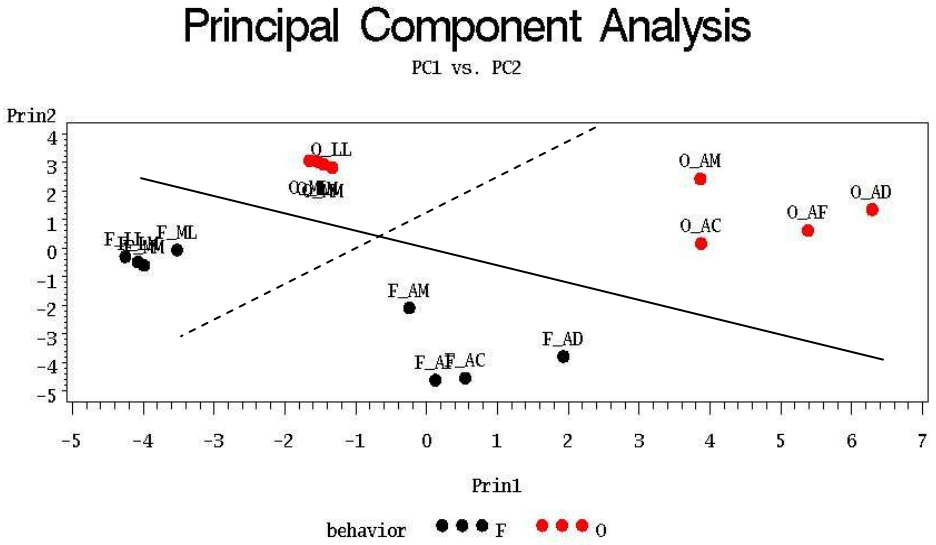


Figure 4.1
Venn diagram of the number of differentially expressed genes detected by one or more individual-study analyses (Ind), the study-level (Study), and the sample-level (Sample) meta-analyses.

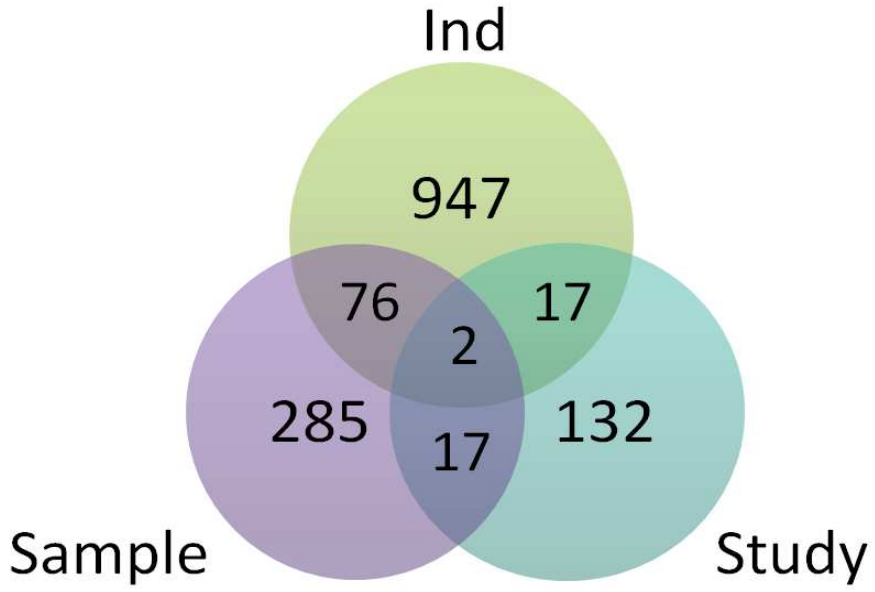


Figure 4.2

Funnel plots of differential expression estimates (square) and 95% confidence interval limits (horizontal line) for genes deoxycytidylate deaminase (*DCTD*), mitochondrial fission regulator 1 (*MTFR1*), vesicle-associated membrane protein 3 (*VAMP3*), and structural maintenance of chromosomes 4-like 1 (*SMC4L1*), by individual-study, study-level (Study), and sample-level (Sample) meta-analyses. Estimates STU denotes study-level meta-analysis of standardized estimates, and SAM denotes sample-level meta-analysis. d25X denotes day 25 extra-embryonic samples (Everts *et al.* 2007a), d25E denotes day 25 embryonic samples (Everts *et al.* 2007b), d280X denotes day 280 or term extra-embryonic sample. (Everts *et al.* 2008), and d7E denotes day 7 embryo samples (Marjani *et al.* 2009; Smith *et al.* 2007). The size of the square denoting the estimate corresponds to the number of observations in the study (d25X n=12; d25E n=11; d280X n=29; d7E n=29; study-level (Study) meta-analysis n = 4; sample-level (Sample) meta-analysis n = 89). Analyses detecting significant (P-value $<1x10^{-3}$, fold change > 1.23) differential expression between AI and NT samples are denoted by an asterisk.

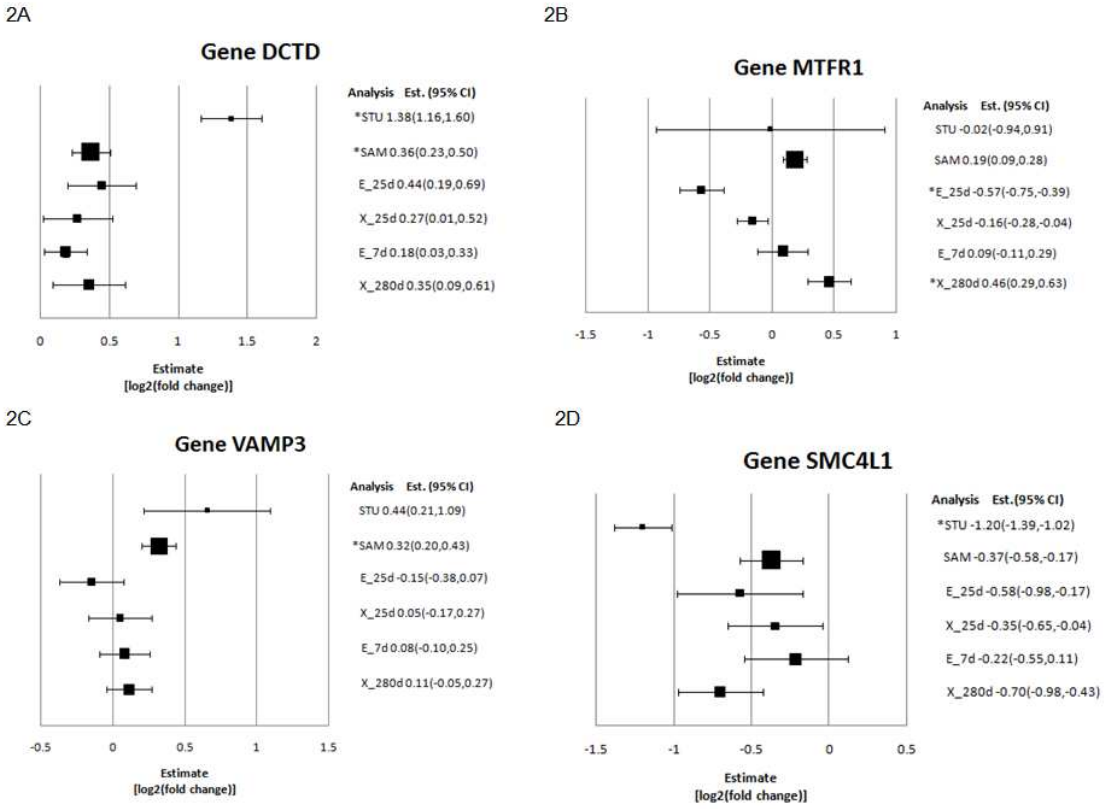


Figure 4.3
Venn diagrams of the number of differentially expressed genes detected **(a)** within embryonic tissue and reproductive technology and between developmental stage, and within embryonic tissue and between reproductive technology and developmental stage; **(b)** within extra-embryonic tissue and reproductive technology and between developmental stage, and within extra-embryonic tissue and between reproductive technology and developmental stage; and **(c)** within reproductive technique and developmental stage and between tissue source, and within developmental stage and between reproductive technique and tissue source.

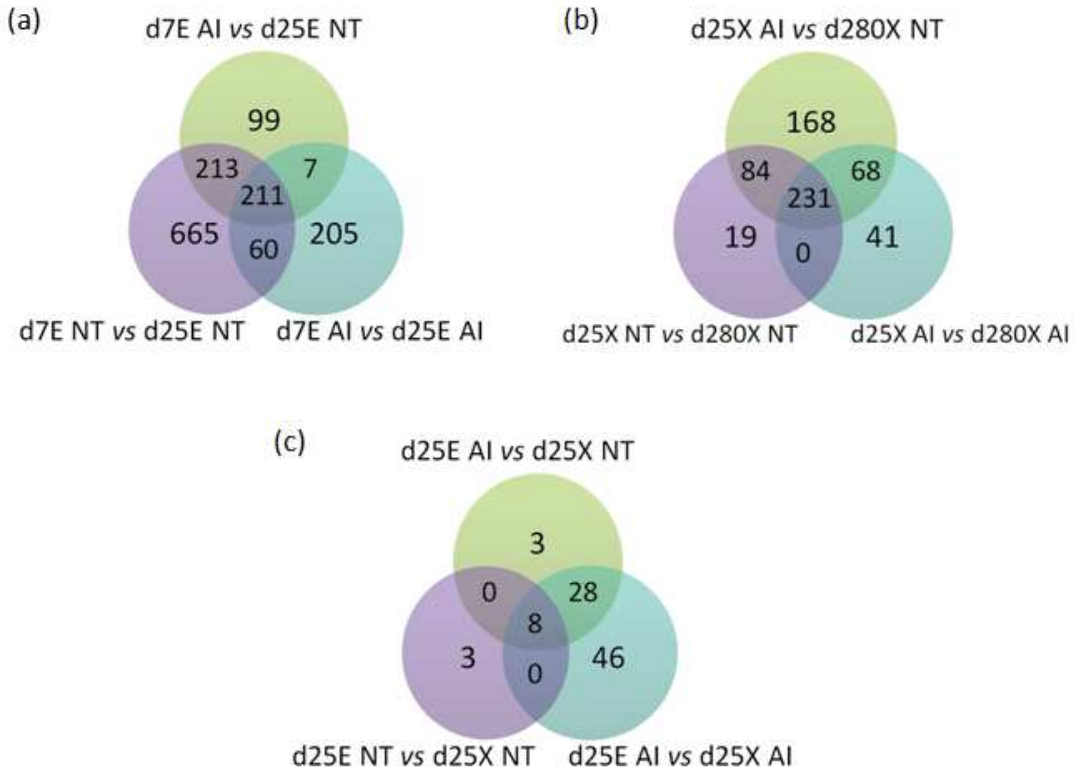
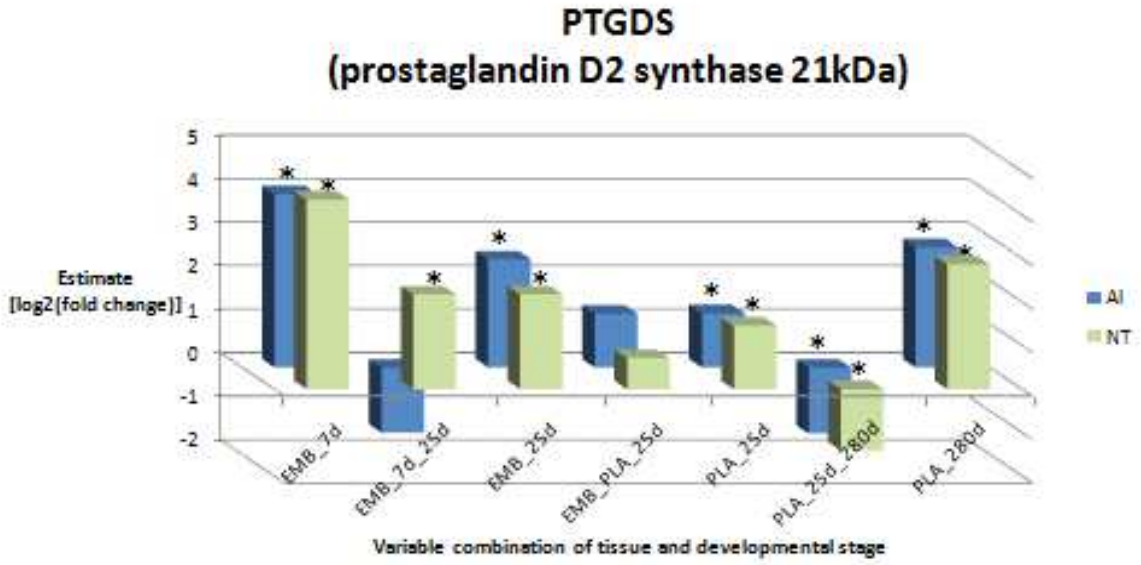


Figure 4.4
Expression of prostaglandin D2 synthase gene within reproductive technology [AI, NT], for different combinations of tissue source [embryonic, extra-embryonic] and developmental stage [7 d, 25 d, 280 d]. The expression estimate is measured as the log2 transformation of the fold change. Columns denoted with asterisks outline significant terms (P-value <.001).



Tables

Table 2.1. Number of transcripts with differential expression (P-value < $1x10^{-3}$) within individual analyses, study-level standardized (Study), and sample-level (Sample) meta-analysis (diagonals), number (upper off-diagonals) and percentage (lower off-diagonals) of transcripts identified differentially expressed in all pairs of analyses relative to the maximum number of significant transcripts that can overlap in both analyses.

	Individual Analyses[1]								**Meta-Analyses**	
	AC	**AD**	**AF**	**AM**	**LL**	**LM**	**ML**	**MM**	**Study**	**Sample**
AC	152	9	55	60	15	21	21	26	6	123
AD	4.3%	65	9	12	5	5	8	6	3	42
AF	38.5%	13.8%	143	53	19	24	12	18	9	99
AM	39.5%	18.5%	37.1%	310	20	35	37	38	6	168
LL	9.9%	7.7%	13.3%	6.5%	422	86	76	78	7	111
LM	13.8%	7.7%	16.8%	11.3%	22.8%	377	73	80	7	114
ML	13.8%	12.3%	8.4%	11.9%	18.0%	19.4%	540	57	6	126
MM	17.1%	9.2%	12.6%	12.3%	18.9%	21.2%	13.8%	413	8	117
Study	18.8%	9.4%	28.1%	18.8%	21.9%	21.9%	18.8%	25.0%	32	22
Sample	80.9%	64.6%	69.2%	54.2%	26.3%	30.2%	23.3%	28.1%	68.8%	853

[1]AC: *Apis cerana* bees raised on *Apis cerana* colonies; AD: *Apis dorsata* bees raised on *Apis dorsata* colonies; AF: *Apis florea* bees raised on *Apis florea* colonies; AM: *Apis mellifera* bees raised on an *Apis* mellifera colony, LL: *Apis mellifera ligustica* bees raised on an *Apis mellifera ligustica* colony; LM: *Apis mellifera ligustica* bees raised on an *Apis mellifera mellifera colony*; ML: *Apis mellifera mellifera* bees raised on an *Apis mellifera ligustica* colony; MM: *Apis mellifera mellifera* bees raised on an *Apis mellifera mellifera* colony.

Table 2.2. Number of transcripts with positive (over expression in forager compared to one-day-old honey bees) and negative (under-expression in forager compared to one-day-old honey bees) differential expression (P-value < $1x10^{-3}$) within individual analysis, study-level standardized (Study) and sample-level (Sample) meta-analysis (diagonals); number (upper off-diagonals), and percentage (lower off-diagonals) of transcripts identified differentially expressed in all pairs of analyses relative to the maximum number of significant transcripts that can overlap in both analyses.

		Individual Analyses[1]																Meta-Analyses			
		AC		AD		AF		AM		LL		LM		ML		MM		Study		Sample	
		+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
AC	+	59	-	6	0	22	0	20	0	3	0	7	0	7	0	7	1	2	0	39	10
	-	-	93	0	3	0	33	0	40	0	12	0	14	1	13	0	18	0	4	7	67
AD	+	13.6%	0	44	-	6	0	7	0	2	0	3	0	2	2	5	0	2	0	22	4
	-	0.0%	14.3%	-	21	0	3	0	5	1	2	0	2	1	3	1	0	0	1	4	12
AF	+	40.0%	0.0%	13.6%	0.0%	55	-	13	0	6	0	8	0	3	0	1	1	3	0	30	7
	-	0.0%	37.5%	0.0%	14.3%	-	88	2	38	0	13	1	15	0	9	2	14	0	6	7	55
AM	+	33.9%	0.0%	15.9%	0.0%	23.6%	2.3%	97	-	4	0	9	1	11	3	6	2	4	0	37	11
	-	0.0%	43.0%	0.0%	23.8%	0.0%	43.2%	-	213	0	16	1	24	3	20	4	26	0	2	26	94
LL	+	5.1%	0.0%	4.5%	4.8%	10.9%	0.0%	4.1%	0.0%	238	-	46	0	48	0	45	0	3	0	57	1
	-	0.0%	12.9%	0.0%	9.5%	0.0%	14.8%	0.0%	8.7%	-	184	0	40	0	28	0	33	0	4	4	49
LM	+	11.9%	0.0%	6.8%	0.0%	14.5%	1.1%	9.3%	0.5%	23.0%	0.0%	200	-	41	0	32	0	3	0	53	1
	-	0.0%	15.1%	0.0%	9.5%	0.0%	17.0%	1.0%	13.6%	0.0%	22.6%	-	177	0	32	0	48	0	4	0	60
ML	+	11.9%	1.1%	4.5%	4.8%	5.5%	0.0%	11.3%	1.4%	20.2%	0.0%	20.5%	0.0%	263	-	34	0	3	0	67	4
	-	0.0%	14.0%	4.5%	14.3%	0.0%	10.2%	3.1%	9.4%	0.0%	15.2%	0.0%	18.1%	-	277	0	23	0	3	6	49
MM	+	11.9%	0.0%	11.4%	4.8%	1.8%	2.3%	6.2%	1.9%	21.5%	0.0%	16.0%	0.0%	16.3%	0.0%	209	-	4	0	49	0
	-	1.7%	19.4%	0.0%	0.0%	1.8%	15.9%	2.1%	12.7%	0.0%	17.9%	0.0%	27.1%	0.0%	11.3%	-	204	0	4	0	68
Study	+	12.5%	0.0%	12.5%	0.0%	18.8%	0.0%	25.0%	0.0%	18.8%	0.0%	18.8%	0.0%	18.8%	0.0%	25.0%	0.0%	16	-	11	0
	-	0.0%	25.0%	0.0%	6.3%	0.0%	37.5%	0.0%	12.5%	0.0%	25.0%	0.0%	25.0%	0.0%	18.8%	0.0%	25.0%	-	16	0	11
Sample	+	66.1%	7.5%	50.0%	19.0%	54.5%	8.0%	38.1%	12.2%	23.9%	2.2%	26.5%	0.0%	25.5%	2.2%	23.0%	0.0%	68.8%	0.0%	437	-
	-	16.9%	72.0%	9.1%	57.1%	12.7%	62.5%	11.3%	44.1%	0.4%	26.6%	0.5%	33.9%	1.5%	17.7%	0.0%	33.3%	0.0%	68.8%	-	416

[1]AC: *Apis cerana* bees raised on *Apis cerana* colonies; AD: *Apis dorsata* bees raised on *Apis dorsata* colonies; AF: *Apis florea* bees raised on *Apis florea* colonies; AM: *Apis mellifera* bees raised on an *Apis* mellifera colony, LL: *Apis mellifera ligustica* bees raised on an *Apis mellifera ligustica* colony; LM: *Apis mellifera ligustica* bees raised on an *Apis mellifera mellifera colony*; ML: *Apis mellifera mellifera* bees raised on a *Apis mellifera ligustica* colony; MM: *Apis mellifera mellifera* bees raised on an *Apis mellifera mellifera* colony.

Table 2.3. Gene Ontology (molecular function or mol. function, biological process or bio. process, and cellular component or cell. component) information for 12 transcripts with significant differential expression in four or more studies.

Transcript	*Computed Gene (CG) Symbol*	*FlyBase ID*	*Gene Name*	*GO*	*GO Category*	*GO Description*
BB170024B20H07	CG6186	FBgn0022355	*Tsf1 Transferrin 1*	GO:0006879	Bio. Process	cellular iron ion homeostasis
				GO:0006952	Bio. Process	defense response
				GO:0006826	Bio. Process	iron ion transport
				GO:0008199	Mol. Function	ferric iron binding
				GO:0005381	Mol. Function	iron ion transmembrane transporter activity
BB160004B20A10	CG17737	FBgn0035423		GO:0006412	Bio. Process	translation
BB170020B10D07	CG9307	FBgn0038180	*Cht5*	GO:0004568	Mol. Function	chitinase activity
BB170013A20H11	CG31997	FBgn0051997		GO:0008150	Bio. Process	unknown
				GO:0005575	Cell. Component	unknown
				GO:0003674	Mol. Function	unknown
BB170012B20G09	CG7470	FBgn0037146		GO:0006561	Bio. Process	proline biosynthetic process
				GO:0005811	Cell. Component	lipid particle
				GO:0005743	Cell. Component	mitochondrial inner membrane
				GO:0005739	Cell. Component	mitochondrion
				GO:0017084	Mol. Function	delta 1-pyrroline-5-carboxylate synthetase activity

Table 2.3 (cont.)						
				GO:0004349	Mol. Function	glutamate 5-kinase activity
				GO:0004350	Mol. Function	glutamate-5-semialdehyde dehydrogenase activity
BB160024A10G05	CG5220	FBgn0038471		GO:0008650	Mol. Function	rRNA (uridine-2'-O-)-methyltransferase activity
BB170003A10E02	CG5840	FBgn0038516		GO:0004735	Mol. Function	pyrroline-5-carboxylate reductase activity
BB160009A20G09	CG30035	FBgn0050035		GO:0008643	Bio. Process	carbohydrate transport
				GO:0016021	Cell. Component	integral to membrane
				GO:0005355	Mol. Function	glucose transmembrane transporter activity
BB170024B10C11	CG10662	FBgn0032832	*sick sickie*	GO:0050829	Bio. Process	defense response to Gram-negative bacterium
				GO:0005575	Cell. Component	ND
BB160007A10B09	CG17762	FBgn0030412	*tomosyn*	GO:0007163	Bio. Process	establishment and/or maintenance of cell polarity
				GO:0007269	Bio. Process	neurotransmitter secretion
				GO:0016079	Bio. Process	synaptic vesicle exocytosis

Table 2.3 (cont.)						
				GO:0016082	Bio. Process	synaptic vesicle priming
				GO:0008021	Cell. Component	synaptic vesicle
				GO:0017075	Mol. Function	syntaxin-1 binding
BB160024B10F06	CG9907	FBgn0003036	*para* paralytic	GO:0045433	Bio. Process	male courtship behavior, veined wing generated song production
				GO:0046680	Bio. Process	response to DDT
				GO:0046684	Bio. Process	response to pyrethroid
				GO:0006814	Bio. Process	sodium ion transport
				GO:0005887	Cell. Component	integral to plasma membrane
				GO:0005886	Cell. Component	plasma membrane
				GO:0001518	Cell. Component	voltage-gated sodium channel complex
				GO:0005248	Mol. Function	voltage-gated sodium channel activity
BB170010A10A08	CG31634	FBgn0051634	*Oatp26F*	GO:0015711	Bio. Process	organic anion transport
				GO:0008514	Mol. Function	organic anion transmembrane transporter activity

Table 2.4. Number of transcripts with non-significant and significant (Sig., P-value < $1x10^{-3}$) differential expression between forager and one-day-old honey bees within studies in Group 1 (AC[1], AD, AF, AM), Group 2 (LL, LM, ML, MM), and both groups combined.

	Group 1	**Group 2**	**Both Groups**
Non-significant	7213	3652	5978
Sig. in 1 study	408	1092	1294
Sig. in 2 studies	69	228	320
Sig. in 3 studies	39	56	100
Sig. in 4 studies	5	9	29
Sig. in 5 studies	N/A	N/A	12
Sig. in 6 studies	N/A	N/A	3
Sig. in 7 studies	N/A	N/A	1
Total	7734	7737	7737

[1]AC: *Apis cerana* bees raised on an *Apis cerana* colony; AD: *Apis dorsata* bees raised on an *Apis dorsata* colony; AF: *Apis florea* bees raised on an *Apis florea* colony; AM: *Apis mellifera* bees raised on an *Apis mellifera* colony, LL: *Apis mellifera ligustica* bees raised on an *Apis mellifera ligustica* colony; LM: *Apis mellifera ligustica* bees raised on an *Apis mellifera mellifera colony*; ML: *Apis mellifera mellifera* bees raised on an *Apis mellifera ligustica* colony; MM: *Apis mellifera mellifera* bees raised on an *Apis mellifera mellifera* colony.

Table 3.1. Numbers of genes that have significant (false discovery rate P-value < 0.1) differential odds of behavior from the study-level (study) and sample-level (sample) meta-analyses of the eight studies.

		Study		**Sample**	
		+	-	+	-
Study	+	22	-	14	0
	-	-	41	0	29
Sample	+			439	-
	-			-	431

The numbers of genes for each meta-analysis approach are in the diagonals, and the numbers of genes that overlap between meta-analyses approaches are in the off-diagonals.
+ = Positive transcripts: increase in gene expression is associated with an increase in the odds that the observation pertains to a forager honey bee.
- = Negative transcripts: increase in gene expression is associated with an increase in the odds that the observation pertains to a one-day-old honey bee.
No gene that has significant association with the odds of behavior was identified in any individual-study analysis.

Table 3.2. The top (most significant) twenty-five genes from the study-level and sample-level meta-analyses.

Honey bee gene identifier	*EST*	*GeneBank Number*	*FlyBase ID*	*Fly Ortholog*	*GO Number*	*GO Description*
GB18624 *	BB170011B20G02	BI508929				
GB15988	BB170003B10A03	BI508870				
GB15106	BB170032B10G12	BI503571				
GB10622 *	BB170018B20G10	BI506143				
GB19745	BB160007B20C09	BI511993	FBgn0022355	CG6186	GO:0005381	iron ion transmembrane transporter activity[3]
			Transferrin 1			
					GO:0005386	N/A
			Tsf1			
					GO:0005576	extracellular region[2]
					GO:0006826	iron ion transport[1]
					GO:0006879	cellular iron ion homeostasis[1]
					GO:0006952	defense response[1]
					GO:0008199	ferric iron binding[3]
GB12396	BB170014B10F12	BI507504				
GB30320	BB170028A20D12	BI506421				
GB10622	BB170018B10A09	BI509691				
GB30213 *+	BB160009B20H09	BI512664				
GB17538	BB170005A20E08	BI508391				
GB14388 +	BB160024B20A02	BI517202				
GB19724 +	BB160012B20G04	BI513447				
GB30033 +	BB160022A10D12	BI516304				
GB19018 +	BB170016B20C12	BI503685				
GB11892 +	BB160009B20H07	BI512662				
GB30514 *+	BB160008B10A08	BI512213	FBgn0026428	CG6170	GO:0000074	N/A
			HDAC6		GO:0003682	chromatin binding[3]

Table 3.2 (cont.)

					GO:0004407	histone deacetylase activity[3]
					GO:0006333	chromatin assembly or disassembly[1]
					GO:0006357	regulation of transcription from RNA polymerase II promoter[1]
					GO:0008270	zinc ion binding[3]
					GO:0016575	histone deacetylation[1]
GB10744 +	BB160017A10D11	BI515017				
GB30434 +	BB160016A20H02	BI514790				
GB11814	BB170014A20G09	BI504893				
GB17831	BB170030B10G08	BI503528	FBgn0000473	CG9438	GO:0004497	monooxygenase activity[3]
			Cytochrome P450-6a2		GO:0005506	iron ion binding[3]
					GO:0005792	microsome[2]
			Cyp6a2		GO:0006118	electron transport[1]
					GO:0008202	steroid metabolic process[1]
					GO:0009055	electron carrier activity[3]
					GO:0016020	membrane[2]
					GO:0017085	response to insecticide[1]
					GO:0020037	heme binding[3]
GB17444	BB170007B20E02	BI504564				
GB12917	BB170024B20H09	BI507137	FBgn0053526	CG33526	GO:0003676	nucleic acid binding[3]

Table 3.2 (cont.)

			PNUTS		GO:0005634	nucleus[2]
					GO:0006605	protein targeting[1]
					GO:0008270	zinc ion binding[3]
					GO:0019888	protein phosphatase regulator activity[3]
					GO:0035304	regulation of protein amino acid dephosphorylation[1]
					GO:0050790	regulation of catalytic activity[1]
GB19938 [+]	BB160017B20H07	BI515276				
GB13652	BB170006A10A12	BI507356				
GB18753	BB170025A20H08	BI507142	FBgn0037262	CG1057	GO:0000119	mediator complex[2]
			Mediator complex subunit 31		GO:0003702	RNA polymerase II transcription factor activity[3]
			MED31		GO:0006367	transcription initiation from RNA polymerase II promoter[1]
					GO:0016455	RNA polymerase II transcription mediator activity[3]
					GO:0045449	regulation of transcription[1]

[1] Biological process; [2] Cellular component; [3] Molecular function; N/A Not Applicable; * Scores over 100; [+] Score1/Score2 from blast hit is < 2; GB numbers are based on Score1; Expressed Sequence Tags (EST) are experimental identifiers; GeneBank Number (Maglott *et al.*, 2005); FlyBase ID and Fly Ortholog based on *Drosophila melanogaster* gene annotations (Tweedie *et al.*, 2009); Gene Ontology (GO) Number and Description based on functional information for *Drosophila melanogaster* genes (Gene Ontology Consortium, 2000)

Table 3.3. Results from the logistic analysis of the complete data set using each of the top 25 genes, identified in the study- and sample-level meta-analyses from stage one, individually. Genes listed in order of most to least significant from stage one analysis.

ID	Est.	SE	P-value	R^2	Error (N of 158)	Class. Error rate	TP	TN	FP	FN	Sens.	Spec.
GB18624	0.45	0.09	$6.2x10^{-7}$	0.18	54	0.34	56	48	31	23	0.71	0.61
GB15988	-0.90	0.18	$4.2x10^{-7}$	0.26	47	0.30	65	46	33	14	0.82	0.58
GB15106	0.60	0.14	$1.3x10^{-5}$	0.14	51	0.32	60	47	32	19	0.76	0.59
GB10622	0.25	0.07	$3.4x10^{-4}$	0.02	57	0.36	47	54	25	32	0.59	0.68
GB19745	0.21	0.08	$7.5x10^{-3}$	0.09	68	0.43	53	37	42	26	0.67	0.47
GB12396	-0.25	0.08	$1.9x10^{-3}$	0.05	67	0.42	48	43	36	31	0.61	0.54
GB30320	0.18	0.05	$7.9x10^{-5}$	0.07	65	0.41	49	44	35	30	0.62	0.56
GB10622	0.37	0.10	$1.5x10^{-4}$	0.11	62	0.39	46	50	29	33	0.58	0.63
GB30213	0.08	0.06	$2.1x10^{-1}$	0.11	75	0.47	28	55	24	51	0.35	0.70
GB17538	0.51	0.12	$3.0x10^{-5}$	0.01	54	0.34	48	56	23	31	0.61	0.71
GB14388	-0.37	0.10	$4.2x10^{-4}$	0.14	67	0.42	35	56	23	44	0.44	0.71
GB19724	-0.46	0.12	$9.0x10^{-5}$	0.08	53	0.34	48	57	22	31	0.61	0.72
GB30033	-0.05	0.05	$3.5x10^{-1}$	0.11	88	0.56	35	35	44	44	0.44	0.44
GB19018	-2.04	0.38	$9.5x10^{-8}$	0.01	40	0.25	61	57	22	18	0.77	0.72
GB11892	-0.33	0.14	$1.\ 8x10^{-2}$	0.22	78	0.49	42	38	41	37	0.53	0.48
GB30514	0.42	0.16	$9.0x10^{-3}$	0.04	65	0.41	47	46	33	32	0.59	0.58
GB10744	-0.46	0.12	$2.1x10^{-4}$	0.05	62	0.39	53	43	36	26	0.67	0.54
GB30434	0.41	0.13	$1.8x10^{-3}$	0.09	64	0.41	51	43	36	28	0.65	0.54
GB11814	0.08	0.08	$3.0x10^{-1}$	0.06	70	0.44	53	35	44	26	0.67	0.44
GB17831	0.05	0.06	$3.6x10^{-1}$	0.01	79	0.50	44	35	44	35	0.56	0.44
GB17444	0.42	0.12	$4.\ 5x10^{-4}$	0.01	61	0.39	54	43	36	25	0.68	0.54
GB12917	0.53	0.10	$2.4x10^{-7}$	0.09	53	0.34	46	59	20	33	0.58	0.75
GB19938	0.14	0.09	$1.1x10^{-1}$	0.20	73	0.46	36	49	30	43	0.46	0.62
GB13652	0.47	0.12	$1.0x10^{-4}$	0.11	59	0.37	52	47	32	27	0.66	0.59
GB18753	-0.00	0.12	$9.7x10^{-1}$	0.00	83	0.53	31	44	35	48	0.39	0.56

ID: honey bee gene identifier; Est. (Estimate): Log odds of forager relative to one-day-old; SE: standard error; R^2: coefficient of determination; Error: Number of incorrectly classified; observations; Class. Error rate: Classification error rate; TP: true positives (foragers); TN: true negatives (one-day-olds); FP: false positives; FN: false negatives; Sens. (Sensitivity): TP/(TP+FN); Spec. (Specificity): TN/(TN+FP)

Table 3.4. Results from the logistic analysis of the complete data set using the top 5 genes simultaneously.

ID	Estimate	SE	P-value
GB18624	3.25	0.78	0.000
GB15988	-2.74	1.02	0.007
GB15106	1.84	0.85	0.030
GB10622	0.44	0.40	0.272
GB19745	0.37	0.51	0.461
Error rate			**0.05063 (8/158)**

ID: honey bee gene identifier; Estimate: Log odds of forager relative to one-day-old; SE: standard error; Error rate: classification error rate (misclassified samples).

Table 3.5. Pearson correlation coefficients (top number) for the top 25 genes, and corresponding P-values (bottom number). Genes are honey bee identifiers (e.g. 18624 corresponds to honey bee gene GB18624).

	18624	**15988**	**15106**	**10622**	**19745**	**12396**	**30320**	**10622**	**30213**	**17538**	**14388**	**19724**	**30033**	**19018**	**11892**
18624	1	0.25 0.0015	-0.44 <.0001	0.46 <.0001	-0.57 <.0001	0.02 0.8149	0.87 <.0001	0.08 0.3098	-0.38 <.0001	-0.02 0.8270	-0.72 <.0001	-0.03 0.7310	0.52 <.0001	-0.19 0.0150	0.55 <.0001
15988		1	-0.57 <.0001	-0.11 0.1602	-0.69 <.0001	0.46 <.0001	0.30 0.0001	-0.31 <.0001	-0.15 0.0666	-0.42 <.0001	-0.09 0.2438	-0.001 0.9900	0.22 0.0048	0.30 0.0002	0.42 <.0001
15106			1	-0.20 0.0114	0.58 <.0001	-0.14 0.0731	-0.39 <.0001	0.11 0.1731	0.40 <.0001	0.40 <.0001	0.26 0.0012	-0.17 0.0308	-0.34 <.0001	-0.16 0.0514	-0.67 <.0001
10622				1	-0.17 0.0306	-0.30 0.000	0.49 <.0001	0.73 <.0001	-0.60 <.0001	0.23 0.0033	-0.51 <.0001	0.32 <.0001	0.59 <.0001	-0.39 <.0001	0.32 <.0001
19745					1	-0.55 <.0001	-0.71 <.0001	0.05 0.5125	0.30 0.0001	0.21 0.0070	0.37 <.0001	0.15 0.0638	-0.56 <.0001	0.02 0.7658	-0.51 <.0001
12396						1	0.27 0.0007	-0.20 0.0110	0.29 0.0002	-0.14 0.0696	0.26 0.0010	-0.61 <.0001	-0.01 0.9235	0.06 0.4232	-0.14 0.0811
30320							1	0.25 0.0017	-0.40 <.0001	0.03 0.6844	-0.60 <.0001	-0.17 0.0299	0.67 <.0001	-0.27 0.0007	0.45 <.0001
10622								1	-0.24 0.0026	0.25 0.0019	-0.02 0.7638	-0.01 0.8852	0.22 0.0055	-0.37 <.0001	-0.11 0.1594
30213									1	-0.02 0.8207	0.53 <.0001	-0.61 <.0001	-0.78 <.0001	0.18 0.0275	-0.62 <.0001
17538										1	-0.11 0.1679	-0.05 0.5027	0.08 0.3236	-0.37 <.0001	-0.15 0.0578
14388											1	-0.38 <.0001	-0.69 <.0001	0.27 0.0007	-0.57 <.0001
19724												1	0.44 <.0001	0.12 0.1441	0.43 <.0001
3003													1	-0.26 0.0009	0.59 <.0001
19018														1	-0.04 0.6497
11892															1

Table 3.5 (cont.)

	30514	10744	30434	11814	17831	17444	12917	19938	13652	18753
18624	-0.34 <.0001	-0.02 0.8479	-0.45 <.0001	-0.66 <.0001	-0.67 <.0001	-0.32 <.0001	0.70 <.0001	-0.19 0.0154	0.19 0.0174	0.37 <.0001
15988	-0.27 0.0007	0.38 <.0001	-0.53 <.0001	-0.55 <.0001	-0.52 <.0001	-0.65 <.0001	0.00 0.9580	-0.26 0.0012	-0.30 0.0001	0.06 0.4511
15106	0.49 <.0001	-0.46 <.0001	0.40 <.0001	0.53 <.0001	0.55 <.0001	0.42 <.0001	-0.16 0.0449	0.15 0.0580	0.19 0.0148	-0.28 0.0004
10622	-0.51 <.0001	0.11 0.1880	-0.06 0.4198	-0.39 <.0001	-0.42 <.0001	-0.01 0.9495	0.27 0.0007	0.27 0.0007	0.32 <.0001	0.49 <.0001
19745	0.48 <.0001	-0.16 0.0426	0.71 <.0001	0.81 <.0001	0.86 <.0001	0.81 <.0001	-0.23 0.0037	0.21 0.0071	0.04 0.6180	-0.23 0.0036
12396	-0.16 0.0475	-0.05 0.5187	-0.55 <.0001	-0.42 <.0001	-0.44 <.0001	-0.71 <.0001	-0.23 0.0032	-0.12 0.1342	-0.30 0.0001	-0.44 <.0001
30320	-0.52 <.0001	0.08 0.3257	-0.54 <.0001	-0.84 <.0001	-0.85 <.0001	-0.56 <.0001	0.47 <.0001	0.05 0.5322	0.28 0.0004	0.33 <.0001
10622	-0.31 <.0001	0.21 0.0072	0.23 0.0035	-0.12 0.1225	-0.15 0.0585	0.06 0.4613	-0.12 0.1386	0.62 <.0001	0.16 0.0419	0.10 0.2197
30213	0.55 <.0001	-0.26 0.0012	0.17 0.0330	0.43 <.0001	0.45 <.0001	0.07 0.3660	-0.29 0.0002	-0.15 0.0558	-0.40 <.0001	-0.74 <.0001
17538	0.19 0.0142	-0.42 <.0001	0.04 0.5825	0.22 0.0059	0.14 0.0786	0.20 0.0108	0.21 0.0085	0.02 0.8112	0.47 <.0001	0.17 0.0316
14388	0.23 0.0044	0.19 0.0182	0.31 <.0001	0.47 <.0001	0.48 <.0001	0.03 0.6987	-0.73 <.0001	0.24 0.0026	-0.49 <.0001	-0.72 <.0001
19724	-0.23 0.0030	0.13 0.0969	0.15 0.0526	0.05 0.5111	0.07 0.3830	0.38 <.0001	0.14 0.0707	-0.04 0.6252	0.28 0.0004	0.64 <.0001
30033	-0.67 <.0001	0.09 0.2584	-0.43 <.0001	-0.72 <.0001	-0.75 <.0001	-0.35 <.0001	0.32 <.0001	0.10 0.1960	0.51 <.0001	0.72 <.0001
19018	0.02 0.7652	0.14 0.0802	-0.03 0.6648	0.04 0.6449	0.12 0.1430	-0.10 0.2023	-0.28 0.0003	-0.23 0.0041	-0.47 <.0001	-0.30 0.0001
11892	-0.37 <.0001	0.27 0.0007	-0.41 <.0001	-0.42 <.0001	-0.48 <.0001	-0.18 0.0278	0.50 <.0001	-0.26 0.0011	0.18 0.0251	0.72 <.0001
30514	1	-0.39 <.0001	0.29 0.0002	0.72 <.0001	0.70 <.0001	0.43 <.0001	0.14 0.0780	-0.33 <.0001	-0.06 0.4288	-0.26 0.0010
10744		1	0.14 0.0818	-0.23 0.0034	-0.18 0.0276	-0.14 0.0700	-0.39 <.0001	0.55 <.0001	-0.15 0.0667	0.11 0.1804
30434			1	0.62 <.0001	0.67 <.0001	0.69 <.0001	-0.31 <.0001	0.44 <.0001	0.11 0.1723	-0.13 0.0918
11814				1	0.96 <.0001	0.76 <.0001	-0.16 0.0448	-0.07 0.3816	-0.07 0.4036	-0.23 0.0035
17831					1	0.77 <.0001	-0.22 0.0065	-0.01 0.9208	-0.10 0.2244	-0.29 0.0003
17444						1	0.09 0.2526	0.07 0.4158	0.23 0.0044	0.12 0.1284
12917							1	-0.49 <.0001	0.32 <.0001	0.52 <.0001
19938								1	0.17 0.0285	-0.01 0.8582
13652									1	0.63 <.0001
18753										1

Table 3.6. Results from the logistic analysis of the complete data set after stepwise selection among the top 25 genes considered simultaneously.

ID	Estimate	SE	P-value
GB18624	10.12	5.73	0.075
GB15988	-7.70	3.97	0.077
GB30033	-2.41	1.41	0.053
GB30434	6.21	3.33	0.063
Error rate			**0.00000 (0/158)**

ID: honey bee gene identifier; Estimate: Log odds of forager relative to one-day-old; SE: standard error; Error rate: classification error rate.

Table 3.7. Linear discriminant functions for behavior for the top 25 genes. Each column represents the vector of coefficients of the classification function that maximizes the ratio of a sample being a forager (F) or a one-day-old (O).

ID (Variable)	Behavior coefficient[1] (F)	Behavior coefficient (O)
GB18624	-1.56483	1.59543
GB15988	-2.71409	-3.70037
GB15106	-0.72034	3.97120
GB10622*	1.83171	3.76634
GB19745	-0.12960	0.63549
GB12396	2.97713	2.02977
GB30320	1.57043	2.84900
GB10622+	-2.84328	-4.24491
GB30213	0.58803	0.02983
GB17538	0.48205	0.59414
GB14388	6.95809	5.41413
GB19724	10.28785	5.53957
GB30033	-2.05599	-3.80879
GB19018	-10.65400	-13.49069
GB11892	-1.41542	-1.83137
GB30514	9.35576	8.73056
GB10744	2.33978	3.28939
GB30434	8.15221	12.10756
GB11814	-7.64590	-5.82944
GB17831	-0.57953	-2.37744
GB17444	-3.38381	-2.16820
GB12917	1.66191	3.93477
GB19938	-2.13896	-1.75997
GB13652	-3.66688	-2.74976
GB18753	4.78901	2.37955

[1]Coefficients are interpreted as regression coefficients, where larger coefficients contribute the most to the prediction of behavior; *Corresponds to BB170018B20G10; + Corresponds to BB170018B10A09.

Table 3.8. Stepwise discriminant analysis of the top 25 genes using stepwise, forward and backward selection methods. Gene identifiers in bold are those unique among all methods combined.

Selection Method	Gene Identifiers
Stepwise	**GB15988, GB18624, GB15106, GB30033, GB19018, GB30434, GB19724, GB10622[1], GB12917, GB19745, GB17444, GB10622[2], GB30320, GB12396**
Forward	GB15988, GB18624, GB15106, GB30033, GB19018, GB30434, GB19724, GB10622[1], GB12917, GB19745, **GB13652**, GB30320, GB10622[2], GB12396
Backward	GB18624, GB15988, GB15106, GB10622[1], GB12396, GB30320, GB10622[2], GB19724, GB30033, GB19018, GB30434, GB17444, GB12917

[1]Corresponds to EST identifier BB170018B20G10; [2]Corresponds to EST identifier BB170018B10A09

Table 3.9. Identification of additional sets of biomarker using the score selection procedure for a logistic regression model with specified candidate subsets of 1-5 genes. Model rank is the given rank from 1-10 (for 10 specified "best" model results) from the output, gene combinations are the results per model for combinations of 1-5 genes, Principal Component Analysis (PCA) gives the results of the first three principal components (PCs) where applicable, and error rate is the percentage of samples misclassified when the corresponding genes are used to classify all 158 observations. The average error rate is the average for all 10 models.

Model rank	***Gene Set***	***PCA (PC variation for PC 1, PC 2, PC3)[proportion]***	***Chi-square***	***Error rate [# misclassified/158 observations]***
		Models with 1 gene		
1	GB10220	-	53.72	25% (40/158)
2	BI515625	-	52.14	31% (49/158)
3	GB10854	-	47.93	26.6% (42/158)
4	BI504888	-	43.73	31% (49/158)
5	GB16009	-	42.12	26% (41/158)
6	GB15988	-	41.56	29% (46/158)
7	GB19468	-	40.31	35% (55/158)
8	GB10310	-	40.03	32% (50/158)
9	GB19018	-	35.39	24.7% (39/158)
10	GB14157	-	35.18	32.3% (51/158)
			Average error rate	**29.26%**

Table 3.9 (cont.)

		Models with 2 genes		
1	GB19547, GB10310	54%, 46%	111.11	9.5% (15/158)
2	BI515625, GB15988	87%, 13%	106.40	7% (11/158)
3	GB10310, BI509882	53%, 47%	106.09	9.5% (15/158)
4	GB14166, BI509882	55.7%, 44.3%	105.32	6.3% (10/158)
5	BI515625, GB10310	91%, 9%	95.75	14% (22/158)
6	GB15988, GB19547	67%, 33%	94.98	11.4% (18/158)
7	BI511603, GB16547	94.7%, 5.3%	93.79	10.8% (17/158)
8	GB15988, GB18624	77.6%, 22.4%	93.61	13.3% (21/158)
9	BI511603, GB30320	85%, 15%	93.39	7.6% (12/158)
10	GB15988, BI509882	66%, 34%	93.36	10% (16/158)
			Average error rate	**9.94%**
		Models with 3 genes		
1	BI516675, GB10310, BI509882	59%, 35%, 6%	121.27	3.8% (6/158)
2	BI515625, GB16694, GB16986	70.3%, 23.4%, 6.3%	120.73	2.5% (4/158)
3	GB14166, GB10310, BI509882	64%, 33%, 3%	120.69	4.4% (7/158)
4	BI515625, GB10310, BI509882	65%, 32%, 3%	120.67	3.8% (6/158)
5	BI515625, GB15988, BI509882	67.1%, 24.7%, 8.2%	120.64	3.8% (6/158)
6	GB30033, GB10310, GB30320	51.4%, 42.4%, 6.2%	118.94	6.3% (10/158)
7	BI515625, GB15988, BI505018	59%, 33%, 8%	118.77	1.3% (2/158)
8	BI516675, GB19547, GB10310	58.6%, 34.7%, 6.7%	118.75	3.8% (6/158)
9	BI515625, GB10310, GB19722	61%, 36%, 3%	118.53	4.4% (7/158)
10	GB19193, GB10220, BI509882	64%, 27%, 9%	117.95	5.1% (8/158)
			Average error rate	**3.92%**
		Models with 4 genes		
1	GB19724, GB10220, BI515625, GB19745	65%, 27%, 7%	132.62	2.5% (4/158)
2	GB14166, BI515625, GB15988, BI509882	66%, 23%, 6%	132.25	1.3% (2/158)
3	GB19724, GB17707, GB18624, GB10310	61%, 31%, 7%	132.20	0% (0/158)
4	GB19724, BI515625, GB10310, BI509882	52%, 43%, 3%	132.18	1.3% (2/158)

Table 3.9 (cont.)

5	BI513123, BI514288, GB10310, BI509882	63%, 32%, 3%	132.15	0.6% (1/158)
6	GB19724, GB18177, BI514288, GB10220	67%, 29%, 3%	131.96	0% (0/158)
7	BI515625, BI515450, GB10310, BI509882	50%, 45%, 3%	131.82	2.5% (4/158)
8	GB19724, GB17707, GB10310, BI509882	56%, 38%, 5%	131.80	2% (3/158)
9	BI516304, GB14166, BI504663, BI509882	53%, 41%, 5%	131.73	2% (3/158)
10	BI513123, GB10220, BI515625, GB16694	68%, 26%, 4%	131.65	0.6% (1/158)
			Average error rate	**1.28%**

		Models with 5 genes		
1	GB20076, GB19724, GB18177, GB10220, BI515625	57.3%, 35.5%, 4.2%	135.50	0% (0/158)
2	GB19724, GB10854, GB10220, BI515625, GB17707	68.1%, 25.6%, 3.6%	135.30	0% (0/158)
3	GB19724, GB18177, GB10220, BI515625, BI508683	64.9%, 27.8%, 3.6%	134.99	0.6% (1/58)
4	GB19724, GB18177, GB10220, BI515625, BI509882	53%, 41%, 3%	134.77	1.9% (3/158)
5	GB19724, GB10220, BI515625, GB17707, BI509882	55.5%, 39.5%, 3.1%	134.64	0% (0/158)
6	GB19724, GB18177, GB10854, GB10220, BI515625	65.5%, 26.9%, 3.6%	133.65	1.3% (2/158)
7	GB19724, GB18177, GB10220, BI515625, GB17707	66.9%, 28.3%, 2.4%	133.09	0.6% (1/158)
8	GB19724, GB18177, GB10220, BI515625, GB18624	65%, 29%, 3%	132.39	1.9% (3/158)
9	GB19724, GB10220, BI515625, GB17707, BI509910	61%, 26%, 10%	132.33	0.6% (1/158)
10	GB19724, GB18177, GB10220, BI515625, BI509910	57%, 27%, 12%	132.25	1.9% (3/158)
			Average error rate	**0.88%**

Table 3.10. Results from the leave-one-study-out analyses after the final logistic model.

Testing Study	Genes	Error Rate	Error Rate
AC	GB18624, GB15988, GB30434	1/24	0.04
AD	GB18624, GB15988, GB30033, GB30434	0/24	0
AF	GB18624, GB15988, GB30033, GB30434	4/24	0.17
AM	GB18624, GB15988, GB30033, GB30434	0/22	0
LL	GB18624, GB15988, GB30033, GB30434	0/16	0
LM	GB18624, GB15988, GB30033, GB30434	0/16	0
ML	GB18624, GB15988, GB10622, GB30033	3/16	0.19
MM	GB18624, GB15988, GB30033, GB30434	0/16	0

Summary (Genes identified in many studies)	
8 studies	GB18624, GB15988
7 studies	GB30033, GB30434
1 study	GB10622

Error rate: classification error rate.
Classification was completed using 7/8 studies as training sets, and the remaining study as the testing set.

Table 3.11. Results from the leave-one-study-out analyses, using three methods to select genes for each individual study.

Testing Study	Gene Identifiers	Error Rate	Error Rate
AC	GB18624, GB15988, GB30434, GB30033, GB19724	0/24	0
AD	GB18624, GB15988, GB30434, GB19724, GB12917	0/24	0
AF	GB18624, GB15988, GB30033, GB30434, GB19724	0/24	0
AM	GB18624, GB15988, GB30033, GB30434, GB19724	0/22	0
LL	GB18624, GB15988, GB30434, GB15106, GB19724	0/16	0
LM	GB18624, GB15988, GB30434, GB19724, GB17444	0/16	0
ML	GB18624, GB15988, GB30033, GB30320, GB12917	3/16	0.19
MM	GB15988, GB30434, GB30320, GB19724, GB12917	0/16	0

Summary (Genes identified in many studies)	
8 studies	GB15988
7 studies	GB18624, GB30434, GB19724
4 studies	GB30033
3 studies	GB12917
2 studies	GB30320
1 study	GB15106, GB17444

Error rate: classification error rate.
Classification was completed using 7/8 studies as training sets, and the remaining study as the testing set. Genes were selected using a combination of backward, forward, and stepwise approaches.

Table 3.12. Results from the leave-one-study-out analyses of the score results.

Testing Study	Genes	Error Rate	Error Rate
AC	GB10310, BI509882, GB14166, BI513123, BI508683	1/24	0.04
AD	BI515625, GB14157, BI509882, GB14166, BI513123	0/24	0
AF	GB14157, BI513447, GB17707, GB14166, BI508683	5/24	0.21
AM	GB14157, BI509882, GB14166, BI513123, BI509910	0/22	0
LL	BI515625, GB10310, BI509882, BI516304, BI510708	0/16	0
LM	GB14157, BI509882, GB16694, GB14166, BI513123	0/16	0
ML	GB14157, BI509882, GB16694, GB14166, BI513123	0/16	0
MM	GB14157, BI509882, GB14166, BI513123, BI508683	1/16	0.06

Summary (Genes identified in many studies)	
8 studies	-
7 studies	BI509882, GB14166
6 studies	GB14157, BI513123
3 studies	BI508683
2 studies	BI515625, GB10310, GB16694
1 study	BI516304, BI513447, BI510708, BI509910, GB17707

Error rate: classification error rate.
Classification was completed using 7/8 studies as training sets, and the remaining study as the testing set. Genes were selected using the score procedure in logistic regression.

Table 4.1. Number of differentially expressed genes between samples created using artificial insemination (AI) and somatic cell nuclear transfer (NT).

		Individual analyses								Meta-analyses			
		d280X[1]		d7E		d25X		d25E		Study		Sample	
		+	-	+	-	+	-	+	-	+	-	+	-
d280X	+	189[2]	-	0[3]	1	1	0	1	1	2	0	3	0
	-	-	657	1	2	0	1	0	0	0	14	1	19
d7E	+	0.0%[4]	1.9%	52	-	0	0	0	4	0	0	26	0
	-	1.4%	2.7%	-	74	0	0	0	1	0	0	1	26
d25X	+	4.5%	0.0%	0.0%	0.0%	22	-	0	1	0	0	0	0
	-	0.0%	7.1%	0.0%	0.0%	-	14	0	0	0	0	0	1
d25E	+	10.0 %	0.0%	0.0%	0.0%	0.0 %	0.0 %	10	-	0	0	0	0
	-	2.6%	0.0%	10.5 %	2.6%	2.6 %	0.0 %	-	38	0	1	4	1
Study	+	3.0%	0.0%	0.0%	0.0%	0.0 %	0.0 %	0.0 %	0.0 %	66	-	9	0
	-	0.0%	13.7 %	0.0%	0.0%	0.0 %	0.0 %	0.0 %	1.0 %	-	102	0	10
Sample	+	1.5%	0.5%	12.7 %	0.5%	0.0 %	0.0 %	0.0 %	2.0 %	4.4 %	0.0 %	20 4	-
	-	0.0%	10.8 %	0.0%	14.8 %	0.0 %	0.6 %	0.0 %	0.6 %	0.0 %	5.7 %	-	17 6

[1]d280X = day 280 extra-embryonic samples, d7E = day 7 embryos, d25X = day 25 extra-embryonic sample, d25E = day 25 embryos.

[2]Number of differentially expressed genes between AI and NT (P-value < $1x10^{-3}$, fold change $\geq$ 1.23) within individual analyses, study-level standardized (Study), and sample-level (Sample) meta-analyses (diagonals).

[3]Upper off-diagonals are the number of genes identified differentially expressed in all pairs of analyses relative to the maximum number of significant genes that can overlap in both analyses.

[4]Lower off-diagonals are the percentage of genes identified differentially expressed in all pairs of analyses relative to the maximum number of significant genes that can overlap in both analyses.

Table 4.2. Number of differentially and non-differentially expressed genes between samples created using artificial insemination (AI) and somatic cell nuclear transfer (NT) within combinations of individual (Ind) studies (d280X, d7E, d25X, d25E), study-level (Study) meta-analysis and sample-level (Sample) meta-analysis.

	Ind	**Ind and Study**	**Ind and Sample**	**Ind, Study and Sample**
Non-significant	12,215	151	302	17
Sig. in 1 study	1,028	17	74	2
Sig. in 2 studies	14	0	4	0
Total	13,257	168	380	19

Table 4.3. Gene information for 8 cattle genes with significant differential expression in two individual studies with opposite signs.

Gene	***Gene Bank Bovine ID***	***Gene Name***	***Over-expressed (+) AI vs NT***	***Under-expressed (-) AI vs NT***
OLIGO_07366	*CK769368*	*PFDN2* - prefoldin subunit 2	d7E[1]	d25E
OLIGO_08124	*CN440810*	*POLG2* - polymerase (DNA directed), gamma 2, accessory subunit	d7E	d25E
OLIGO_00784	*CR454884*	*POLE2* - polymerase (DNA directed), epsilon 2 (p59 subunit)	d7E	d280X
OLIGO_03653	*DV784696*	*HAND2* - heart and neural crest derivatives expressed 2	d280X	d7E
OLIGO_05820	*BC111227*	*POLR3H* - polymerase (RNA) III (DNA directed) polypeptide H	d25X	d25E
OLIGO_06019	*DT722240*	*BRRN1* - non-SMC condensin I complex, subunit H	d7E	d25E
OLIGO_06762	*CX948597*	*MTFR1* - mitochondrial fission regulator 1	d280X	d7E
OLIGO_08413	*M74083*	*FNTA* - farnesyltransferase, CAAS box, alpha	d7E	d25E

[1]d280X = day 280 extra-embryonic samples, d7E = day 7 embryos, d25X = day 25 extra-embryonic sample, d25E = day 25 embryos.

Table 4.4. Differentially expressed genes between conditions within developmental stages (7 d, 25 d, 280 d), sample sources (extra-embryonic or X, embryonic or E) and reproductive technologies (AI, and NT).

Variables and corresponding levels		***Within technology and source***				***Within technology and stage***		***Within source***		***Within source and stage***				***Within stage***
Reproductive technology	***AI***	✓[1]		✓		✓		✓	✓	✓	✓	✓	✓	✓
	NT		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓
Sample source	***E***	✓	✓			✓	✓	✓		✓	✓			✓
	X			✓	✓	✓	✓		✓			✓	✓	✓
Developmental stage	***7 d***	✓	✓					✓		✓				
	25 d	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓
	280 d			✓	✓				✓				✓	
No. of genes		**483**	**1149**	**340**	**334**	**82**	**11**	**530**	**551**	**16**	**4**	**1**	**21**	**39**

[1]Checks depict the developmental stages, sources and technologies contrasted, and the last row contains the number of differentially expressed genes (P-value < $1x10^{-3}$ and fold change >1.23).

APPENDIX A: CHAPTER 2 GENE ONTOLOGY AND FRUIT FLY ANNOTATION INFORMATION FOR SAMPLE-LEVEL META-ANALYSIS RESULTS

Gene Ontology (molecular function or mol. function, biological process or bio. process, and cellular component or cell. component) and fruit fly information for 125 *Apis mellifera* transcripts with significant differential expression in only the sample-level meta-analysis may be found in a supplemental file named '**Appendix A.pdf**'.

APPENDIX B: CHAPTER 4 RESULTS FOR INDIVIDUAL STUDIES COMPARING AI AND NT

Results of comparison of AI versus NT for individual studies (d7E, d25E, d25X, d280X) may be found in a supplemental file named '**Appendix B.pdf**'.

APPENDIX C: CHAPTER 4 RESULTS FOR STUDY-LEVEL AND SAMPLE-LEVEL META-ANALYSES COMPARING AI AND NT

Results of comparison of AI versus NT for study-level and sample-level meta-analyses may be found in a supplemental file named '**Appendix C.pdf**'.

APPENDIX D: CHAPTER 4 RESULTS OF CONTRASTS OF DEVELOPMENTAL STAGE WITHIN REPRODUCTIVE TECHNOLOGY AND TISSUE SOURCE

Results of contrasts based on developmental stage (8 d, 25 d, 280 d) for AI and NT embryonic and extra-embryonic samples may be found in a supplemental file named '**Appendix D.pdf**'.

APPENDIX E: CHAPTER 4 RESULTS OF CONTRASTS OF TISSUE SOURCE WITHIN REPRODUCTIVE TECHNOLOGY AT 25 DAYS GESTATION

Results of contrasts based on tissue source (embryonic or extra-embryonic sample) for AI and NT at 25 d may be found in a supplemental file named '**Appendix E.pdf**'.

APPENDIX F: CHAPTER 4 RESULTS OF CONTRASTS OF REPRODUCTIVE TECHNOLOGY FOR 25 DAY AND 280 DAY EXTRA-EMBRYONIC SAMPLES

Results of contrasts of AI versus NT for extra-embryonic samples at 25 d and 280 d may be found in a supplemental file named '**Appendix F.pdf**'.

APPENDIX G: CHAPTER 4 RESULTS OF CONTRASTS OF REPRODUCTIVE TECHNOLOGY FOR 8 DAY AND 25 DAY EMBRYONIC SAMPLES

Results of contrasts of AI versus NT for embryo samples at 8 d and 25 d may be found in a supplemental file named '**Appendix G.pdf**'.

Author's Curriculum Vitae

Heather Ann Adams
10488 Lynn Road
Byron, IL 61010
(217) 369-3081
hadams1@illinois.edu

Education

Ph.D., **Animal Sciences**, University of Illinois at Urbana-Champaign, Urbana, IL, 2010
Specialization: Bioinformatics
Dissertation title: "Enhanced transcriptome profiling and biomarker discovery using meta-analytical techniques"
Advisor: Dr. Sandra L. Rodriguez-Zas, Ph.D.

M.S., **Animal Sciences**, University of Illinois at Urbana-Champaign, Urbana, IL, 2006
Specialization: Quantitative Genetics
Thesis title: "Analyzing the impact of inbreeding and heterozygosity on predicted transmitting ability survival through sire family analysis and simulation"
Advisor: Dr. Roger D. Shanks, Ph.D.

B.S., **Animal Sciences**, University of Illinois at Urbana-Champaign, Urbana, IL, 2004
Minor: Chemistry

Continuing Education

Meta-Analysis: Statistical Methods for Combining the Results of Independent Studies
American Statistical Association; Workshop diploma, Salt Lake City, UT, 2007

Workshop on Statistical Analysis of Microarray Data
University of Illinois at Urbana-Champaign, Urbana, IL, 2007

Teaching Experience

Instructor, Animal Sciences 444: Applied Animal Genetics, University of Illinois at Urbana-Champaign, Spring 2009
- Taught quantitative genetics (applications to animal improvements, including methods of selection and mating designs)

Teaching Assistant, Animal Sciences 444: Applied Animal Genetics, Spring 2005/2006
- Held weekly office hours
- Assisted professor in teaching simulation project for genetic improvement of dairy cattle

Informal Teaching Experience

Guest Lecturer, Animal Sciences 400: Dairy Herd Management, University of Illinois at Urbana-Champaign, Fall 2008

- Taught SIMBULL (computer simulation of genetic improvement in dairy cattle)
- Developed class project, and graded final projects

Guest Lecturer, Animal Sciences 545: Statistical Genomics, University of Illinois at Urbana-Champaign, Spring 2008

- Taught Beehive (suite of web-based tools to analyze gene expression experiments) to graduate students

Mentor, Department of Animal Sciences, University of Illinois at Urbana-Champaign, Spring 2004

- Supervised high school student in internship project (dairy cattle breeding and genetics)

Professional Experience

Graduate research assistant, Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 2007-2009

- Analysis of microarray gene expression data

Graduate research assistant, Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, 2007-2010

- Meta-analysis of gene expression experiments

Graduate research assistant, Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, 2004-2006

- Analysis of sire, daughter and individual inbreeding values on PTA survival of registered Holstein bulls

Honors, Recognitions

Gamma Sigma Delta, Honor Society of Agriculture – University of Illinois Chapter, 2010

David H. and Norraine A. Baker Graduate Fellowship in Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, 2009-2010

Graduate College Travel Award, University of Illinois at Urbana-Champaign, Urbana, IL, 2007

Invited Lectures and Invited Conference Presentations

Alltech Distinguished Lecture Series in Nutrigenomics, University of Kentucky, Lexington, KY, 2008
Title: "An introduction to Beehive"

2nd International Meeting on Mammalian Embryogenomics, INRA-Jouy en Josas, Paris, France, 2007
Title: "Statistical analyses of gene expression data"

Conferences/Symposia Attended

5th Annual BeeSpace Workshop (presentation), University of Illinois at Urbana-Champaign, Urbana, IL, May 2009

Statistical Genetics of Livestock for the Post-Genomic Era Symposium (poster), University of Wisconsin–Madison, Madison, WI, May 2009

Joint Statistical Meetings, American Statistical Association (presentation), Denver, CO, August 2008

IGB Fellows Symposium (poster), University of Illinois at Urbana-Champaign, Urbana, IL, April 2008

2nd International Meeting on Mammalian Embryogenomics (presentation and poster), INRA-Jouy en Josas, Paris, France, October 2007

Joint Statistical Meetings, American Statistical Association (presentation), Salt Lake City, UT, July 2007

CSHL Workshop on Honey Bee Genomics & Biology, Cold Spring Harbor, NY, May 2007

IGB Fellows Symposium, University of Illinois at Urbana-Champaign, Urbana, IL, February 2007

Technical Skills

MS Windows, UNIX; MS Office, **SAS**; R, PERL, C++

Articles in Journals

1. **Adams H. A.**, B. R. Southey, R. E. Everts, S. L. Marjani, X. C. Tian, H. A. Lewin, and S. L. Rodriguez-Zas. 2010. Meta-analysis of gene expression studies provides insights into cattle embryo development across reproductive techniques, developmental stages and sources. *Funct Integ Genomics* (Submitted).
2. Alaux C., Y. Le Conte, **H. Adams**, S. Rodriguez-Zas, C. Grozinger, S. Sinha, and G. Robinson. 2009. Regulation of brain gene expression in honey bees by brood pheromone. *Genes Brain Behav* 8(3): 309-319.
3. **Adams H. A.**, B. R. Southey, G. E. Robinson, and S. L. Rodriguez-Zas. 2008. Meta-analysis of genome-wide expression patterns associated with behavioral maturation in honey bees. *BMC Genomics* 9: 503.
4. Rodriguez-Zas S. L., Y. Ko, **H. A. Adams**, and B. R. Southey. 2008. Advancing the understanding of the embryo transcriptome co-regulation using meta-, functional, and gene network analysis tools. *Reproduction* 135: 213–224.

Bulletins, Reports, or Conference Proceedings

1. **Adams H.A**., S.L. Rodriguez-Zas, and B.R. Southey. 2008. A novel multi-stage approach to classify samples across multiple gene expression experiments. 2008 Proceedings of the American Statistical Association, Biometrics Section, Alexandria, VA: American Statistical Association.
2. **Adams H. A.**, S. L. Rodriguez-Zas, and B. R. Southey. 2007. Comparison of meta-analytical approaches for gene expression profiling. 2007 Proceedings of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Abstracts

1. **Adams H. A.**, S. L. Rodriguez-Zas, and B. R. 2008. A novel multistage approach to classify samples across multiple gene expression experiments. On the Abstract CD of the Joint Statistical Meetings 2008. August 3 - August 7, Denver, CO.
2. **Adams H. A.**, S. L. Rodriguez-Zas, and B. R. Southey. 2007. Comparison of meta-analytical approaches for gene expression profiling. In the Abstract Book of the Joint Statistical Meetings 2007. July 29 - August 2, Salt Lake City, UT. 76p.
3. **Adams H. A.**, B. R. Southey, and S. L. Rodriguez-Zas. 2007. Gene expression profiling of mouse embryos using meta-analysis techniques. In the Abstract Book of the 2nd International Meeting on Mammalian Embryogenomics. October 17-20, 2007. Paris, France.
4. **Adams H. A.** and R. D. Shanks. 2006. Quantifying the impact of multiple independent heterozygous loci on survival. *J. Dairy Sci.* 89(Suppl. 1): 273.

5. **Adams H. A.** and R. D. Shanks. 2006. Effect of genetic defects and inbreeding on PTA survival in a Holstein sire family. 8th World Congress of Genetics Applied to Livestock Production. August 13-18, 2006 BeloHorizonte, MG, Brazil. Communication 01-45, p. 17.

Professional Societies or Associations

American Association for the Advancement of Science

American Statistical Association

Research Goals

My goals are to work towards being a top statistical geneticist, collaborating with researchers in industry, federal agencies and academia. I have mastered a large number of statistical, informatics and computation tools to mine genomic data and support genomic-assisted improvement. My previous research experience has proven my exceptional teamwork abilities, and also my ability to lead independent projects, working towards short and medium term deadlines while pursuing long term goals. Statistical genetics is key to understanding and interpreting the vast amounts of data outputted from gene expression experiments. I aim at working towards interpreting gene characterization of numerous transcriptomes, allowing for the understanding of gene functions, characteristics, and overall biological systems for genetic improvement, sustainability and improved agricultural production, while maintaining health and reproductive performance.

References

Dr. Sandra L. Rodriguez-Zas, Ph.D., Associate Professor of Bioinformatics
Department of Animal Sciences
University of Illinois at Urbana-Champaign
(217) 333-8810, rodrgzzs@illinois.edu

Dr. Roger D. Shanks, Ph.D., Professor of Genetics
Department of Animal Sciences
University of Illinois at Urbana-Champaign
(217) 244-3155, rdshanks@illinois.edu

Dr. Harris A. Lewin, Ph.D., Professor of Immunogenetics
Department of Animal Sciences
Director, Institute for Genomic Biology
University of Illinois at Urbana-Champaign
(217) 244-3034, h-lewin@illinois.edu