Fall 12-2010

# Toxicogenomics Analysis of Non-Model Transcriptomes Using Next-Generation Sequencing and Microarray

Arun Rawat
*University of Southern Mississippi*

The University of Southern Mississippi

# TOXICOGENOMICS ANALYSIS OF NON-MODEL TRANSCRIPTOMES USING NEXT-GENERATION SEQUENCING AND MICROARRAY

by

Arun Rawat

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

December 2010

ABSTRACT

TOXICOGENOMICS ANALYSIS OF NON-MODEL TRANSCRIPTOMES

USING NEXT-GENERATION SEQUENCING AND MICROARRAY

by Arun Rawat

December 2010

With the advent of next generation technologies like Roche/454 Life Sciences that require low cost and less time for sequencing will help in providing a workable draft of non-model species genomes. Availability of high throughput microarray technologies for gene expression profiling provides low-cost tools for investigation of highly-integrated responses to various stimuli. These advancements along with bioinformatics processing have led to an increasing number of non-model species having well-annotated transcriptomes. The project focuses on the life cycle of development, functional annotation, and utilization of genomic tools for the avian wildlife species to determine the molecular impacts of exposure to munitions constituents (MCs).

Massively parallel pyrosequencing is created from the normalized multi-tissue library of Northern bobwhite (*Colinus virginianus*) and Japanese quail (*Coturnix coturnix*) cDNAs. The assembly of next generation sequencing for transcriptomes of these organisms is challenging. High number of ESTs and longer read length require high computational memory and management between the sensitivity and accuracy to assemble correctly. The researcher developed a new pipeline "Contigs Assembly Pipeline using Reference Genome" (CAPRG) to assemble long reads for non-model organisms that have available reference genome. The results were benchmarked by employing parameter space for

different available methods that utilize de novo strategies like overlap-layout-consensus (OLC) and graphs for long reads. It was observed that CAPRG performance was better or near equivalent in the two transcriptomic datasets based on different benchmarks but also completes the assembly in a fraction of the time as compared to assemblers that yield competitive results.

The researcher performed statistical analysis to generate differentially expressed genes and utilized metabolic maps, biological networks, pathway analysis and GO enrichment to the differentially-expressed genes in the livers of birds exposed for 60 days (d) to 10 and 60 mg/kg/d 2,6-DNT. These revealed insights into the metabolic perturbations underlying several observed toxicological phenotypes. The impacts were validated by RT-qPCR including: a shift in energy metabolism toward protein catabolism via inhibition of control points for glucose and lipid metabolic pathways, PCK1 and PPARGC1, respectively.

To greatly expand the information-base for Northern bobwhite that has little supporting information in Genbank, the researcher initiated and developed the web-based knowledgebase (www.quailgenomics.info). The Quail genomics share and develop functional genomic data for Northern bobwhite to allow researchers to perform analysis and curate genomic information for this non-model species.

The University of Southern Mississippi

# TOXICOGENOMICS ANALYSIS OF NON-MODEL TRANSCRIPTOMES USING NEXT-GENERATION SEQUENCING AND MICROARRAY

by

Arun Rawat

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

Mohamed O. Elasri
_____
Director

Edward J. Perkins
_____

Tim McLean
_____

Jonathan Sun
_____

Preetam Ghosh
_____

Susan A. Siltanen
_____
Dean of the Graduate School

December 2010

ACKNOWLEDGMENTS

I am grateful to the committee chair, Dr. Mohamed O. Elasri, and committee members, Dr. Edward J. Perkins, Dr. Tim McLean, Dr. Jonathan Sun and Dr. Preetam Ghosh, for their advice and support for the project.

The laboratory experiments cDNA library, microarray experiments, and RT-qPCR were conducted at Engineer Research and Development Centre (ERDC), Vicksburg, and I am thankful to Dr. Kurt A. Gust and Dr. Edward J. Perkins for providing the data and useful inputs for the analysis for the project.

I am also thankful to Glover George for providing access to parallel computing facilities at School of Computing Sciences and to my lab members, Dr. Tanwir Habib and Dr. Jagan Thodima, for their help in the project. Special thanks to Antony Swartz and other team members in the lab for useful feedback that helped me with the presentation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

# CHAPTER I

# INTRODUCTION AND BACKGROUND

## Motivation

Advancement in ultra high throughput technologies like next generation sequencing and microarray is shaping the landscape of life sciences. The utility of these for advancing disciplines within the life sciences has been broadly documented. The next generation sequencing is greatly expanding the sequencing depth and coverage (Heng Li and Nils Homer, 2010; Margulies *et al*., 2005). This has great significance for non-model organism that are of interest to respective scientific communities as development of non-model species has lagged relative to model species (Ellegren, 2008; Meyer *et al*., 2009; Vera *et al*., 2008). The low cost and less time required for next generation sequencing will help in providing a workable draft of these non-model species genomes (Heng Li and Nils Homer, 2010; Papanicolaou *et al*., 2009). A number of commercial sources (e.g. Agilent Technologies, Nimblegen, Affymetrix) provide on-demand synthesis of high-density oligonucleotide arrays directly from users' annotated sequence databases yielding low-cost tools for investigation of highly-integrated responses to various stimuli (Rawat *et al*., 2010c). Such expression-profiling tools are the engine for expanding the "universal language" with which to describe cellular responses (Lamb *et al*., 2006).

Diverse group of species may be at risk of exposure to munitions compounds (MCs) when utilizing habitat present on military training, munitions manufacturing, and demilitarization facilities (Gust *et al*., 2009; Quinn, Jr. *et al*., 2007). High concentrations of these MCs are detected near military training facilities and in soil next to detonations at army ranges (Jenkins *et al*., 2006). Few of these MCs like dinitrotoluenes are used

commercially in the production of polyurethane foams, coatings, and dyes (Quinn, Jr. *et al*., 2007). People exposed to these energetic compounds have reported health problems like anemia, abnormal liver function and skin irritation (Sabbioni *et al*., 2005; Tchounwou *et al*., 2001). The environmental agency and army is concerned about the impacts on human and wildlife. The analysis from these studies will provide risk-assessment paradigm to land managers that will assist in making remediation decisions (Quinn, Jr. *et al*., 2007).

The wildlife bird species Northern bobwhite and Japanese quail has good representation in habitat near military facilities that have MCs found as soil and water contaminant posing potential risks to these ground foraging birds. The Northern bobwhite has been shown to be important for numerous ecological (Perea *et al*., 2009; Quinn, Jr. *et al*., 2007), methodological (Johnson *et al*., 2005) and environmental management reasons (Sauer *et al*., 2005) making it an excellent experimental avian wildlife model . The Japanese quail is considered to be reproductive model organism (Balthazart *et al*., 1996). Robust understanding of MEC effects in avian receptors is lacking. Advancement in ecotoxicological assessment for MEC effects in avian species is critical to insure the protection of wild populations. Genomics approaches are recognized to provide a data rich/information intensive platform for investigating chemical effects allowing broad characterization of ecotoxicological effects. These genomic investigations are being leveraged to holistically describe MEC effects to improve field risk assessments for wildlife birds.

Development of multi-tissue microarray tools is required to characterize systemic impacts that may result from MC exposure and for discovery of mechanisms of action (MOA) toxicological effects. MOA information is a critical component of ecological risk

assessment and the broad and integrated examination of molecular targets and functional pathways that microarray assays enable is recognized as a key strength for deciphering MOA (Ankley *et al*., 2006). The utility of Northern bobwhite as an avian-wildlife genomic model was demonstrated in a neurotoxicogenomic investigation where the MOA underlying seizures were determined in response to exposure to the munition hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) (Gust *et al*., 2009).

In concert, *in silico* methods for high-throughput sequence assembly and annotation including expressed sequence tag (EST) characterization, Gene Ontology (GO) classification and assignment to functional pathways have risen to the challenge of maximizing biological information connected to the transcriptome (Hu *et al*., 2003; Udall *et al*., 2006; Wang *et al*., 2007). Data intensive applications like information retrieval and data mining pose challenges and parallelization can address these issues. Recently there has been development in bioinformatics to address cost-effective methods for the fast solution of computationally large and data-intensive problems. Various algorithms and programs are built that address the processor and memory issues by parallelization that helps to perform the task.

The project primarily focuses on understanding, utilizing and management of these high throughput technologies, next generation sequencing and microarray to thoroughly assess MC effects in the non model species (Figure 1). This will also include workflow on interaction, development and utilization of various biological data models and genomic tools to understand the systemic perturbation in wildlife birds and development of knowledgebase.

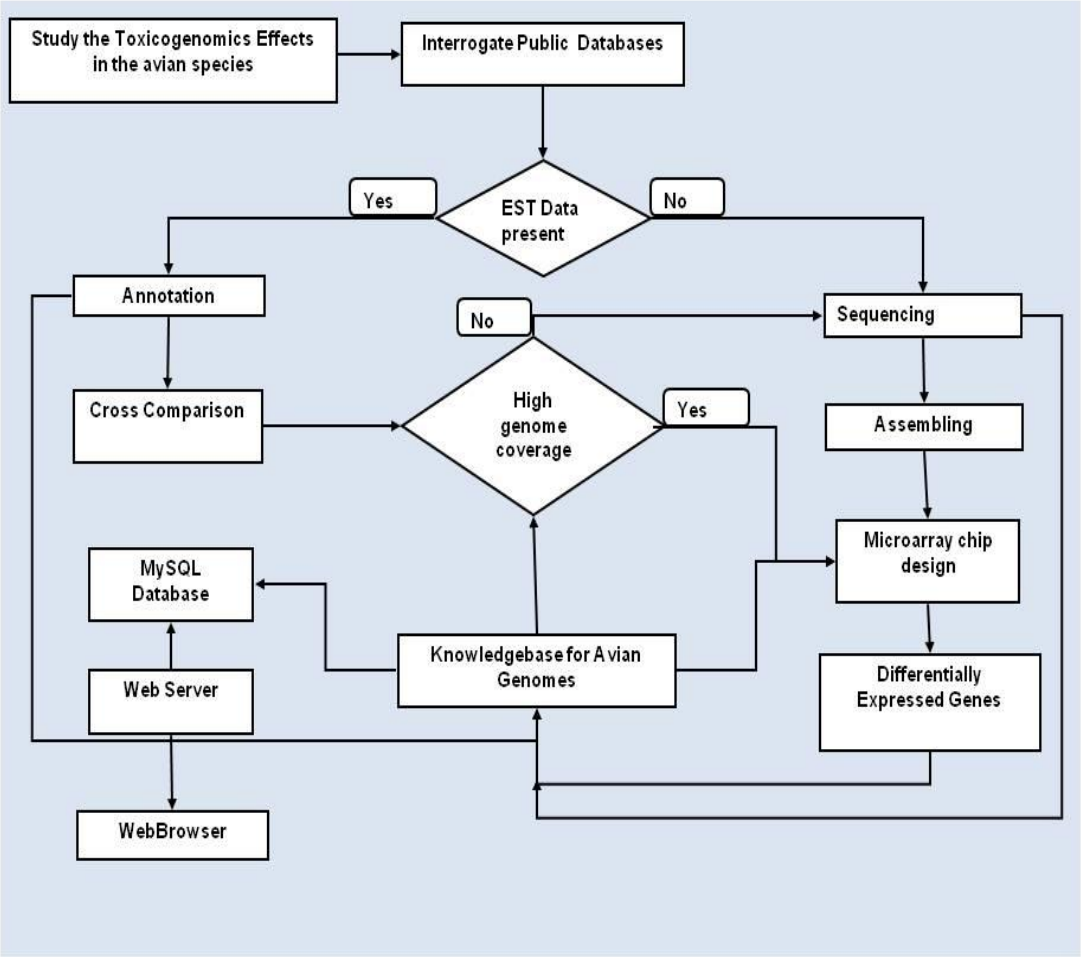


*Figure 1.* The project flowchart representing data components.

The study includes genomic responses to the MC compounds like 2,6-dinitrotoluene (2,6-DNT) for the toxicological model species, the Northern bobwhite and the Japanese quail and Zebrafinch (*Taeniopygia guttata*) (Figure 1). So far, the genomic information for non model organism like Northern bobwhite and Japanese quail is virtually absent in public databases like NCBI with limited number of ESTs. By comparison, avian species such as chicken and zebra finch have been robustly described and whole genome shotgun (WGS) data of chicken (Hillier *et al*., 2004) is available. For most of the analysis, the researcher extensively utilized the existing information from chicken as chicken has

achieved model species status (Cogburn *et al*., 2007). The Northern bobwhite and Japanese quail were sequenced while the existing information of zebrafinch was used after genomic comparisons. The flow chart is an iterative process subsuming various analysis and components described further.

## Background

### *Transcriptome Overview*

Transcriptome can be defined as set of expressed genes for a defined tissue cells modulated by any external or internal factor (Velculescu *et al*., 1997). The genome is static and the gene expression level acts as link between the genome of an organism and physical characteristics (Velculescu *et al*., 1997). A cell responds to external cues like chemical/mechanical/environmental stimulus through mechanism called signal transduction (Reece *et al*., 2002) (Figure 2). Initiation of signal transduction cascades produces transcription factors that activate genes (Lalli and SassoneCorsi, 1994). The initial stimulus triggers the expression of different genes that trigger myriad complex physiological events and eventually change in cell function.

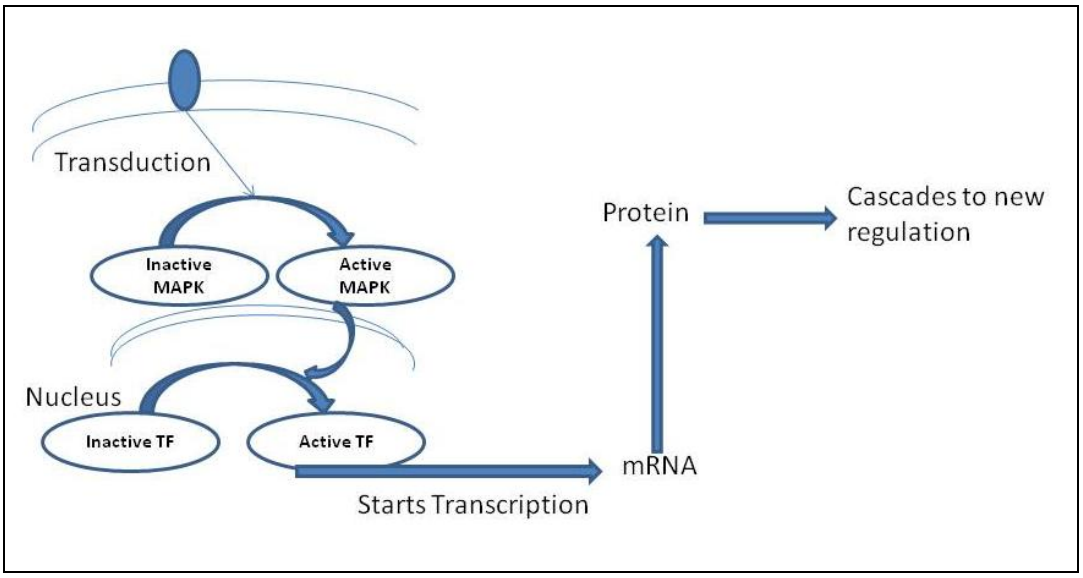


*Figure 2.* The initiation of signal transduction activates transcription. The exposure to stimulus cascades signal transduction activating transcription factor to form mRNA changing state of a cell**.**

Genes are expressed through a process commonly referred to as transcription. Transcription of DNA results in the formation of mRNA through a series of events (Lodish *et al*., 2004) starting with initiation when transcription factors (tf) and RNA Polymerase bind to a specific site of DNA called promoter and locally melts the double stranded DNA. The elongation of mRNA starts as RNA Polymerase moves along the entire length of coding DNA, sequentially melting the DNA and adding nucleotides to the RNA strand. Finally the transcription is terminated when the RNA Polymerase meets the termination sequence in the DNA. Gene expression is controlled at various levels (Lodish *et al*., 2004) like transcription, mRNA splicing, mRNA export from the nucleus, translation control and post-translational modifications. Regulation of gene expression at any of the above steps except for post-translational control, directly affects the abundance of gene products i.e. proteins. However for most genes, transcriptional control is the most important step for gene expression (Lodish *et al*., 2004).

*Toxicogenomics*

Toxicogenomics is a branch of science that deals with observing toxic responses using high throughput profiling technologies like transcriptomics, proteomics and metabolomics (http://en.wikipedia.org/wiki/Toxicogenomics). The study of toxicology involves different types of exposures like subacute, sublethal and subchronic to determine the impact of toxic substance on the organism. The toxicity end points for these experiments is assessed with life span, mortality, body weight, blood chemistry, and histological assessment of the different tissues like brain, liver, kidney (Quinn, Jr. *et al.*, 2007). The toxicological phenotypes are observed and the RNAs are extracted from the different tissues. The toxicological experiments are set up based on these observations and RNA expression profile can then be recorded with high throughput transcriptome profiling technologies for further investigations to understand genomic impacts. The mRNAs that are expressed represents the transcriptome for particular condition and are extracted from the nucleus and cDNA libraries are produced. The normalized cDNA library is sequenced using Roche/454 Life Sciences next generation parallel pyrosequencing for the Northern bobwhite and Japanese quail.

*Next Generation Sequencing*

ESTs are fragments of DNA coding regions that are generated by sequencing at one or both ends of a gene. The ESTs provide quick route to discover genes in absence of full genome sequences. With the advent of the next generation sequencing that provide sequencing depth and coverage (Margulies and *et al.*, 2005) and have been utilized to explore and study genome wide variation (Dalca and Brudno, 2010), identification of protein binding sites, quantitative analysis of transcriptome (RNA-Seq) (Pepke *et al.*,

2009), and assembly of genome and transcriptome (Papanicolaou *et al*., 2009). The non-model species transcriptomics project has grown phenomenally to address biological objectives of interest (Heng Li and Nils Homer, 2010; Papanicolaou *et al*., 2009). The low cost and less time required for next generation sequencing will help in providing a workable draft of these non-model species genomes (Meyer *et al*., 2009; Papanicolaou *et al*., 2009; Vera *et al*., 2008). The next generation sequencing like SOLiD, Roche/454, Illumina and Helicos generate data in order of giga-bytes (Metzker, 2010). To keep pace with these high throughput technologies, new alignment tools have been developed in past few years.

*Sequence Assembling*

The assembly of next generation sequencing for genomes and transcriptomes of these organisms is challenging (Heng Li and Nils Homer, 2010). Most of the transcriptomics studies for non-model organisms take sequence alignment as first step to generate contiguous sequences (contigs) that consist of overlapping reads by assembly (Papanicolaou *et al*., 2009). The sequence alignment is performed de novo or against reference genome (Heng Li and Nils Homer, 2010). Various de novo alignment methods use overlap-layout-consensus (OLC) strategy like CAP3 (Huang and Madan, 1999), graph based methods like NEWBLER (454 Life Sciences, Branford, CT), VELVET (Zerbino and Birney, 2008) algorithms are based on suffix trees (Heng Li and Nils Homer, 2010). For long reads alignment, BLAT (Kent, 2002), SASHA2 (Ning *et al*., 2001) have been used to align against reference genome. The alignment of short sequencing reads like Illumina (Illumina, Inc., San Diego, CA), SOLiD (Life Technologies Corporation, Carlsbad, CA) and Helicos (Helicos BioSciences Corporation, Cambridge, MA) to reference genome has

been applied to reads less than 200 bp that usually have less sequencing error (Heng Li and Nils Homer, 2010; Heng Li and R Durbin, 2010). The development of short read assemblers was quick and led to development of SOAP (Li *et al*., 2008b; Li *et al*., 2009b), MAQ (Li *et al*., 2008a), BOWTIE (Langmead *et al*., 2009). Long read aligners like BLAT and SASHA2 are relatively slower as compared to short read aligners (Heng Li and R Durbin, 2010). As the next generation technology is developing, the length of the reads are getting longer and long reads will dominate again (Heng Li and R Durbin, 2010). The Roche/454 ESTs are now considered as long reads. High number of ESTs and longer read length require high computational memory and management between the sensitivity and accuracy to assemble correctly. Parallel computing/multiprocessor is often utilized to access shared memory and partition problem that is sent to different CPU's for execution that results in high performance.

To keep up with the pace of increasing memory requirement for the long reads assembly with OLC, different tools/pipeline like TGICL (Pertea *et al*., 2003) , PAVE (Soderlund *et al*., 2009) are created to manage memory and perform clustering using megablast followed by assembling. Graph based methods based on K-mer are considered to be less memory intensive (Miller *et al*., 2010). Alternate methods like MIRA take different strategy for high and low confidences regions and take SNPs into account (Chevreux *et al*., 2004). The reference genome data is more readily available and therefore instead of denovo assembling approaches followed by conventional pipeline and tools, the researcher tried to assemble against the reference genome of close phylogenetic neighbour (Rawat *et al*., 2010a).

*CAPRG*

The section describes the development of "Contig Assembly Pipeline against Reference Genome" (CAPRG) that first map long reads against the reference genome followed by assembly (Rawat *et al*., 2010a). ESTs juxtaposed to one another spanning across the chromosome are binned in same window (Figure 3). For each window, sorted with anchor position on the chromosome, a group of ESTs are aligned and the contigs are generated at high percentage of identity.

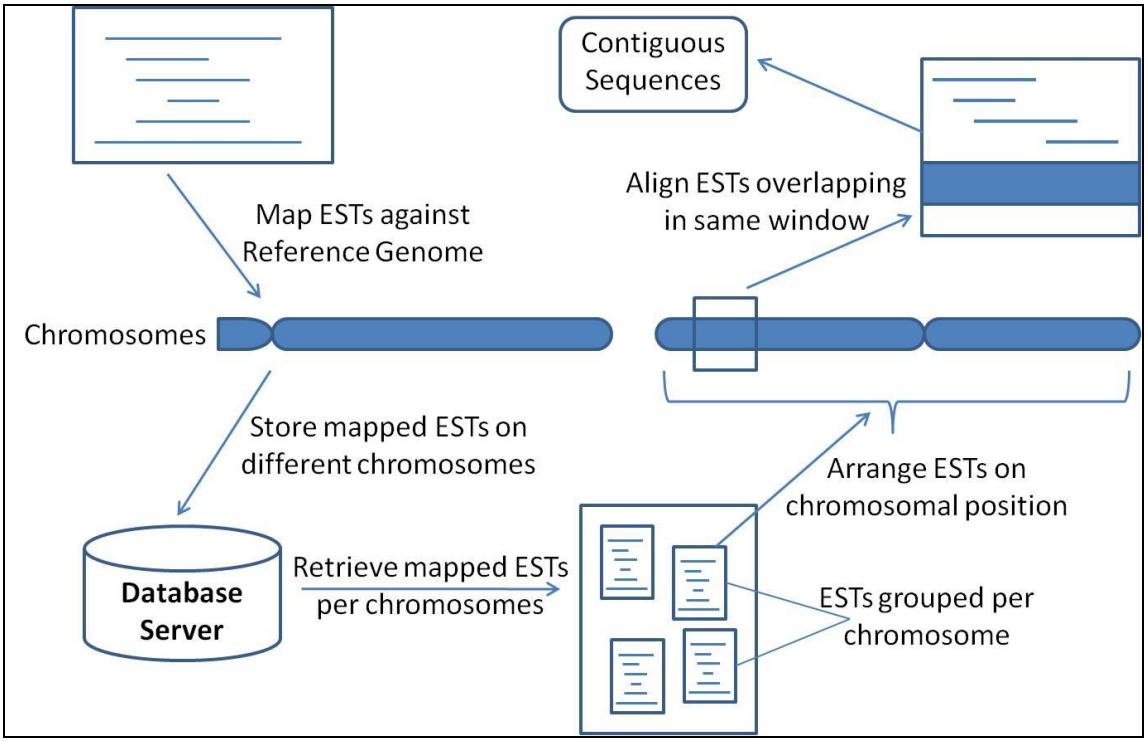


*Figure 3.* The flow chart of CAPRG representing the mapping of reads to generate Contigs.

The new implementation is tested against close phylogenetic neighbor using chicken reference genome. Gallus gallus and Japanese quail belong to same family Phasianidae while Northern bobwhite is more distant to chicken and belong to family Odontophoridae. It is found that CAPRG performed near equivalent or better in the two

transcriptomic datasets of Northern bobwhite and Japanese quail and finishes the assembly in fraction of time as compared to assemblers that yield competitive results (Rawat *et al*., 2010a). Other advantages are that it generates lesser number of superfluous overlaps and due to restrictive window based approach where limited number of ESTs seek overlap, have lesser chance of chimeric contigs. One of the limitation of mapping reads however to a reference genome is the availability of a close phylogenetic neighbor (>94% identity) (Dutilh *et al*., 2009).

*Microarray*

Microarray is a valuable tool that permits detection of small differences in transcript abundance. The DNA microarray can therefore measure the difference in transcriptional activity for every gene by comparing their mRNA levels when grown at different conditions (Ye *et al*., 2001). Each microarray has thousands or tens of thousands of genes represented on the chip. Any assay to be surveyed is comprised of one or more samples of mRNA expressed in any developmental stage, stress condition or specific tissue that is labelled with distinguishable marker. These are then hybridized to "spots" representing features/genes on the array (Causton *et al*., 2003).

The arrays are scanned by a laser scanner and images are produced and analyzed to obtain an intensity value for each probe. These intensities represent the extent of hybridization occurred for each oligonucleotide probe set (Causton *et al*., 2003) . The data therefore represents expression conditions with sensitivity. The oligonucleotide arrays are generally preferred over other kinds of microarray due to some of the advantages like no reverse transcription or amplification is involved resulting in fewer chances for contamination due to non specific amplification and mishandling, the mRNA is labelled

directly which is considered to represent the natural distribution of RNA species within the sample and there are lesser chances of cross hybridization which is considered to add noise (Ye *et al*., 2001).

The analysis of microarray experiments is challenging not only due to large size of data but also various types of variations that can be introduced at different stages of the experiments (Li and Wong, 2001). There is no golden rule for microarray analysis, and the application of various techniques might yield different results. The gene expression profiling with the microarray is widely used to understand the differences in expression between two groups (control and experiment) to identify differentially expressed genes (DEG's). Application of techniques suitable for microarray analysis begins with preprocessing and calculation of DEG's through various techniques like Bayesian method, neural network, support vector machine that utilize various feature selection methods like Pearson correlation and Euclidean distance which can be applied to seek genes with unknown function (Brazma and Vilo, 2001).

Variances help us to determine whether variance that arises in a sample is due to the random nature of the sampling process or is due to any variable of interest. Proper estimation of variances is an important aspect to calculate differentially expressed genes and replicates play an important role (Baldi and Long, 2001). The *t*-test does not detect differential expression effectively in the case of few replicates and leads to underestimation of variances, losing the sensitivity. To overcome these, Bayesian probabilistic framework for array data implemented as Regularized *t*-test (Baldi and Long, 2001), has been suggested. Regularized *t*-Test uses Bayesian probabilistic framework to calculate a background variance for each of the genes under analysis. Regularized *t*-test has been

found to perform better than conventional *t*-test when the number of replicates is low (2-3) (Baldi and Long, 2001).

Cross validation of DEG's generated from different statistical tests is another powerful way to consolidate DEGs list and overcome noise and experimental artifacts. Also false discovery methods like Receiver Operator Characteristic (ROC) curves are helpful to correct for multiple comparisons.

*Toxicological Phenotypes*

After identifying DEGs from microarray experiments, the toxicogenomic responses are utilized to discover MOA for MCs based on perturbation and develop new insights for metabolic pathways observed in phenotypic responses to these exposures. In a subacute toxicity study (Johnson *et al*., 2007), the responses to RDX was found more toxic than 2,6-DNT and both RDX and 2,6-DNT showed different observation with RDX accumulating in brain while 2,6 DNT causing gastrointestinal distress, dehydration and loss in body mass. To further study these impacts, a Northern bobwhite microarray for brain tissue was developed to study the neurotoxicogenomic effects due to both RDX and 2,6-DNT (Gust *et al*., 2009). As expected, RDX showed high number of differentially expressed genes while 2,6-DNT showed few genes were affected. For the common dose of RDX, the birds exhibiting RDX-induced seizures accumulated over 20x more RDX in brain tissues in comparison to non-seizing birds. Expression patterns among seizing and non seizing birds were different and genes that are involved in neuronal electrophysiology and signal transduction were found to be significantly affected in birds experiencing seizures.

In study for subchronic effects on adult Northern bobwhite for 2,6-DNT (Quinn, Jr. *et al*., 2007), these exposures affected blood chemistry by decrease in red blood cells and

hemoglobin, plasma concentration of protein, albumin, globulin, aspirate aminotransferase and potassium, sodium and chlorine ions however uric acid levels were increased significantly. The phenotypic observations were decrease in body weight, enlarged gallbladders, edematous gastrointestinal tracts, pale kidneys and fibrous liver and loose stools. Overall effect along with histopathological abnormalities observed suggested that the liver and kidneys are major targets for 2,6 DNT.

To evaluate sub-lethal effect on Northern bobwhite, the birds were exposed to doses of RDX (1,3,5-trinitro-1,3,5-triazine) (Quinn, Jr. *et al*., 2009) and dose dependent mortality rate was found for many dose groups. Other effects observed were weight loss due to gastrointestinal effects, death preceded by clonic and tonic convulsions and degeneration of testicular and splenic tissue.

The project focus to understand the sub-lethal exposures of 2,6-DNT with microarray analysis for Northern bobwhite, the 10 and 60 mg/kg/d doses of 2,6-DNT (Rawat *et al*., 2010c). The doses were selected to represent the minimum dose at which significant impacts on toxicological endpoints were observed and a highly affective dose that elicited numerous significant impacts on toxicological endpoints approaching the threshold of mortality (Quinn, Jr. *et al*., 2007) .

After DEGs are recognized for the microarray experiment, the expression can be evaluated by GO enrichment, pathways, metabolic maps and gene networks. The gene ontology (GO) describes gene/gene product in three categories biological processes, cellular components, and molecular functions as directed acyclic graph (DAG) in which a term may have more than one parent and more than one child (Ashburner *et al*., 2000; Harris *et al*., 2006a). GO enrichment are frequently used to analyze set of genes that might

have shared GO terms for large data such as microarray experiments (Boyle *et al*., 2004). The enrichment is performed with the help of algorithms that determine whether the observed annotation from microarray experiment is significant compared to background genome with appropriate *p*-value. Pathway information provides an integrated representation of biochemical reactions and functional interactions in comparison to the gene-centric GO analysis (Thomas *et al*., 2007; Werner, 2008). The researcher collected pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database using KEGG Orthology (KO) to annotate unique gene sequences (Kanehisa *et al*., 2006). KO is a classification of ortholog and paralog groups based on highly confident sequence similarity scores and is directed acyclic graph with hierarchy layout in four tiers (Mao *et al*., 2005). Gene Map Annotator and Pathway Profiler (GenMAPP) is a program that allows users to visualize microarray data in context to pathways stored as MAPPs and allow meaningful interpretation to expression changes (Dahlquist *et al*., 2002; Salomonis *et al*., 2007).

The system biology focuses on holism by studying the interactions and dynamics of organismal function, instead of reductionist approach focusing on isolated parts of a cell. Any biological system consist of genes and proteins however we are at juncture of biology that most of the genes and proteins for model organism are already known (Kitano, 2002). This creates opportunity to shift the paradigm to system level approach to develop understanding of interactions and dynamics of biological system. One of the key properties that can be utilized to understand to perform system level analysis is system structure that focuses on pathways and gene interactions network and the effect of these on intracellular and multicellular structures (Kitano, 2002). Large scale, comprehensive database storing

gene-regulatory and biochemical network is key to understand system structures. One such system, Ingenuity Pathways Knowledgebase (IPKB) stores known interactions among gene objects that can be superimposed on DEGs to build networks (www.ingenuity.com).

These tools and knowledgebase were utilized to interpret and find significant changes in several physiological pathways at 10 and 60 mg/kg/d that were both dose responsive and dose dependent corresponding with the observed toxicological phenotypes. The impacts in 2,6-DNT treatments were investigated to determine potential mechanisms underlying observed gross-level effects and effects on blood chemistry and, further, explored genomics-directed observations to provide a systemic understanding of the general pharmacology of 2,6-DNT in Northern bobwhite.

*Knowledgebase Design and Development*

Annotation provides functional context to the unknown sequences. The first step is to perform pair wise alignment against the Genbank database using heuristic programs like BLAST. With the high number of contigs generated from the next generation sequencing, homology detection is major bottleneck to perform annotation. The parallelization of sequence similarity can be done with Message Passing Interface (MPI) Blast using High Performance Computing (HPC) (Darling *et al*., 2003). The data is partitioned into "n" segments and querying is parallelized running in n nodes concurrently resulting in considerable speedup. As the time lag to detect homology is sufficiently reduced by utilizing parallel infrastructure, this allows comparing different assembling outputs and performing comparative genomics.

After the homology detection, associating the unigenes with various biological models is useful. Multiple sequence alignment can be performed with Interproscan that is

used to scan for protein signature (protein families, domains and functional sites) from various databases PROSITE, PRINTS, Pfam and ProDom (Quevillon *et al*., 2005; Zdobnov and Apweiler, 2001). Prediction of protein-coding regions was established using ESTScan which uses a hidden Markov model to identify coding regions, even if the quality is low and contains frame shifts, a common sequencing error (Iseli *et al*., 1999). Different approaches for annotation like gene ontology, pathway assignment helps to develop insight into the genome. Gene ortholog detection is extensively utilized to mine the increasing amount of sequence data generated by complete or partial genome projects which provides increasing accuracy in predicting ortholog and paralog relationships (Conte *et al*., 2008; Engelhardt *et al*., 2005). As the genomics data of model organisms is available in public resources like Genbank, the data can be used to perform comparative genomics. Recently chicken annotation has achieved model organism status (Cogburn *et al*., 2007), the comparative genomics across six model organism and chicken to understand similarities and differences in the proteins is performed. The availability of genomic data and phylogeny can be assessed with the comparative genomic analysis that helps to recognize chicken as closest phylogenetic neighbour that is used as reference/background organism for further studies.

So far, the genomic information for non model organism like Northern bobwhite lags behind model organism and other avian species (Rawat *et al*., 2010b). By comparison, avian species such as chicken and zebra finch have been robustly described and various tools have been developed to allow in-depth bioinformatics investigations. For example, various public repositories (www.uniprot.org, www.geneontology.org, www.genome.jp/kegg) host protein, gene ontology and pathway information for chicken in

addition to specialized chicken databases which are also freely accessible (http://geisha.arizona.edu, www.chick.manchester.ac.uk, http://mpss.udel.edu/gga). The researcher recognized the need to greatly expand the information-base for Northern bobwhite and develop the species as a robust avian-wildlife genomic model. The knowledgebase is initiated to share and develop functional genomic data, initially with Northern bobwhite (www.quailgenomics.info).

Data entities such as protein information and gene ontology annotation are integrated in a common platform with a web interface for comprehensive parameter searching. The linkage among these data entities is provided by unigene ID that connects the entities internally allowing the user to perform flexible query searches. The result of our effort is a web-accessible knowledgebase for Northern bobwhite which includes user friendly navigation tools and provides EST assembling information, sequence and structural properties and complex search utilities, bundled with an alternative method to generate sequence scaffolds to "stitch" transcripts against a reference organism. The data represented in the Quail Genomics knowledgebase have provided novel insights into the systemic perturbations of 2,6-DNT in Northern bobwhite (Rawat *et al*., 2010c). Similarly, the knowledgebase can provide researchers the ability to perform analysis and curate genomic information to further their own research pursuits.

*Genomic Comparisons between Avian Species*

The comparison among toxicological model species, the Northern bobwhite and Japanese quail and passerine bird, the zebra finch to the MEC compounds 1,3,5-trinitro-1,3,5-triazacyclohexane (RDX) and 4-amino-2,6-dinitrotoluene (4A-DNT) of avian responses will be conducted. The goal of this research is to assess the similarity in

responses of the avian species to RDX and 4A-DNT using clinical toxicological methods, analytical chemistry and genomic inquiry (K.A. Gust *et al*., manuscript in preparation).

The Northern bobwhite and Japanese quail were sequenced as there is limited availability of data however zebrafinch have available data present in dbEST (http://www.ncbi.nlm.nih.gov/dbEST). To make a more accurate comparison among these avian species, it is proposed that atleast 50% of the targets should be common among them (K.A. Gust *et al*., manuscript in preparation). The available zebrafinch data is processed and genomic comparison study performed among these avian species.

# CHAPTER II

# METHODS AND MATERIALS

## Preparation of cDNA Library

Tissue samples used to construct the normalized cDNA library for Northern bobwhite were collected in exposure studies conducted at the US Army Center for Health Promotion and Preventative Medicine in Edgewood, MD.

In a subacute toxicity study, each of the nine groups of Northern bobwhite with mixed sex were separately exposed for 14 days with corn oil (control), 50, 100, 190, 350 mg 2,6-DNT/kg body weight/d and 20, 80, 125, 180 mg RDX/kg/d (Johnson *et al*., 2007).

In study for subchronic effects on adult Northern bobwhite, both sex were exposed for 60 days to one of 0, 5, 10, 40, 60 mg/kg/d dose for 2,6-DNT (Quinn, Jr. *et al*., 2007). Based on hematological measures, it was found that the lowest-observed-adverse-effect level is 40 mg/kg/d based on hematological measures, and the no-observed-adverse-effect level is 10 mg/kg/d based on the absence of results that show adverse effects.

To evaluate sublethal effect on Northern bobwhite, the birds were exposed to 0, 0.5, 3, 8, 12, or 17 mg/kg of RDX (1,3,5-trinitro-1,3,5-triazine) in corn oil daily for 14 day (Quinn, Jr. *et al*., 2009). Dose dependent mortality rate was found for 17, 12 and 8 mg/kg/d dose groups. Other effects observed were weight loss due to gastrointestinal effects, death preceded by clonic and tonic convulsions and degeneration of testicular and splenic tissue. The 3.0 and 8.0 mg/kg/d were determined for no-observed-adverse-effects and lowest-observed-adverse-effects levels, respectively.

In total, the RNA pool used to construct the Northern bobwhite cDNA library represents 179 tissue samples taken from 56 individual birds. Tissue types represented

include brain, liver, testes, duodenum, colon and feather pulp. All protocols were conducted consistent with Good Laboratory Practices and approved by the Institutional Animal Care and Use Committee at the U.S. Army Center for Health Promotion and Preventative Medicine.

The cells consist of following components: DNA, protein, RNA, tRNA, rRNA(most abundant RNA) and mRNA(1-5% is transcriptional expression). The RNA extraction procedure starts with the RNA storage and stabilization. The gene expression profile can be altered due to RNA degradation in samples and stabilization of RNA is necessary. This is performed by storing the tissue samples with RNA Later. The lysis/homogenization step unfolds protein and denatures them and also distorts the tertiary structure of RNAses. The binding process involves binding negatively charged RNA/DNA to the silica membrane so that all the other contaminates are separated. This is followed by washing after centrifugation to remove the residual proteins and salt and digestion of DNA with the help of DNase enzyme. Finally elution helps in separating the RNA from the silica surface with the help of elution buffer (http://bitesizebio.com/2010/06/28/how-silica-spin-column-dna-and-rna-preps-work).

Immediately following euthanasia by $CO_2$ asphyxia, tissue samples were fixed in RNA Later™ (Ambion, Austin, TX) following manufacturers recommendations. RNA extraction was conducted using RNeasy Mini RNA extraction kits (Qiagen Inc., Valencia, CA). RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) with RNA 6000 Nano LabChips® RNA. Only samples with a 28s/18s ratio $\geq$1.4 and RNA integrity number (RIN) $\geq$6 were used for downstream applications. The majority of RNA samples greatly exceeded these minimum

requirements. The RNA compilation included 800ng of total RNA from each of the 179 RNA samples which was purified for poly(A) RNA using a NucleoTrap mRNA purification kit (Macherey-Nagel, Germany).

This RNA pool is used to construct normalized cDNA library for Northern bobwhite. The cDNA library generation for each dose and stressor for the tissue is done with the SMART™ PCR cDNA Synthesis Kit (Clonetech Laboratories Inc. Mountain View, CA) and utilized to reverse-transcribe 0.5 µg of the poly(A) RNA into full length cDNAs with these steps:

i. First strand cDNA synthesis
ii. double stranded (ds) cDNA synthesis
iii. ds cDNA polishing
iv. cDNA library construction

Most cDNA synthesis methods utilize reverse transcriptase (RT) to transcribe mRNA to single stranded (ss) cDNA. Many times the RT is unable to transcribe entire mRNA sequence especially for long mRNAs leading under representation of 5' ends. The SMART protocols help to transcribe full length cDNAs. When the RT reaches the 5' end of the mRNA, it adds deoxycytidine to the 3' end of cDNA. The SMART II has oligo(G) sequence at its 3' end and it hybridizes with deoxycytedine of RT. As a result, this extended template continues replication and the full length ss DNA that covers the entire length of RNA are generated.

0.5 µg of total RNA along with the 3' SMART CDS Primer II, SMART II A Oligonucleotide and Reverse transcriptase enzyme are used for this step. cDNA synthesis methods involves reverse transcriptase (RT) enzyme to transcribe mRNA into ss cDNA in

the first-strand reaction. The double stranded cDNA synthesis is done with master mix (PCR buffers, DNA polymerase enzyme, dNTP mix, 5' PCR primer) from ss cDNA. Double strand cDNA polishing step is used to inactivate DNA polymerase with the Proteinase K. Next, the T4 DNA polymerase is used to produce blunt ended cDNA that can be ligated to any adaptor. Finally the cDNA with adaptors can be now inserted into the plasmid to be cloned in E Coli (SMART PCR cDNA Synthesis Kit User Manual).

## cDNA Library Normalization

After the cDNA library has been constructed from all the tissues and conditions (dose, stressor), the entire library is normalized. The expressed genes in eukaryotic cells generate mRNA varying from few copies to 200,000. The random sequencing of EST from the cDNA library will have greater probability to select abundant transcripts to be sequenced and leading to inefficient detection of rare transcripts. Normalization is multi-step process performed using the Trimmer cDNA Normalization Kit prior to sequencing to increase the random sequencing efficiency and detection of rare genes (http://www.evrogen.com/kit-user-manuals/Trimmer.pdf).

The first step is hybridization that denatures cDNA at high temperature (98°C) and then allows to renature at lower temperature (68°C). Denaturation produces single stranded cDNA fragments of both the rare and abundant transcripts. As the temperature is lowered, the ss cDNA starts to reassociate. In a given amount of time, the abundant transcript anneal more rapidly and in greater number as compared to rare transcript. The second step is use of enzyme Duplex Specific Nuclease (DSN), which has a strong affinity for cleaving ds DNA as compared to ss DNA. This helps in removing most of the ds DNA that belong to abundant transcripts though some might be rare ds DNA. The cDNA library is normalized,

leaving behind majority of the ss DNA that belong to rare transcripts (some may be single stranded abundant cDNA).

The cDNA fragments are then amplified with the PCR and the normalization efficiency is checked on agarose gel by using marker gene of known abundance. Successful normalization is represented in form of a smear with no distinguishable bands in between showing that the abundant transcripts are normalized at par with the rare transcript.

The cDNA library was normalized prior to sequencing to capture both high and low abundance transcripts using the Trimmer cDNA Normalization Kit (Evrogen JSC, Moscow, Russia).

## Sequencing

The pyrosequencing workflow is based on principle of :

One Fragment = One Bead = One Read

The pyrosequencing begins with each fragment (ss DNA) that is attached with specially designed beads (http://454.com/products-solutions/how-it-works/index.asp). Each unique fragment attached to the bead is amplified millions time thus each bead now has million copies of fragment fixed with it. Each bead is deposited in different well and immobilized and enzymes like DNA polymerase, ATP sulfurylase, luciferase and apyrase, and the substrates adenosine 5´ phosphosulfate (APS) and luciferin fill the well (Figure 4).

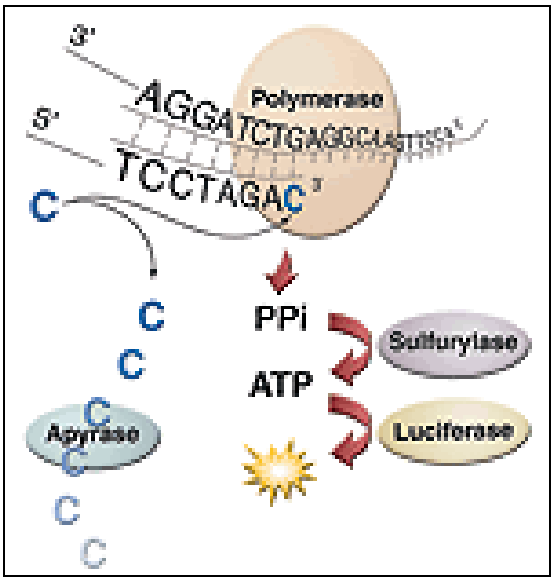


*Figure 4.* The flow of the pyrosequencing**.** Figure courtesy of http://www.har.mrc.ac.uk/services/GEMS/mapping.html

The deoxy-nucleotide triphosphates (dNTPs) are flowed one by one for each (of the four) A, T, G, C nucleotide over the wells. The complementary dNTP will be hybridized with the template by DNA polymerase. For each hybridized dNTP, equivalent amount of pyrophosphate (PPi) is released (http://en.wikipedia.org/wiki/Pyrosequencing). The PPi is converted to ATP in the sulfurylase mediated reaction with adenosine 5´phosphosulfate. The generated ATP is consumed by bioluminescent luciferase to produce visible light and this signature is captured by charge coupled device (CCD) camera. The signature of the light identifies the number of nucleotides (Margulies *et al*., 2005). Finally, the dNTP not utilized for hybridization needs to be degraded and this is done by apyrase enzyme. This cycle is repeated by adding new dNTP for further sequencing reads.

The normalized cDNA library was sequenced by 454 Life Sciences (20 Commercial St., Branford, CT) using massively parallel pyrosequencing on a GS-FLX

sequencer. The normalized cDNA library was sequenced by 454 Life Sciences using massively parallel pyrosequencing on a GS-FLX sequencer. Similar exposures were conducted on the Japanese quail to construct cDNA library that is sequenced at Berkeley University using massively parallel pyrosequencing on a GS-FLX sequencer. Approximately 478,142 expressed sequence tags (ESTs) comprising 114 megabases were sequenced for Northern bobwhite and 559,833 comprising 189 megabases for Japanese quail (Table 1).

Table 1

*The Sequencing Data Generated with 454 Life Sciences Parallel Pyrosequencing*

| | Northern bobwhite | Japanese quail |
|---|---|---|
| Passed Filter Wells | 478,142 | 559,833 |
| Total Bases | 114,877,461 | 189,239,672 |
| Average Length | 240.26 | 338.03 |
| Longest Read Length | 466 | 686 |
| Shortest Read Length | 28 | 11 |
| Median Read Length | 249 | 388 |

## EST Analysis

EST processing typically consists of two steps, sequence cleansing and assembly, the assembly (Figure 3). Sequence cleansing is removal of adapters, homo-polymer, vector contaminations and low quality (base-call) sequences. Many tools and scripts like STADEN (staden.sourceforge.net), CODONCODE (Dedham, MA), PHRED (www.phrap.com), SeqClean (http:// compbio.dfci.harvard.edu/tgi/software) are available

to perform sequence cleansing. These allow us to remove vector sequences, homo-polymer and adapters and screen the low quality sequences. The cleansing step is important part of EST pre-processing as any unwanted fragments can result in ambiguous overlaps. Also, the EST reads have low-quality regions generally at the either ends (Chou and Holmes, 2001). These low quality regions represent the low confidence call to the nucleotide base. The removal of these are generally done for base call with quality score QV<20. After removal of these regions, the high quality EST sequences can be used to build contiguous sequences.

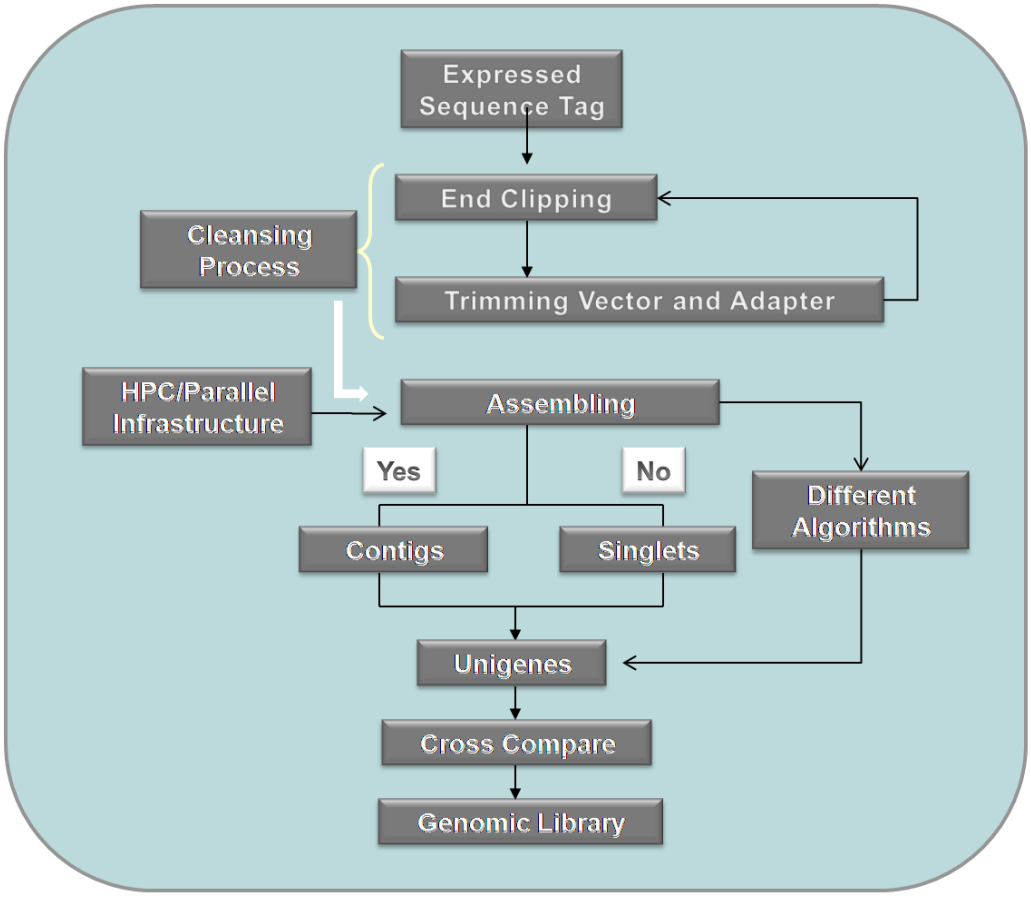


*Figure 5.* The flowchart for EST Analysis.

With the introduction of next generation sequencing, the assembly of high number of ESTs ranging from half million to one million requires high memory usage and computational infrastructure. Various programs using strategies like overlap-layout-consensus or graph theory are available. The overlap-layout-consensus like CAP3 (Huang and Madan, 1999), TGICL is memory intensive while graph based algorithm like VELVET are considered to be less memory intensive. The assembly output like total number of contiguous sequences, average length and number of homologous matches not only differs with different algorithm but also different parameters used for same program.

*Overlap-Layout-Consensus Assemblers*

The OLC assemblers form contigs in three stages. The overlap stage performs all-against-all, pair wise read comparison (Miller *et al*., 2010). The seed and extend heuristics is generally used for faster execution. The overlap between sequences is sensitive to the initial seed size, minimum overlap length and percent identity. Higher parameter values are more accurate but leads to shorter contigs. The layout stage constructs and manipulates to build approximate read layout. The consensus stage performs multiple sequence alignment (MSA) to construct precise layout and the consensus sequence. There are different known methods to perform multiple sequence alignment.

*Graph Algorithms for Assembly*

Graphs represent relationships with the set of nodes and edges that connects the nodes (Miller *et al*., 2010). A K-mer graph consists of fixed length sub-sequences from a larger sequence (Figure 6).

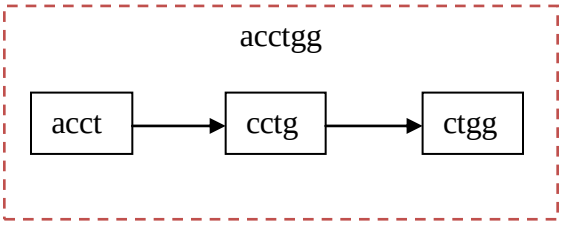


*Figure 6.* Directed graph representing K-mer with K=4 for sequence.

The nodes represent fixed length strings and the edges as suffix-to-prefix overlaps. Each read has its own path. The paths through the graph are considered as potential contigs. The reads with perfect overlaps result in common path and overlaps are detected without any pair-wise comparisons (Miller *et al*., 2010). The shared K-mers are easily detected than the overlaps and the computational cost of assembly is reduced. K-mer graphs are more sensitive to repeats and sequencing errors by handling the overrepresentation and underrepresentation of the K-mer population (Miller *et al*., 2010). The repeats will be overrepresented with higher number of K-mer while the underrepresented K-mer might be the random sequencing errors. The length of the K-mer size determines the accuracy of the contigs with higher K-mer size represent higher sensitivity (Zerbino and Birney, 2008) but will lead to shorter contigs.

*Indexing the Reference Genome*

The Smith-waterman algorithm solved for the local alignment problem between two sequences (Smith and Waterman, 1981). The FASTA (Pearson and Lipman, 1988) and BLAST family of alignment programs like NCBI BLAST (Altschul *et al*., 1995), MegaBLAST (Zhang *et al*., 2000) and WU-BLAST (Gish and States, 1993) gave fast alignments involving large sequence databases. The BLAST builds index of the query sequence and then scan across the database with "seed and extend strategy" for the high-scoring pairs (HSPs) (Altschul *et al*., 1995). However, with the introduction of high

throughput next generation sequencing, assembling and annotation of millions of ESTs pose challenging problem (Kent, 2002). With the availability of reference sequences, the BLAT, SSAHA (Ning *et al*., 2001) index the entire genome database and then scan through the query sequence. Recently, the Burrows Wheeler Aligner's Smith Waterman (BWA-SW) (Heng Li and R Durbin, 2010) uses FM-indices as compared to hash table based algorithms such as SSAHA and BLAT. The alignment is accelerated resulting in substantial saving the time (Heng Li and R Durbin, 2010).

*Assembling Pipeline*

To cope with increasing number of the ESTs for assembling of the next generation sequencing, many tools like TGICL and PAVE have been utilized that uses multi-processor for faster execution. The ESTs are clustered with the help of MegaBLAST by grouping sequences with high similarity and then assembling these groups with CAP3 (Pertea *et al*., 2003; Soderlund *et al*., 2009). This strategy reduce the number of pair wise sequence comparisons (Bragg and Stone, 2009) and results in better memory management and execution. The reads length is increasing and with the availability of genome data for close phylogenetic neighbour, CAPRG is developed that aligns the reads with the help of long read aligner BWA-SW. The database management MySQL is used to create cluster based on overlapping ESTs spanning across chromosome and ESTs are assembled with the CAP3 (Rawat *et al*., 2010a).

It is crucial to recognize the best dataset by cross comparing the assembly outputs to quality control as the further study like genome coverage, microarray probe design are dependent on this step. The researcher compared the statistics among assembly output and

establish the best result set among these assembling program by union, intersect or superior among different assembling dataset.

## Microarray Data Design and Analysis

The Agilent platform provide custom oligonucleotide microarray platform and the array design is an important step of microarray analysis. Printing a diverse set of putative transcripts provides a broader coverage of a genome and help to understand the systemic perturbation in more holistic manner. The flowchart represents the design of the microarray, preprocessing, statistical analysis leading to RT-qPCR validation with the help of different tools (Figure 7).

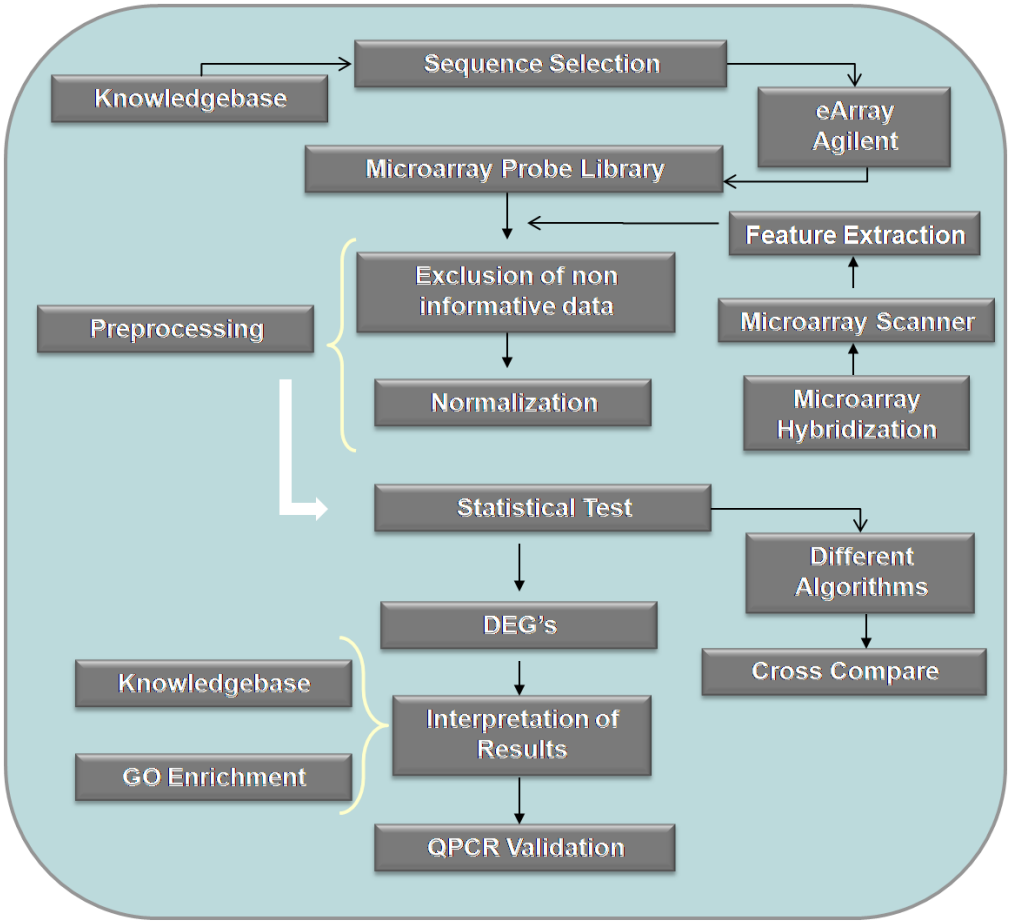


*Figure 7.* The flowchart for microarray design and analysis.

*Preprocessing*

In microarray analysis, the preprocessing involves the detection and elimination of non-informative probe sets from the dataset and normalization. These steps take place before the statistical validation of the differential expression. The microarray helps to understand the variation of gene expression for any response. These variations generally termed as interesting, which are biologically important, may be hidden by the superimposition of "obscuring variation" or systematic variation (Irizarry *et al*., 2003). These obscuring/technical variations might be induced during mRNA extraction, manufacture of array, hybridization and scanning labels. These variations in a DNA microarray experiment can affect the measurement of mRNA levels, making direct comparisons difficult and generating misleading results. Normalization is therefore used to identify and reduce the effects of this systematic/obscuring variation to enable direct comparisons between different chips (Causton *et al*., 2003). There are various approaches to perform normalization are generally categorized as "within" array and "between" array, and studies have shown that different normalization method effect the DEG's. In quantile normalization, there is no reference chip involved in normalization process and information from all arrays is used (Irizarry *et al*., 2003). Methods like scale and non linear normalization use reference chips for normalization, and other methods perform "within" (also called per-chip) normalization.

*Identification of DEGs*

The DEGs are identified with the *p*-value from statistical model like the *t*-test and fold change that indicates whether the gene is up-regulated or down-regulated. The *p*-value cutoff decision to identify differentially expressed genes is arbitrary as there is always a

tradeoff between true positives and false positives. The cutoff is empirical and there is no standard rule for cutoff selection however the FDR like ROC can help in *p*-value cutoff decision.

*t-test*

The means $m_c$ and $m_t$ and variances $s_c^2$ and $s_t^2$ are used to compute a normalized distance between two populations by calculation:

$$t = m_c - m_t / \sqrt{(\frac{S_c^2}{n_c} + \frac{S_t^2}{n_t})}$$

Two samples are considered to be different, when *t* exceeds a certain threshold depending on the confidence level selected.

*Regularized t-test*

A Bayesian probabilistic framework-based *t*-test calculates differences in gene expression by combining the empirical variance with the local background variance associated with neighbouring genes and calculates the confidence associated with the differential expression (Baldi and Long, 2001). This is supposed to compensate for the experimental noise with the limited number of replicates. Those genes which have near similar expression can be considered neighbouring gene and the window size is user defined (100 or so) called the sliding window. The variance within any given treatment is estimated by the prior variance for that gene (obtained from local weighted average of the variance of other neighbouring genes) called background variance and the empirical estimate of the variance for that gene (Baldi and Long, 2001). Bayes confidence estimate value is the confidence on the weight given to Bayesian prior estimate.

*ROC Plots*

The Receiver Operating Characteristics (ROC) plots represent the true positive and false positive to describe sensitivity of an experiment as shown in Figure 8.

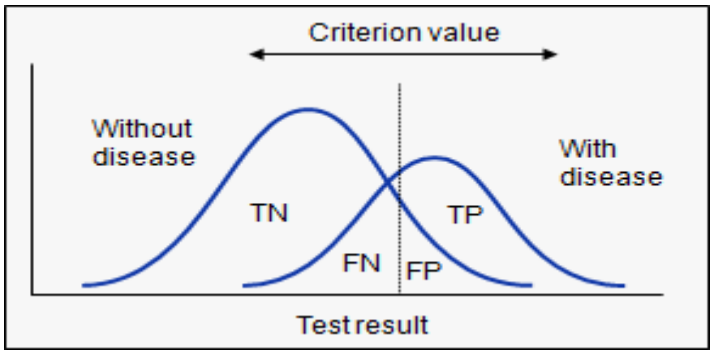


*Figure 8.* The classification of two populations.

The prediction of the ROC plots is decided by the diagonal that divides the ROC space (Figure 9). For a point on the line, it means there are X % False Positives and Y % True Positives, given the p-value threshold below which genes can be considered to be significant.

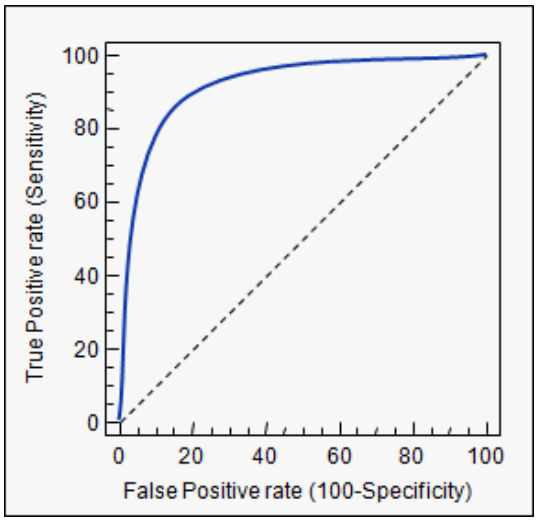


*Figure 9.* The ROC space and the prediction outcome **.**

All the points that are represented above the diagonal are considered good classification with higher sensitivity (less false negatives) and higher specificity (less false positives). The points below the diagonal are considered as poor prediction (http://en.wikipedia.org/wiki/Receiver_operating_characteristic).

## Life Cycle of Development for the Northern Bobwhite

In this section, the methods for life cycle of sequence assembly, development and functional annotation and microarray design and analysis for Northern Bobwhite are described.

### *Sequence Assembly*

Sequences were trimmed for flanking SMART Adapters using Codon Code Aligner with the sequences shorter than 50 bp discarded. Poly (A/T) tails were removed from raw EST sequences with Codon Code Aligner resulting in 467,708 high-quality sequences out of a total 478,142 reads. Sequence assembly was performed using CAP3 with the minimum percent identity (the minimum percentage of identical bases in the aligned region) ≥ 80% and minimum alignment score (overlap match plus counting mismatches) ≥ 30 bps. Contiguous sequences (contigs) and singlets less than 200 bp were removed. The resulting 71,384 sequences (35,904 contigs and 35,480 singletons) were considered putative unique genes.

### *Sequence Analysis, ORF Prediction, and Domain Analysis*

The homologs for the unique genes were interrogated with non-redundant database from NCBI with Message Passing Interface (MPI) Blast (Darling *et al*., 2003) using high performance computing (http://cluster.vislab.usm.edu) on a 44 node cluster. Comparative genome analyses were performed using Refseq id for protein, downloaded in February

2008 from http://www.ncbi.nlm.nih.gov. The open reading frame (ORF) and frame direction of unique genes are derived with the help of ESTScan (Iseli *et al*., 1999; Nagaraj *et al*., 2007). Ortholog detection was performed with an in-house perl script to generate best Blast E value and bit score pairs for chicken sequences (downloaded from NCBI, February 2008) and Northern bobwhite unique transcript dataset and further mapping between these two datasets performed in the database package MySQL (www.mysql.com). Blastx orientations were preferentially used over frame directions from ESTScan in determining correct orientations for microarray probe design. For Interproscan domain analysis, the frame direction (and translated sequence) from ESTScan was used. Interproscan was used to search PROSITE, PRINTS, Pfam, ProDom, SMART and Panther protein signatures for unique genes with significant blast hits (Quevillon *et al*., 2005).

*Gene Ontology and KEGG Pathway Annotation*

For the functional annotation of the Northern bobwhite orthologs, the GO annotation against *Gallus gallus* annotation (downloaded from GOA (http://www.ebi.ac.uk) Chicken 39.0, released December 15th 2008) is performed using Web Gene Ontology Annotation Plot (WEGO) (Ye *et al*., 2006). The annotation derived was used to define second level GO function categories within each primary GO level (molecular function, cellular component and biological process). The chicken and quail dataset was compared to find statistically significant relationship by Pearson Chi-Square test performed by WEGO on 2x2 matrixes. KEGG Pathway analysis was conducted using KEGG Orthology-Based Annotation System (KOBAS) web server (Wu *et al*., 2006) investigating raw sequences that had a significant blast hit. Sequences with significant KO

were compared used for pathway analysis against the *Gallus gallus* as a reference organism.

*Northern Bobwhite Oligonucleotide Microarray Design*

A 15,000-probe microarray is created using a 8x15K custom oligonucleotide microarray platform (Agilent Technologies, Santa Clara, CA). The array was developed from a database consisting of sequences identifying putative transcript identities from blastx matches ($E \leq 10^{-5}$) to Northern bobwhite as well as from matches to closely-related avian species including *Gallus gallus*, *Numida meleagris*, *Coturnix japonica*, *Meleagris gallopavo*, *Callipepla gambelii* and *Gallus varius* to prioritize unique transcripts for inclusion on the array. The reads for next generation sequencing originate from random locations within each cDNA and have orientations in both the positive and negative frame. The Agilent microarray platform required 5' to 3' orientation (i.e. positive frame), therefore the sequences are to be segregated based on frame direction. A total of 8,454 non-redundant sequences with positive-frame orientations were identified and incorporated into the microarray design as described in the Supplemental text. For the remaining positions available on the microarray, a total of 3,272 unique Refseq IDs from the remaining homologous sequences (non-avian species) in order of lowest E-value where $E \leq 10^{-5}$ were incorporated. Complementary sequences were created for each due to lack of confidence in frame direction yielding the 6,546 probe sequences needed to complete the microarray probe set (See below for anti-sense strand exclusion methods). The 15,000 target sequences were uploaded to eArray (Agilent Technologies) where 60mer oligonucleotide probes were developed to represent each putative transcript on the microarray.

*Microarray Experimental Design, Hybridizations, and Data Extraction*

Based on experimental design, a one-color microarray hybridization experiment was conducted investigating the effects of subchronic (60d) exposures through oral gavage to the MC, 2,6-DNT on gene expression among liver and feather pulp (a potential non-invasive-marker) tissues in Northern bobwhite. Male birds were selected for the investigation due to higher sensitivity to 2,6-DNT exposure relative to females. A completely randomized design was utilized incorporating a 3 (Treatment) x 2 (Dose) factorial treatment arrangement to test for differences in gene expression among exposures (control, 10 and 60 mg/kd/d, 2,6-DNT) and among tissue types (liver and feather pulp). Data were extracted from microarray images using Agilent Feature Extraction software (Agilent Technologies). Each exposure type included 4 biological replicates from which both liver and feather-pulp tissues were derived (Table 2).

Table 2

*The Experimental Design of Liver Versus Feather for Northern Bobwhite Microarray for Control and Doses 10,60mg/kd/d for 2,6-DNT*

| Treatment | Liver Sample Id | Feather Sample ID |
|---|---|---|
| 0 mg | L-110 | F-110 |
| 0 mg | L-138 | F-138 |
| 0 mg | L-145 | F-145 |
| 0 mg | L-169 | F-169 |
| 10 mg | L-121 | F-121 |
| 10 mg | L-126 | F-126 |
| 10 mg | L-141 | F-141 |
| 10 mg | L-156 | F-156 |
| 60 mg | L-112 | F-112 |
| 60 mg | L-116 | F-116 |
| 60 mg | L-142 | F-142 |
| 60 mg | L-149 | F-149 |

*Microarray Analysis*

Background subtracted adjusted median signal intensities were normalized on a per-chip basis with R (http://cran.r-project.org/) that transforms the signal intensity by dividing signal intensity for all the genes with the mean intensity in each array. The normalized data was imported into HDArray using Bioconductor (http://www.bioconductor.org/) to measure the confidence value associated with fold

change for each gene ($p < 0.01$, unless stated otherwise). Results of Bayesian analysis were compared against a parametric, non-paired *t*-test. The *t*-test was conducted using GeneSpring™ version 7.2 (Agilent Technologies) where data were first normalized with per-chip median scaling and cross-gene error model. The t-test ($p < 0.05$) assumed non-equal variance and incorporated a >1.8 fold-change requirement investigating the 9,711 probes that had present flags for all four replicate samples for all conditions.

The ROC plots were constructed for these four experiments. The first two plots (Figure 10A and 10B) representing F10 and F60 indicate that there are no true significantly differentially expressed genes, because extremely high *p*-value cutoff is required. The plots (Figure 10C and 10D) for L10 and L60 show that the results have higher sensitivity with reliable prediction at low *p*-value.

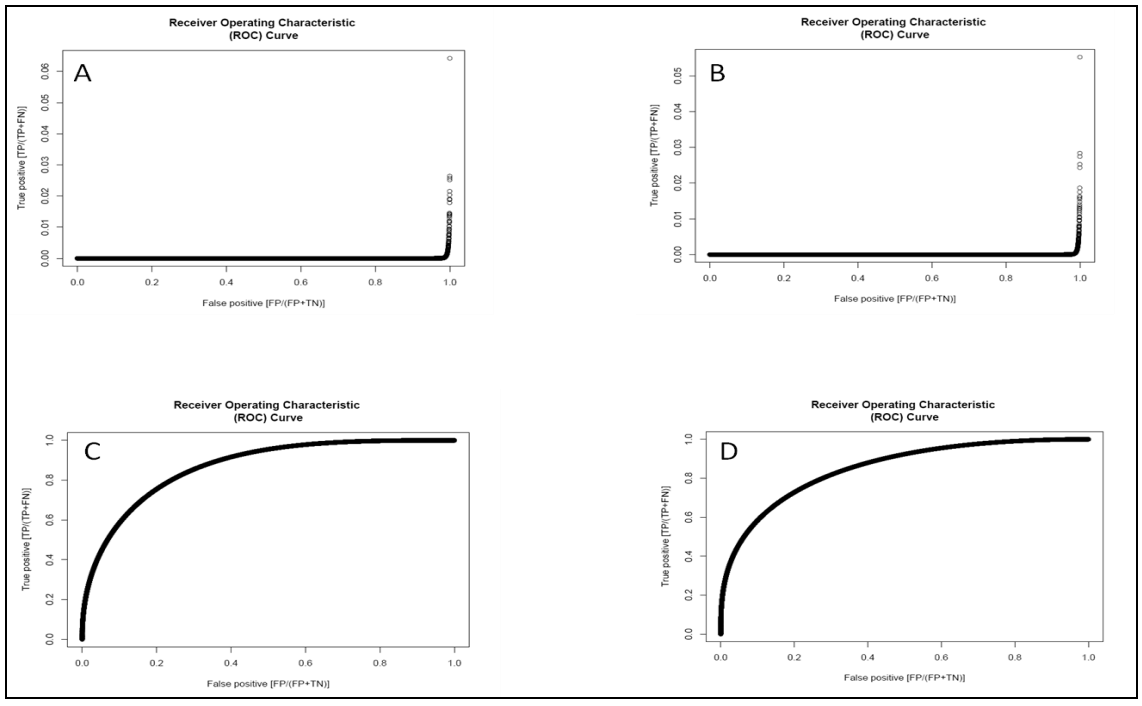


*Figure 10.* The ROC plots for the Liver and Feather at 10/60mg/kg/d for 2,6-DNT.

*Additional Quality Control for Microarray Data Analysis*

The 3,272 sense-antisense probe-pairs printed on the microarray were utilized to QC the microarray analysis for cross hybridization. One probe in each probe pair represents a nonsense sequence and our expectation was that no target should have specific binding to it. When differentially expressed genes (DEGs) were examined, 4 and 19 probe pairs were co-expressed in liver (10mg and 60mg, respectively). These non-specific DEGs were removed from our DEG list. The number of non-specific DEGs were limited relative to the overall number of properly-functioning probe-pairs which provided confidence in the representation of unique probes on the microarray, hybridization quality and microarray analysis.

*Reverse-Transcription, Quantitative Polymerase Chain Reaction (RT-qPCR)*

RT-qPCR allow us to amplify and quantify DNA at the same time, the amplified sample is measured at the end of each PCR cycle. The data thus generated can be used to calculate and compare gene expression in various tissue samples and also to determine the presence and abundance of a particular gene in the sample.

Transcript-expression levels were examined using DNase (Qaigen, Valencia, CA) treated total RNA reverse transcribed using Random Primers (Invitrogen, Carlsbad, CA) in a SuperScript III (Invitrogen) catalyzed reaction. Applied Biosystems (ABI) Prism 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA) was used to run SYBR® Green-based reactions. SYBR Green used here is a fluorescent dye that intercalates with double stranded DNA. Thermo-cycling began with a 10 minute, 95°C hot start followed by 45 cycles of 95°C for 15 seconds and 59°C for one minute. As the DNA

starts to accumulate after each cycle the fluorescent intensity increases and can be measured to quantify DNA concentration.

*Public Availability of Gene Expression and Toxicology Data*

Microarray and toxicological data are uploaded at the National Institute of Environmental Health Sciences (NIEHS), Chemical Effects in Biological Systems (CEBS) database (http://cebs.niehs.nih.gov) under the investigation title "Bobwhite Quail 2,6-DNT," CEBS accession number: 011-00001-0001-000-4. Microarray data and description is submitted with GEO archive: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18940

## Knowledgebase Design and Development

The Quail Genomics knowledgebase (www.quailgenomics.info) is implemented as a web-based tool with PERL 5.10.0, CGI, PHP 5.3, and BioPERL 1.6 script programs developed in-house interfacing with MySQL 5.4.3 database through PERL-DBI and integrate with class packages and modules of Go-Dev project (Figure 11). The user interface is supported in HTML that is hosted on Apache 2.2.13 webserver (UNIX version).

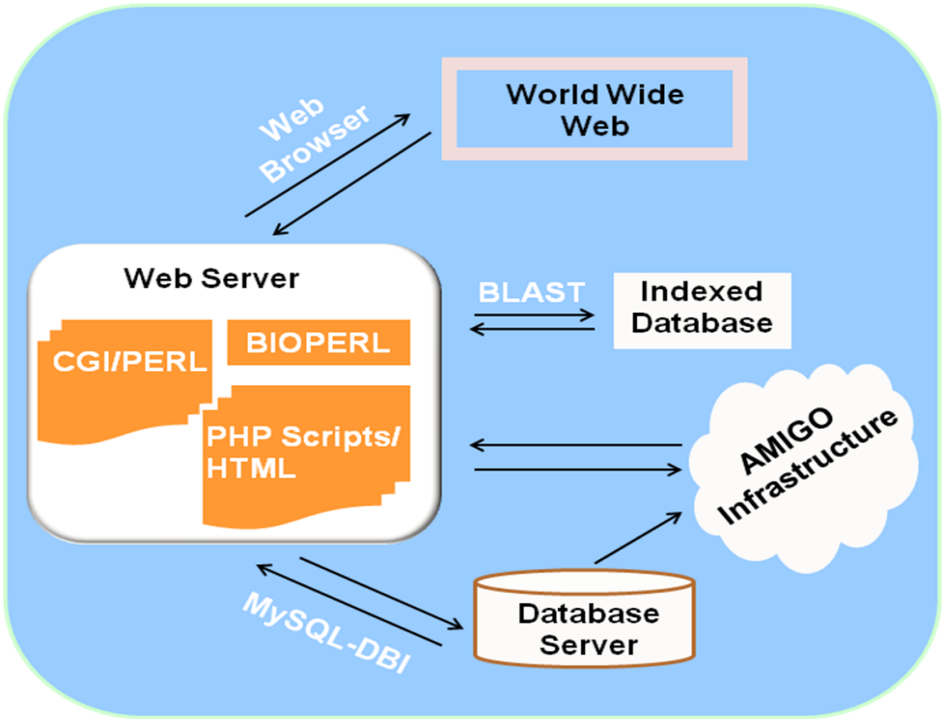


*Figure 11*. The web architecture of Quail Genomics.

The Quail Genomics knowledgebase (Rawat *et al*., 2010b) currently runs on a duo 2.26GHz Quad core Intel Xeon that uses the 64 bit Snow Leopard v1.6 operating system. The assembled sequences from Northern bobwhite and the annotation information are stored in the database. Users can access various features and data by PHP, PERL/CGI-BIN scripts hosted in Apache and retrieve information from database and/or indexed text files to display results (Figure 12). Hyperlinks are provided on the display results for retrieval of additional information.

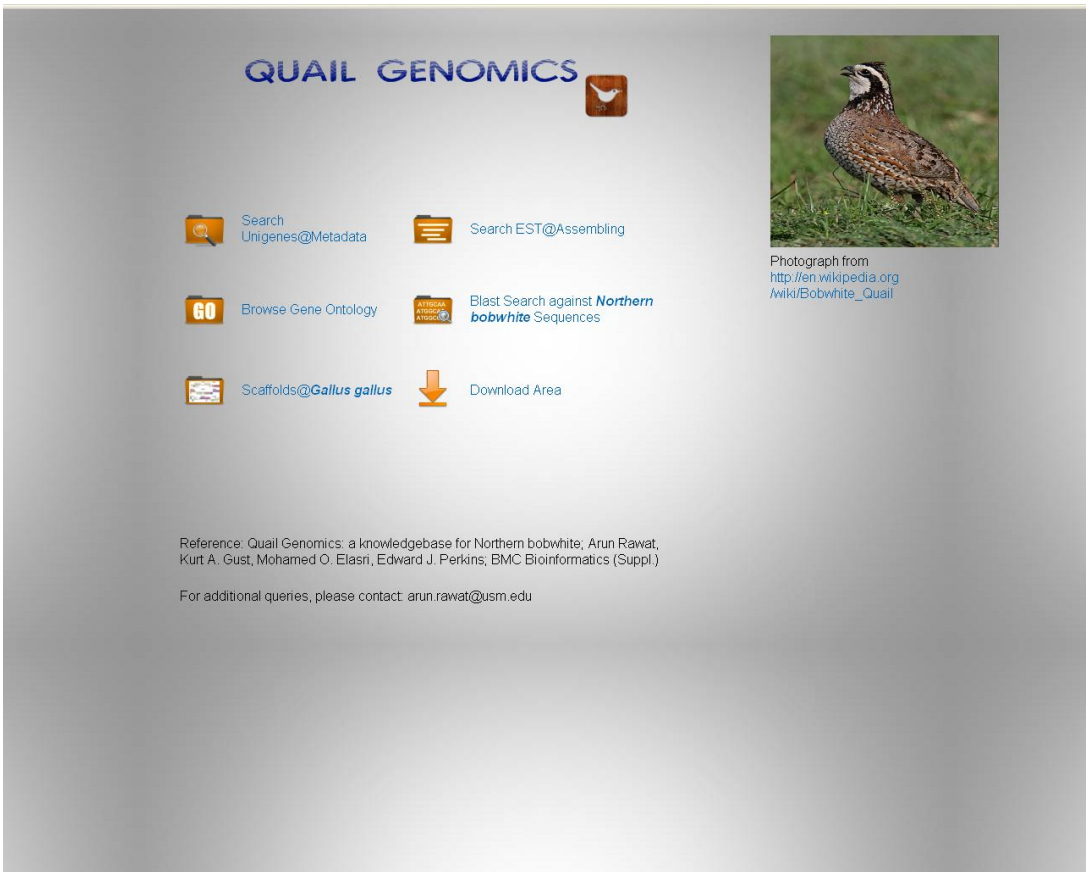


*Figure 12.* The web interface of the Quail genomics.

*Database Schema and Implementation*

The conceptual data representation of annotation and association of data entities is summarized in Figure 13. The database schema design and development was performed based on this interaction among the data entities and is stored in the relational database.

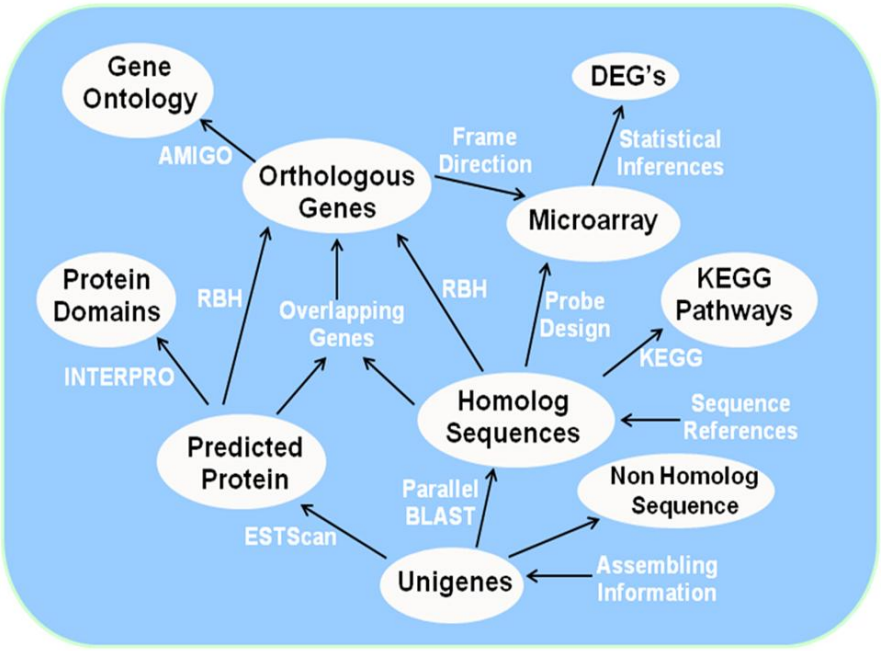


*Figure 13.* The data-flow diagram and interaction among the data entities stored in database.

The protein database represents Refseq sequences for chicken downloaded from Entrez in fasta format (http://www.ncbi.nlm.nih.gov/sites/gquery). The assembled sequences and corresponding chicken protein sequences are indexed for fast querying and stored as flat files.

*Nucleotide Assembly and Annotation*

Quail genomics at present hosts the information as described in Table 3.

Table 3

*Composition of Data Available in Quail Genomics Knowledgebase*

| ***Data Content in Quail Genomics*** | |
|---|---|
| Sequencing | 478,142 EST |
| Post assembly | 71,384 Unigenes (35,904 contigs, 35,480 singlets) |
| Putative homologs | 21,912 hits |
| Predicted protein regions | 39,400 potential ORFs |
| Protein domains | 15,057 Interproscan hits |
| Ortholog detection | 8,825 putative orthologs |
| Gene Ontology | 4,786 GO terms |

The relationship among these data models requires data conversion and formatting with the help of PERL, BIOPERL parse/extract/format scripts and programs so that the output of one data model be the input of another entity. The output is stored in the database and can be retrieved via querying for visual inspection of protein domains in HTML format. Hyperlinks are provided on the display results for retrieval of additional information. Microarray-based gene expression data is also included in the Quail Genomics knowledgebase where *p*-value and fold changes are stored as persistent data for each experiment.

*Gene Ontology Browser*

To functionally annotate the 8,825 Northern bobwhite orthologs, the researcher investigated and inferred putative GO (Ashburner *et al*., 2000) annotations finding

matching annotations for 4,786 (54.2%) genes. The GO-Dev project from Amigo is implemented locally to provide Tree Browser to represent the Northern bobwhite orthologs. GO-Dev is an open-source platform that consists of CGI/perl modules, database structure and web interface (http://wiki.geneontology.org/index.php/AmiGO_Manual:_Installation). GO-Dev and dependencies including GraphViz (http://www.graphviz.org/) which are required to successfully run the Amigo infrastructure locally were downloaded and installed. The Amigo infrastructure runs on the Quail Genomics webserver and interacts with our local MySQL database where database dumps are imported into the GO table structures. The GO data for Northern bobwhite orthologs is exported to the local Amigo database and the data can be viewed via tree browser from Quail Genomics.

*Genomic Scaffolds*

The 71,384 unigenes identified for Northern bobwhite are an over representation of the total protein-coding genes that are expected to make up the Northern bobwhite genome. Frequently, due to missing EST sequences, the ESTs from a single gene may not overlap to assemble to a contiguous sequence resulting in non-overlapping contigs and singletons or splits in genes. Also, stringent parameters during assembly of ESTs into contigs might lead to unassembled sequences especially when the sequences have low genome coverage (Potter *et al.*, 2004). These issues of missing sequences or fragmentation might lead to partial representation of a protein-coding sequence. Many of the sequence fragments might actually represent the same protein leading to redundancy in the assembled sequences.

For model organisms, methods such as Blat (Kent, 2002) and Bowtie (Langmead *et al.*, 2009) can be utilized to annotate against a reference genome; however inadequate

representation of a reference sequence for a non-model organism makes annotation difficult. Other methods like Ensemble gene-build pipeline (Potter *et al*., 2004) are also available however these do not allow user to select "gene of interest" and allow visualization. Finally, methods such as Genescript (Hudek *et al*., 2003) do provide visualization features, however integration of these with our web interface and database is not practical due to workflow and operating system incompatibility.

Therefore, a pipeline is built that generates scaffolds by aligning unigenes that might represent partial sequence fragments against specific coding regions of gene to generate scaffolds consisting of multiple-unigenes (Figure 14). The user can select from the list of all the Northern bobwhite genes stored in the database that have more than six unigenes/fragments (arbitrarily set) that represent same protein-coding region. The scaffolds can be built by clicking 'gene of interest' which fetches these similar unigene fragments from the database. The built in bioperl parsers extract the BLAST information and hit details and these are utilized to align the unigenes against the proteome database of chicken and visualized in a web browser.

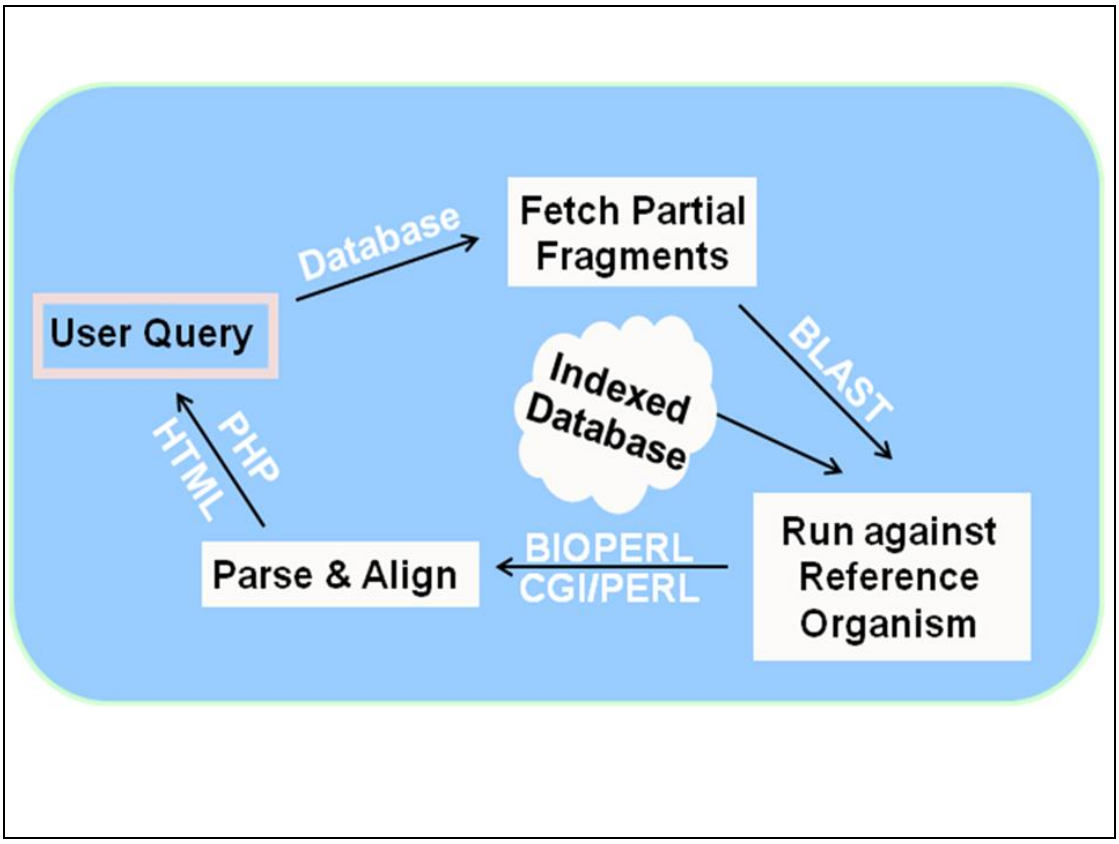


*Figure 14.* The flow chart representing the building of genomic scaffolds.

As described above, the protein database consists of Refseq sequences for chicken and stored in our local server. The output of BLAST includes information including: start position, end position, frame direction and sequence homology against the reference sequence for each unigene which is parsed and extracted with BioPerl script. This information is used to align each unigene that might represent a partial fragment against the chicken protein sequence. The temporary file handler added in the script allows multi-user access and deletes files created during scaffold building, to maintain housekeeping on the server.

Sequence Alignment Strategies and Development of CAPRG

Within couple of years, there have been lot of development in the next generation sequencing. After introduction of Roche/454, many short read sequencing like SOLiD,

Illumina and Helicos were developed. However, as the read length is increasing, the assembling paradigm will again shift to long read assembling (Heng Li and Nils Homer, 2010). To keep pace with these high throughput technologies, new alignment tools have been developed in past few years.

The cross comparison study is undertaken to utilize different strategies for assembling and assembly method are broadly classified as follows, PAVE, VELVET, MIRA using de novo strategies like OLC and graphs and CAPRG that uses reference genome to build initial clusters. To standardize the comparison study, same dataset for Northern bobwhite and Japanese quail after cleansing is used for assembling with different algorithms. The read preprocessing is done by masking adaptors, base calling, and removal of unwanted sequences like mitochondrial, rDNA and contaminants and homopolymer.

The output of graph based method like Velvet is highly dependent on K-mer size (Zerbino and Birney, 2008) while the OLC assemblers like CAP3 are affected by identity percent (Miller *et al*., 2010). The parameter space for the assembling output with different methods is taken into consideration and compared the output of these different assemblers against CAPRG.

The assembling for the four assemblers is performed on duo 2.26GHz Quad core Intel Xeon (Intel Corporation, Santa Clara, CA) that uses the 64 bit Snow Leopard v1.6 (Apple Computer Inc. Cupertino, CA) operating system with 16GB RAM. The multi-processor option is available for PAVE and MIRA and 8 processors are used for assembling. So far the CAPRG is single threaded application development of multi-threaded application will result in considerable speedup.

The CAPRG pipeline is implemented with PERL 5.10.0, and BioPERL 1.6 script programs interfacing with MySQL 5.4.3 database (www.mysql.com) through PERL-DBI. Other dependencies includes BWA-SW, SAMTools (Li *et al*., 2009a) and CAP3. The chicken reference genome build May'2006 downloaded from Golden Path (http://genome.ucsc.edu/) is used.

Singlets were excluded and only contigs with length greater than 200 are used for comparison study between different methods. These contigs were annotated against non-redundant protein database with using Parallel Blast with HPC (cluster.vislab.usm.edu) and BLAST programs against chicken database.

Genomic Comparisons between Avian Species

The comparison among toxicological model species, the Northern bobwhite and Japanese quail and passerine bird, the zebra finch to the MEC compounds 1,3,5-trinitro-1,3,5-triazacyclohexane (RDX) and 4-amino-2,6-dinitrotoluene (4A-DNT) of avian responses will be conducted.

Due to unavailability of Northern bobwhite and Japanese quail data, these organisms were sequenced. However, the zebrafinch have been sequenced and data is available in dbEST. However, before generating microarray for zebrafinch, two factors to evaluate the suitability of the available zebrafinch data for cross-comparison are considered. The first is the quality of the available data and second, coverage with the Northern bobwhite and Japanese quail data. Primary databases have redundancy due to separate submission by different groups for same gene and replication (each gene sequence in the database represented in separate form i.e. mrna, est) in the data. Also these databases have too many sequences for microarray probes and therefore we want to limit these

according to our studies. Secondary resources like Refseq can be ideal place to construct microarray libraries (Stekel, 2003). However, these resources have limited information for the non model sequences like zebrafinch.

Unigene are built by clustering available mRNA and EST sequences of an organism in the GENBANK and assigned to a cluster for overlapping sequences above a threshold of identity. Sufficient redundancy can still be found in unigenes clusters as and therefore many clusters might represent the same gene. TIGR gene indices have slightly higher advantage as it contains consensus sequences (contigs) unlike unigenes and therefore these consensus sequences are of higher quality and better for oligonucleotide design (Stekel, 2003). The zebrafinch assembly data is downloaded on Aug-2009 with UniGene (http://www.ncbi.nlm.nih.gov/unigene/) built # 12 and the total number of clusters was 14,432. The TIGR index (http://compbio.dfci.harvard.edu/tgi/) Release 1.0 is used for zebrafinch with total number of tentative sequences to be around 14,384.

# CHAPTER III

# RESULTS AND DISCUSSION

This Section discusses the toxicogenomics analysis for Northern bobwhite lifecycle development from sequence assembly of NGS, development and functional annotation, microarray design and analysis. The Quail Genomics knowledgebase is established to share and curate this species along-with tools like scaffold building that maps unigenes against reference proteome. As most of the analysis for Northern bobwhite was performed in 2007, with new sequencing data for Japanese quail in 2010, the researcher revised my strategies for sequence assembly. Using contemporary methods and data, the researcher introduces CAPRG that perform better or near equivalent in the two transcriptomic datasets based on different benchmarks. In the last section, the Japanese quail and songbird genomic comparisons is discussed and microarray design that would be used to interpret and assess the toxicological impacts across these avian species.

## The Lifecycle of Toxicogenomics Analysis for Northern Bobwhite

### *cDNA Sequencing, Sequence Assembly, and EST Processing for Northern Bobwhite*

Sequencing of the normalized cDNA library yielded 114.9 megabases of sequence in 478,142 reads with an average read length of 240 bp and approximately 95% of the total sequence read lengths distributed between 200 and 300 bp. A total of 82,064 unique sequences were identified comprised of 37,685 contigs and 44,379 singlets after vector cleaning, adapter and polyA tract removal and CAP3 sequence assembly. Nearly 64.4% and 19.8% of the contigs were comprised of greater than two and nine ESTs respectively (Figure 15).

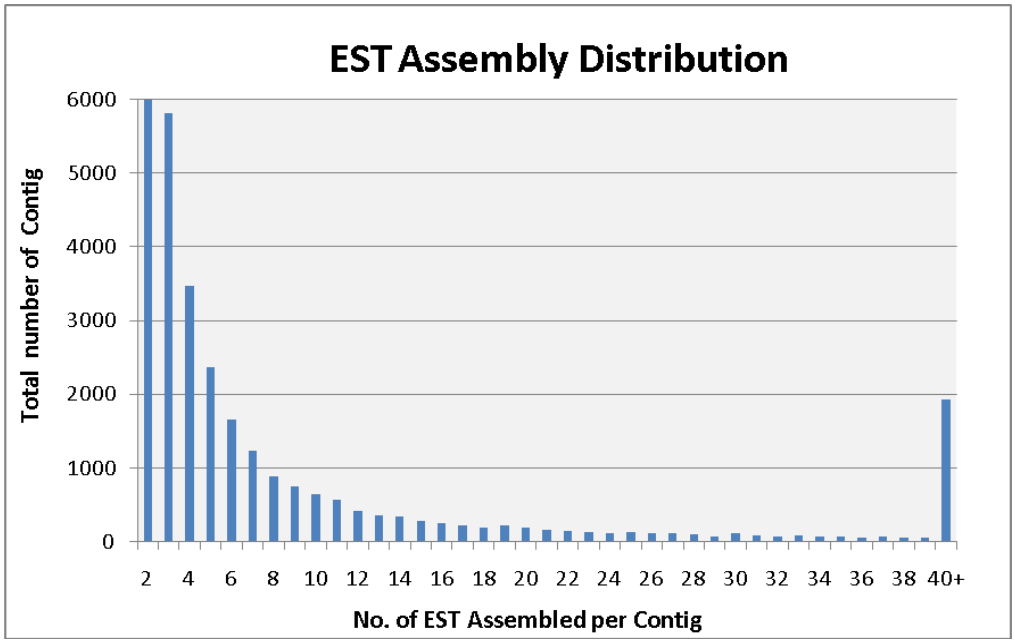


*Figure 15.* The EST distribution representing number of EST assembled per contig.

Contigs and singlets having length greater than 200 bp, totaling 71,384 unigenes, were selected for further analysis and annotation. The average length of the assembled contigs was 500bp (min = 202 bp, max = 6524 bp, and stdev = 403). The number of transcripts having significant matches to known or postulated protein sequences, where Blast expectation scores (E) were $\leq 10^{-5}$, was 39.2% of contigs (14,081 out of 35,904) and 22% (7,831 out of 35,480) of singlets (Table 4).

Table 4

*Overview of Blastx Matches Derived from Contig Assemblies Conducted Using CAP3 Assembly and Newbler Assembly. The Results Depict the Distribution of EST Data Searched against nr Protein Database by Blastx*

| | **CAP3 Contig** | | **CAP3 Singleton** | | **Total CAP3 Assembly** | | **Total Newbler Assembly** | |
|---|---|---|---|---|---|---|---|---|
| Homology | N | % | N | % | N | % | N | % |
| $0 \leq E \leq 10^{-100}$ | 1145 | 3% | 0 | 0% | 1145 | 2% | 698 | 2% |
| $10^{-100} < E \leq 10^{-50}$ | 3280 | 9% | 2 | 0% | 3282 | 5% | 2020 | 4% |
| $10^{-50} < E \leq 10^{-20}$ | 6641 | 18% | 4883 | 14% | 11524 | 16% | 7896 | 17% |
| $10^{-20} < E \leq 10^{-5}$ | 3015 | 8% | 2946 | 8% | 5961 | 8% | 5560 | 12% |
| Total significant match ($E \leq 10^{-5}$) | 14081 | 39% | 7831 | 22% | 21912 | 31% | 16174 | 35% |
| No hit or $E > 10^{-5}$ | 21823 | 61% | 27649 | 78% | 49472 | 69% | 30254 | 65% |

The researcher compared the unique gene transcripts derived from two unique sequence assembly approaches, CAP3 and Newbler (454 Life Sciences, Branford, CT) to determine the optimal assembly method for our sequence dataset. The results of the two sequence assembly algorithms were compared by searching each set against the NCBI non-redundant protein databases. The CAP3 assembly generated a greater number of significant matches than the Newbler assembly (21,912 vs 16,175) at blastx, $E \leq 10^{-5}$ (Table 4) as well as a greater number of Refseq accession protein IDs (11,667 vs 8,766).

The researcher also compared the gene and protein products in common among the two assemblies and found that 97% (6,158/6,349) of genes identified by Newbler (performed at 454 Life Sciences) were also present in the CAP3 assembly Figure 16.

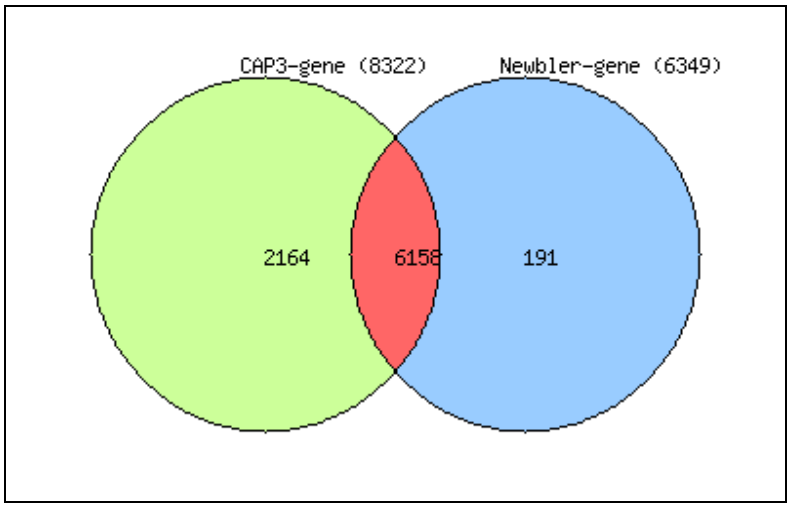


*Figure 16.* Assembling comparison between CAP3 and Newbler against chicken proteome.

The CAP3 assembly provided an additional 2,164 distinct genes as compared to 191 identified by Newbler. Finally, in cross-species protein database comparisons, 80% (7,007/8,784) of Newbler protein hits were similar to CAP3 (Figure 17).

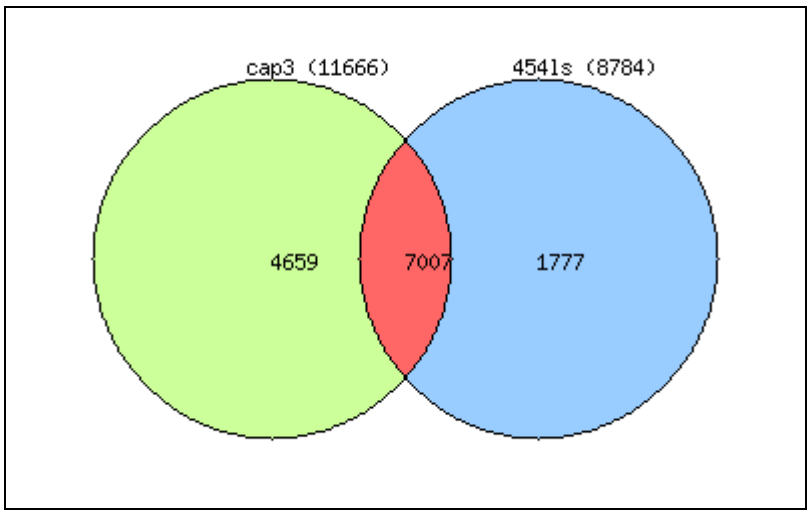


*Figure 17.* Cross species protein database comparison between CAP3 and Newbler assemblies.

The CAP3 assembly resulted in a higher number of significant matches, a greater number of distinct genes, and incorporated most of the protein database matches identified using Newbler. Given these optimal assembly results, the CAP3-assembled sequences were utilized for all the downstream analyses and annotations of ESTs.

*Comparative Sequence Analysis*

The researcher examined the degree to which genes present in Northern bobwhite are conserved across various species by comparison of the unique transcripts to genomes of several model species. Unique transcripts for Northern bobwhite were similar to 18,968 (26.57%), 16,579 (23.23%), 16,144 (22.62%), 6,100 (8.55%), 4,551 (6.38%), and 1,906 (2.67%) genes of chicken, human, mouse (*Mus musculus*), *Drosophila melanogaster*, *Caenorhabditis elegans* and yeast *(Saccharomyces cerevisiae)*, respectively at $E \leq 10^{-10}$ (Table 5).

Table 5

*Genomic Comparison of Northern Bobwhite Transcriptome against nr Database for Model Organisms*

| **Organisms** | **Human** | | **Mouse** | | **Fly** | | **C. elegans** | | **Yeast** | | **Chicken** | | **All Organisms** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Homology | N | % | N | % | N | % | N | % | N | % | N | % | N | % |
| $E \leq 10^{-100}$ | 735 | 1.03% | 714 | 1.00% | 155 | 0.22% | 101 | 0.14% | 46 | 0.06% | 1090 | 1.53% | 1145 | 1.60% |
| $E \leq 10^{-50}$ | 3160 | 4.43% | 3050 | 4.27% | 791 | 1.11% | 549 | 0.77% | 226 | 0.32% | 4229 | 6.39% | 4561 | 6.39% |
| $E \leq 10^{-20}$ | 11938 | 16.72% | 11506 | 16.12% | 3260 | 4.57% | 2299 | 3.22% | 927 | 1.30% | 15166 | 22.90% | 16347 | 22.90% |
| $E \leq 10^{-10}$ | 16579 | 23.23% | 16144 | 22.62% | 6100 | 8.55% | 4551 | 6.38% | 1906 | 2.67% | 18968 | 26.57% | 20297 | 28.43% |
| $E \leq 10^{-5}$ | 18386 | 25.76% | 17964 | 25.17% | 7912 | 11.08% | 6180 | 8.66% | 2770 | 3.88% | 20734 | 29.04% | 21912 | 30.70% |
| No Match | 52998 | 74.24% | 53420 | 74.83% | 63472 | 88.92% | 65204 | 91.34% | 68614 | 96.12% | 50650 | 70.95% | 49472 | 69.31% |

Approximately 850 genes (1.19%) were similar to all six model organisms as well as chicken and 15,498 genes (21.71%) in mammalian model organisms (human and mouse), at $E< 10^{-10}$ (Table 6).

Table 6

*Similar Genes among Model Organisms Represented in Northern Bobwhite Transcriptome*

| | **$E \leq 10^{-100}$** | | **$E \leq 10^{-50}$** | | **$E \leq 10^{-20}$** | | **$E \leq 10^{-10}$** | |
|---|---|---|---|---|---|---|---|---|
| **Organism** | N | % | N | % | N | % | N | % |
| **Human + Mouse** | 697 | 0.98 | 2899 | 4.06 | 10958 | 15.35 | 15498 | 21.71 |
| **Human+Fly+Mouse** | 152 | 0.21 | 757 | 1.06 | 3087 | 4.32 | 5749 | 8.05 |
| **Human+Fly+Mouse+Chicken** | 150 | 0.21 | 741 | 1.04 | 3021 | 4.23 | 5627 | 7.88 |
| **Human+Fly+Mouse+C.elegans** | 98 | 0.14 | 503 | 0.7 | 2016 | 2.82 | 3959 | 5.55 |
| **Human+Fly+Mouse+C.elegans+ Yeast** | 40 | 0.06 | 170 | 0.24 | 559 | 0.78 | 949 | 1.33 |
| **Human+Fly+Mouse+C.elegans+ Yeast+Chicken** | 40 | 0.06 | 168 | 0.23 | 550 | 0.77 | 929 | 1.3 |
| **Human+Fly+C.elegans+Yeast+ Mouse+A.thaliana** | 35 | 0.05 | 151 | 0.21 | 501 | 0.7 | 866 | 1.21 |
| **Human+Fly+C.elegans+Yeast+ A.thaliana+Mouse+Chicken** | 35 | 0.05 | 151 | 0.21 | 494 | 0.69 | 850 | 1.19 |
| **Fly+C.elegans** | 99 | 0.14 | 504 | 0.71 | 2029 | 2.84 | 3991 | 5.59 |
| **Fly+C.elegans+Yeast** | 40 | 0.06 | 170 | 0.24 | 563 | 0.79 | 953 | 1.34 |
| **Fly+C.elegans+Yeast+A.thaliana** | 35 | 0.05 | 151 | 0.21 | 505 | 0.71 | 870 | 1.22 |
| **C.elegans+Yeast** | 40 | 0.06 | 170 | 0.24 | 568 | 0.8 | 969 | 1.36 |
| **C.elegans+Yeast+A.thaliana** | 35 | 0.05 | 151 | 0.21 | 506 | 0.71 | 879 | 1.23 |
| **Yeast+A.thaliana** | 37 | 0.05 | 159 | 0.22 | 533 | 0.75 | 928 | 1.3 |

Not surprisingly, many genes with functions in central metabolism and protein synthesis such as eukaryotic translation initiation factor 1, glyceraldehyde-3-phosphate dehydrogenase, ribosomal protein, methylenetetrahydrofolate dehydrogenase , actin, lysyl/glycyl/threonyl tRNA synthetase, proteasome 26S ATPase, UDP-glucose pyrophosphorylase, heat shock protein, fumarate hydratase were found to be conserved across all organisms at $E<10^{-100}$. Overall, Northern bobwhite had the greatest protein coding gene homology to chicken when comparing against genomes of model organisms (Table 5) therefore we selected the chicken genome to further investigate phylogeny.

*Assessment of Ortholog Detection*

Gene ortholog detection is utilized to mine the sequence data that will provide higher accuracy in predicting ortholog and paralog relationships. Prediction of protein-coding regions was established using ESTScan which uses a hidden Markov model to identify coding regions, even if the quality is low and contains frame shifts, a common sequencing error (Iseli *et al*., 1999). Open-reading frames (ORF) were found for 39,400 (22,432 contigs and 16,968 singlets) of the 71,384 unique transcripts with 20,660 located in the 5' end and 18,740 in the 3' end of the transcripts (see Protein Domain section below). The average read length for the ORFs was 352.3 bp (min = 60, max = 4500; SD = 232.2). A total of 18,647 proteins from translated ORFs (47.3% of total) were found to have significant blastp matches at $E \leq 10^{-5}$ and 18,499 (99.2%) of these mapped to the putative homologs. These predicted proteins were analyzed using a high-throughput automated annotation pipeline yielding annotation for an additional 2.3% of proteins not recognized by blastp. The remaining predicted proteins (50.4%) detected by the ESTScan model may

represent false positives or homologs that might be detected by less stringent blast cutoff values.

Phylogenetic trees provide accurate, however computationally intensive detection of orthology for putative homologs. A less computationally demanding method is reciprocal blast hit (RBH) analysis, an alternative method frequently used as a shortcut for ortholog detection (Moreno-Hagelsieb G, 2008). The rationale is that two genes in different genomes are considered orthologous if they match as best hits when each full genome is queried against the other (Bork *et al*., 1998). RBH to assess ortholog detection among Northern bobwhite and chicken (*Gallus gallus*, an intra-order phylogenetic relative of Northern bobwhite) across both the putative homologs and predicted ORFs that might lead to orthologous genes is performed (Figure 18).

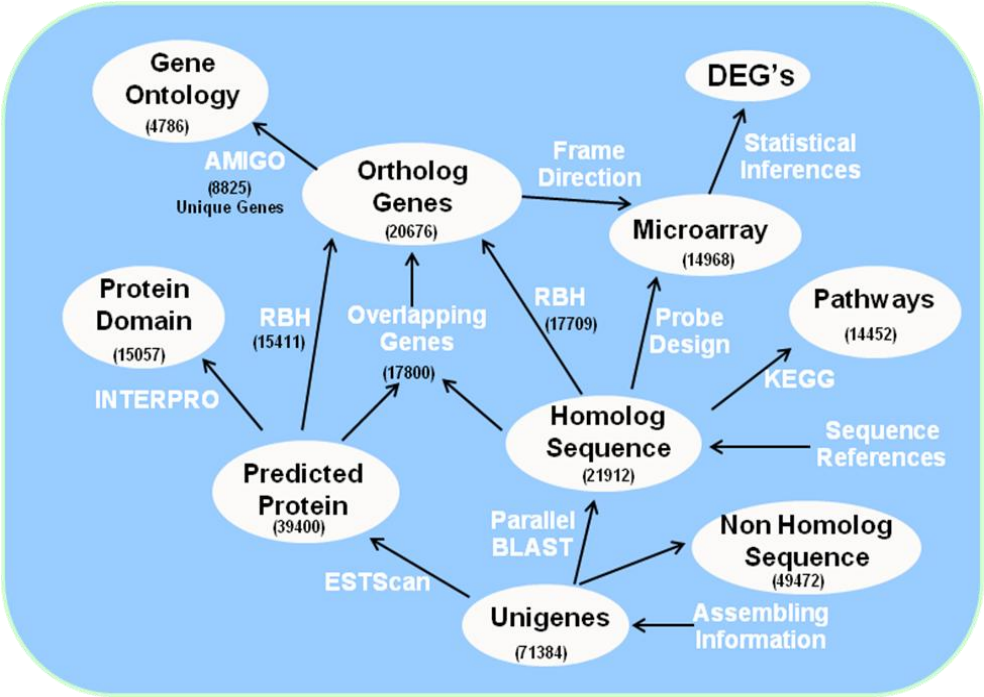


*Figure 1.* Flowchart describing the work process of the ortholog detection and annotation.

The blast hits ($E \leq 10^{-5}$) from chicken-bobwhite and bobwhite-chicken were sorted on minimum E value (and maximum bit score if more than one hit with same E value existed). The researcher found 17,709 orthologs for the putative homologs and 15,441 orthologs for the predicted proteins from translated ORFs found by ESTScan. Out of orthologs predicted by ESTScan, 14,724 (95.3%) were also present in orthologs from the putative homologs and the remaining in each dataset were either unique or had different ortholog matches. It is observed that 17,800 putative homolog-ORF pairs (where matching protein identifiers were sorted on minimum E value for blast hits) were complementary in 85% of comparisons to the RBH pairs. To maximize the ortholog count for the Northern bobwhite, the conjoined ortholog sets derived from each method resulting in 20,676 non-redundant orthologous unique transcripts that represent 8,825 unique gene products (Figure 18). The number of orthologs corresponds to approximately 48% of the *Gallus gallus* proteome, suggesting that approximately half of coverage of chicken proteome have been achieved by the Northern bobwhite transcriptome. The summation of these orthologs provides a broad representation of predicted molecular functions inherent in the Northern bobwhite transcriptome that is utilized to establish functional annotations as described in the following section.

*Functional Annotation*

To functionally annotate the 8,825 Northern bobwhite orthologs, the researcher investigated and inferred putative GO annotations finding matching annotations for 4,786 (54.2%) genes. The distribution of GO annotation within the first-level molecular function, biological process and cellular component is shown in Figure 19A.

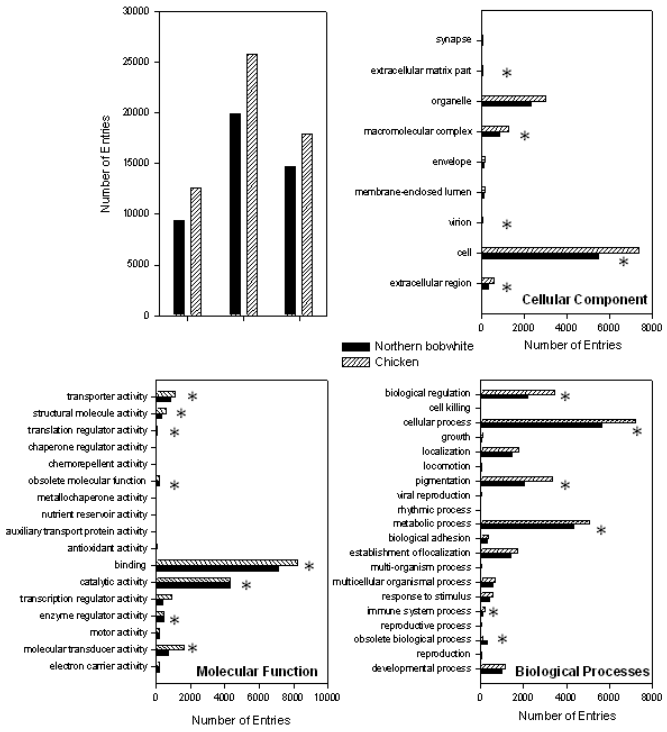


*Figure 2.* Gene ontology (GO) comparison between Northern bobwhite (*Colinus virginianus*) and domestic chicken (*Gallus gallus*). Panel A represents first level GO information. Panel B, C and D represent GO at the 2nd level for cellular component, biological process and molecular function respectively. A Pearson Chi-Square test was used to determine statistically significant matches (p value < 0.05) among species for GO categories (statistically significant categories denoted by "*").

The number of second-level GO sub-categories within each first-level GO category is represented for Northern bobwhite and chicken in Figure 19B-2D. Statistically significant matches were observed among species for a variety of second-level GO categories including the top 4 most abundant categories for both for biological processes ("cellular process," "metabolic process," "biological regulation" and "pigmentation") and molecular functions ("binding," "catalytic activity," "molecular transducer activity" and "transporter activity"). These results indicate similar enrichment in these ontology categories among Northern bobwhite and chicken. Annotations for these transcripts were

inferred from the electronic annotation (IEA) evidence level however with manual intervention, these orthologs are candidates that can be updated to Inferred from Sequence Orthology (ISO) evidence level (http://www.geneontology.org/GO.evidence.shtml). The GO categories of these orthologs can be browsed through the GO browser implemented from amigo (Harris *et al*., 2006b) that is available at www.quailgenomics.info.

*High-Throughput Automated Annotation*

The automated annotation of the Northern bobwhite transcript dataset is performed to efficiently identify protein domains and incorporate proteins within functional molecular pathways (Mao *et al*., 2005). The different approaches for annotation, pathway assignment and protein domain prediction are applied to develop insight into the Northern bobwhite genome which has limited information in public repositories.

*Pathway Assignment*

The researcher assigned metabolic pathways based on sequence similarity (E-value $\leq 10^{-5}$) to sequences with known KEGG pathways. Pathway identification of these unique transcripts were performed by statistical enrichment with Fisher exact test and Benjamini and Hochberg FDR correction against the *Gallus gallus* pathways resulting in 14,452 KO entries for 5,907 unigenes (Figure 20).

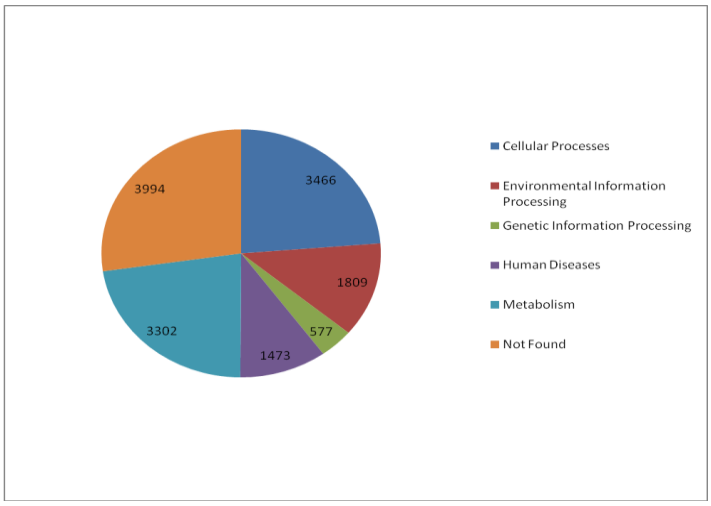


*Figure 3.* Distribution of KEGG orthology (KO) for Northern bobwhite for the first level of hierarchical organization.

In the second level of the KO (Figure 21), endocrine and immune system in cellular processes, signal transduction in environmental information processing, translation in genetic information processing, cancer in human diseases, amino acid and carbohydrate metabolism in metabolism were found to be most abundant.

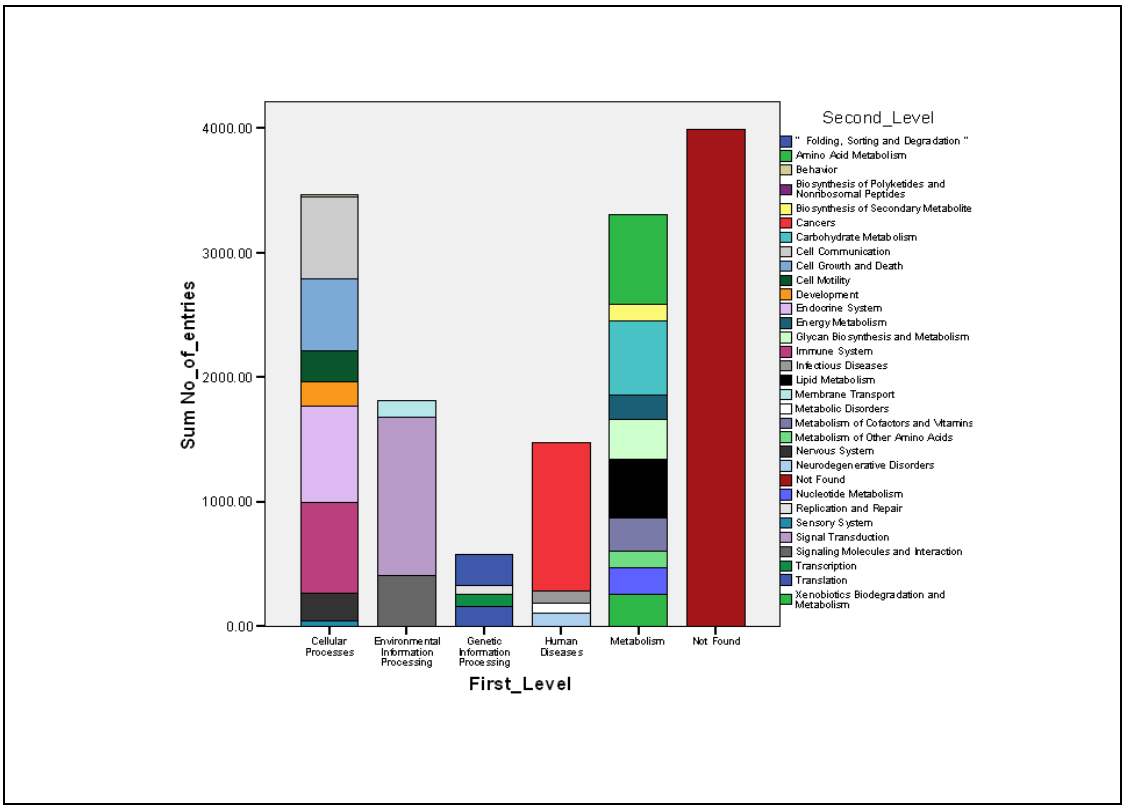


*Figure 4.* Distribution of KEGG orthology (KO) for Northern bobwhite represents the second level of hierarchical organization.

Among the most frequently resolved identities at the third level of organization (Figure 22) were receptors and channels, protein kinases, MAPK signaling pathway, calcium signaling pathway and neuroactive ligand-receptor interaction.

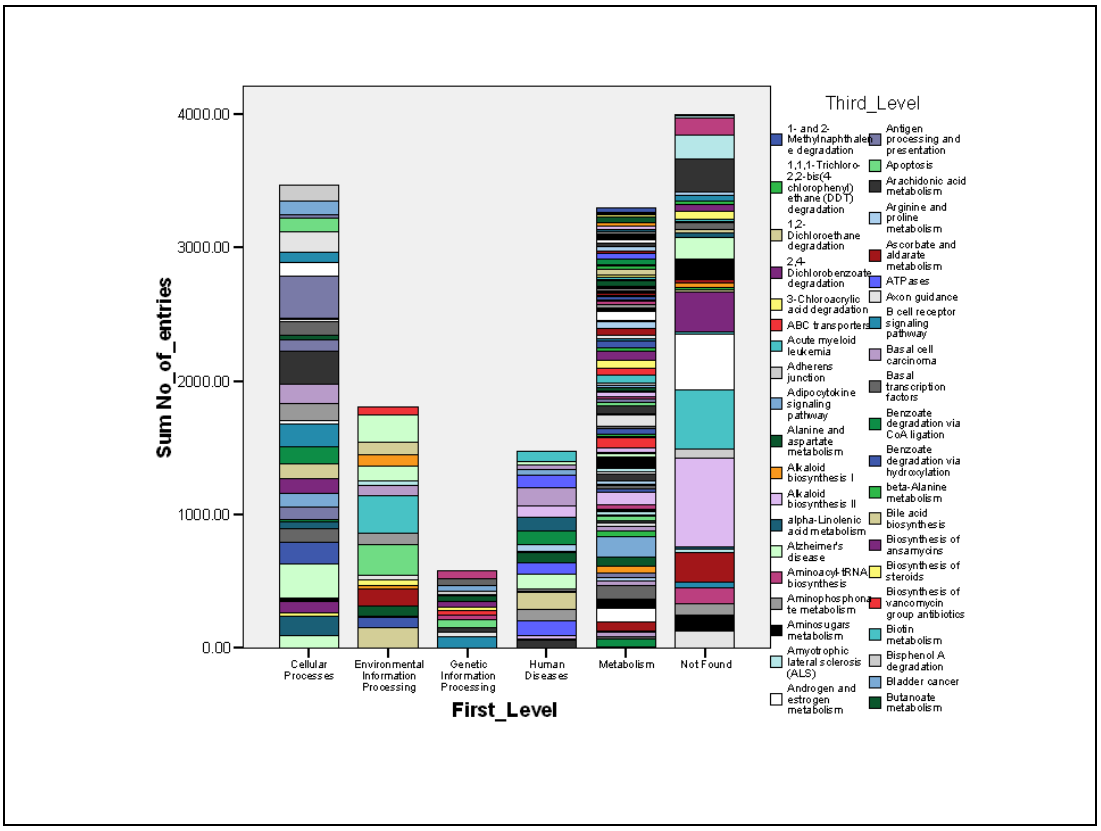


*Figure 5.* Distribution of KEGG orthology (KO) for Northern bobwhite represents the third level of hierarchical organization.

*Protein Domain Prediction*

The 20,954 putative coding regions predicted by ESTSCAN is compared for 21,912 unigenes ($E \leq 10^{-5}$) and it is found that 7,240 (33% of significant blast hits) were not assigned into any domain and family. The remaining 14,672 (66.95%) grouped into 2,453 domains and protein families. The most common domains were protein kinase (245), WD-40 repeats (122), ankyrin (96), zinc finger (91) and RNA recognition motif (84).

The 18,446 predicted coding regions for the remaining 49,475 non-significant unigenes ($E > 10^{-5}$) queried resulted in 385 (.78 %) unique matches in Interpro database. The most common domains were TonB box N-terminal (52), Phosphopantetheine

attachment site (18) & PS00014 (16) involved in post-translation modification and beta tubulin autoregulation binding site (12).

*Microarray Analysis: Effects of 2,6-DNT Exposure on Transcript Expression*

Several toxicological impacts have been observed in Northern bobwhite dosed with 2,6-DNT (Quinn, Jr. *et al*., 2007) at the 10 and 60 mg/kg/d doses referred to as L10 and L60. The microarray results from Bayesian *t*-test with the results of parametric *t*-test are compared and a substantial overlap of differentially expressed genes (DEG) among the two analyses is found including 182 and 396 common genes relative to controls for L10 and L60 respectively (Figure 23A and 23B). The researcher proceeded by utilizing the results of Bayesian *t*-test for the purposes of investigating 2,6-DNT effects.

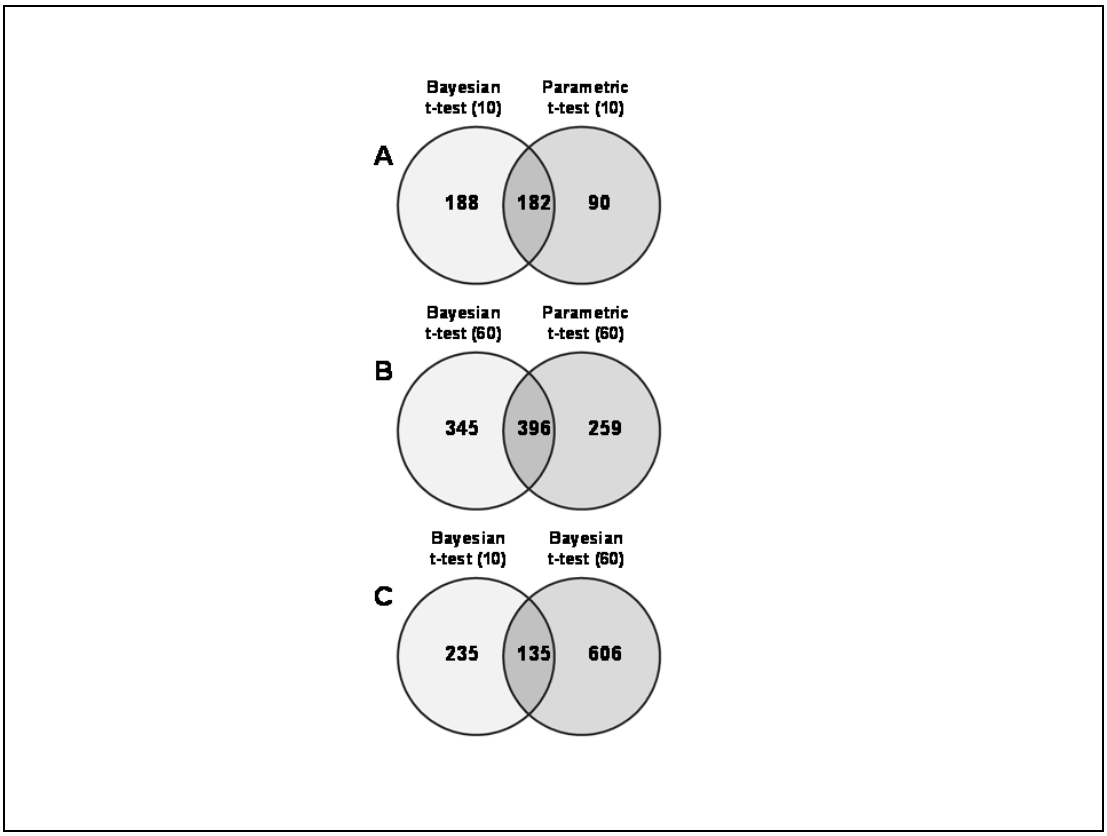


*Figure 6.* Results of microarray analyses identifying significant differential expression of transcripts relative to controls in response to a 60d exposure to 2,6-DNT. Panels A and B compare the results of a parametric t-test (p-value < 0.05 and fold change = 1.8) and Bayesian *t*-test (p-value < 0.01) investigating differential expression in liver tissues at the 10 and 60 mg/kg/d doses respectively. Panel C provides comparison of differentially expressed transcripts among 10 and 60 mg/kg/d 2,6-DNT doses.

The number of DEG relative to unexposed controls (0 mg/kg/d) at $p < 0.01$ nearly doubled with the increase in 2,6-DNT dose from 372 to 750 in the L10 and L60 treatments, respectively, with 138 DEG in common among doses (Figures 23C). KEGG pathway and GO associations were mined from the Northern bobwhite annotated library for all differentially expressed genes (Appendixes A and B). Further enrichment of GO associations to establish biologically meaningful contexts connected to expression changes were performed with the help of the database for annotation, visualization, and integrated discovery (Dennis *et al*., 2003; Huang *et al*., 2009) and GenMAPP.

*Validation of Microarray Analysis*

The correspondence among RT-qPCR and microarray results was generally good at 73.5% and 76.5% for the L10 and L60 treatments, respectively (Table 4). The correlation of fold-change results among the expression assays was highly significant ($P < 0.001$) and approximated a 1:1 relationship as evidenced by regression equations $y = 1.05x - 0.09$, $R^2 = 0.71$ for L60 and $y = 0.80x - 0.12$, $R^2 = 0.86$ for L10 (Figure 24).

| Gene Symbol | Liver, 2,6-DNT (60 mg/kg/d) Array | Liver, 2,6-DNT (60 mg/kg/d) qPCR | Liver, 2,6-DNT (10 mg/kg/d) Array | Liver, 2,6-DNT (10 mg/kg/d) qPCR |
|---|---|---|---|---|
| HMGCS1 | -2.16 | -1.78 | -2.54 | -1.51 |
| ACACA | -1.70 | -1.83 | -0.83 | -0.91 |
| AGMAT | 1.51 | 1.44 | 0.54 | 0.33 |
| AKR1B10 | 1.90 | 1.48 | 1.30 | 0.36 |
| ALB | 4.79 | 3.58 | -0.93 | -0.71 |
| ANXA13 | 3.43 | 4.49 | 0.56 | 0.69 |
| ANXA2 | 3.03 | 3.74 | 0.22 | 0.41 |
| BMP2 | 2.75 | 2.74 | 1.04 | 0.95 |
| CDC16 | -1.92 | 0.38 | -0.70 | 0.38 |
| CPOX | 0.98 | 2.10 | 0.17 | 0.34 |
| CYP7A1 | 0.92 | 0.68 | 2.34 | 2.25 |
| ENO2 | 2.72 | 4.93 | 0.27 | -2.25 |
| FASN | -1.83 | -1.76 | -1.00 | -1.16 |
| GOT1 | -0.80 | -0.35 | -0.14 | -0.16 |
| GSTA | 1.52 | -0.55 | -0.22 | 0.10 |
| HMG | -1.29 | -0.61 | -0.42 | 0.04 |
| IARS | 0.93 | -0.29 | 0.55 | 0.37 |
| LDHA | 1.10 | 0.94 | 0.99 | 0.77 |
| LOC418187 | -1.67 | -1.93 | -1.12 | -1.34 |
| LOC427984 | 2.63 | 2.88 | 0.29 | 0.37 |
| LOC429084 | 4.47 | 4.35 | 0.31 | -0.23 |
| MDH1 | -1.22 | -0.53 | -0.41 | 0.25 |
| ABCB4 | 1.24 | 0.18 | 0.12 | -0.72 |
| NDUFB5 | -0.09 | -0.56 | 0.71 | 0.09 |
| ODC1 | 1.61 | 1.77 | 0.11 | -0.75 |
| PAH | -1.72 | -1.30 | -0.26 | 0.01 |
| PCK1 | -3.58 | *NA | -4.17 | *NA |
| PPARGC1A | -1.10 | -4.10 | -1.13 | -1.87 |
| RYR3 | -1.16 | -5.42 | -0.90 | -0.55 |
| SLC5A1 | 1.39 | 3.90 | 0.58 | 1.65 |
| TRIP12 | 0.49 | -0.14 | -0.51 | 0.20 |
| UGDH | -0.93 | -1.46 | -0.76 | -0.77 |
| UROS | 1.20 | 1.02 | 0.58 | -1.27 |
| XDH | -1.72 | -1.76 | -1.01 | -0.55 |

*NA - No Amplification

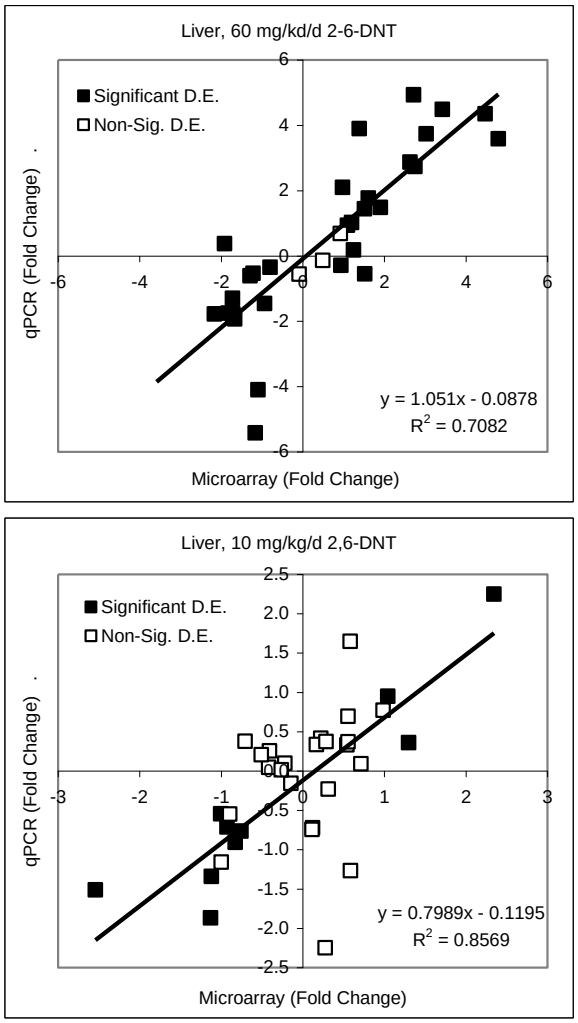


Figure 7. Comparison of RT-qPCR and microarray results. Values represent fold change ($Log_2$) in transcript copy number relative to controls. Red and green highlighted cells represent statistically significant increases and decreases in copy number, respectively. Regression analyses represent correlations in fold change among microarray and RT-qPCR results. Regression equations and $R^2$ values provide correlations for significant differentially expressed transcripts. Gene names and primer sequences are provided in Appendix C.

These results indicate that the microarray analysis was generally accurate for liver tissue lending confidence to the targets that were specifically assayed with RT-qPCR and to

the greater microarray results set used to develop the global interpretation of 2,6-DNT toxicity in liver tissue.

*Functional Impacts of 2,6-DNT Exposure*

Exposure to 2,6-DNT caused a variety of toxicological impacts including gross-level effects and physiological effects as evidenced by alterations in blood chemistry in Northern bobwhite (Quinn, Jr. *et al*., 2007). Based upon observable physiological effects, the lowest level at which 2,6-DNT adversely effected both male and female quail was 40 mg/kg/d based on hematological measures while the only statistically significant change observed at 10 mg/kg/d was reduced blood-glucose levels in males. The majority of impacts occurred in a dose-response manner where the 10 mg/kg/d dose was affected in the same relative direction as the 60 mg/kg/d dose, however not significantly. Here, we found significant changes in several physiological pathways at 10 and 60 mg/kg/d that were both dose responsive and dose dependent corresponding with the observed toxicological phenotypes. Further, genomic results detected 2,6-DNT impacts on gene expression (Figure 24) below the threshold at which adverse effects were manifested. We investigated pathways, GO terms and individual gene targets significantly impacted in 2,6-DNT treatments to determine potential mechanisms underlying observed gross-level effects and effects on blood chemistry and, further, explored genomics-directed observations to provide a systemic understanding of the general pharmacology of 2,6-DNT in Northern bobwhite (Figure 25).

## Gross-level Effects:

**Edema in gastrointestinal tract**

GenMapp: Prostaglandin Synthesis and Regulation (L60: 4 up, 1 down) - Figure 5A

**Liver lesions in 11 of 12 males and 10 of 12 females in the 60 mg/kg/d treatments. Oval cell / biliary hyperplasia**

GO:0006950: response to stress (L10: 5 up, 5 down)
GO:0009611: response to wounding (L60: 2 up, 5 down)
GO:0042060: wound healing (L60: 2 up, 3 down)

**Brown pigmentation accumulation in the Kupffer cells (Liver) in all birds @ 60mg/kg/d and Dark brown granular pigmentation accumulation in spleen.**

KEGG:gga00860:Porphyrin and chlorophyll metabolism (L60: 2 up)
GO:0007596: blood coagulation (L60: 2up, 3 down)
GO:0007599: hemostasis (L60: 2 up, 3 down)

**Diarrhea & weight loss as well as death in 3 and 4 birds in the 40 and 60 mg/kg/d treatments respectively**

KEGG:gga00860:Porphyrin and chlorophyll metabolism (L60: 2 up)

## Blood Chemistry Results:

**Reduction in plasma glucose levels**

KEGG:gga00010: Glycolysis / Gluconeogenesis (L60: 2 up, 5 down)
KEGG:gga00020: Citrate cycle (TCA cycle) (L60: 1 up, 5 down; L10: 1 down. 1 common gene and response among doses)
KEGG:gga00030: Pentose phosphate (L60: 1 down)
KEGG:gga00040: Pentose and glucuronate interconversions (L60: 1 up, 1 down; L10: 1 up, 1 down). 2 common genes and responses among treatments)
KEGG:gga00051: Fructose and mannose metabolism (L60: 1 up, 1 down; L10: 1 up, 1 down. 1 common gene and response among treatments)
KEGG:gga00052: Galactose metabolism (L60: 1 up, 1 down; L10: 1 up. 1 common gene and response among treatments)
KEGG:gga00190: Oxidative phosphorylation (L60: 3 up, 2 down, L10: 1 down; 1 common gene and response among treatments).
KEGG:gga00500: Starch and sucrose metabolism (L60: 1 up, 5 down; L10: 1 down. 1 common gene and response among treatments)
KEGG:gga00620: Pyruvate metabolism (L60: 2 up, 4 down; L10, 1 up, 2 down. 3 common gene responses and gene responses among treatments)
GO:0006091: generation of precursor metabolites and energy (L60: 7 up, 15 down)
GO:0016491: oxidoreductase activity (L60: 14 up, 25 down; L10: 9 up, 9 down)

**Reduction in RBCs and Hemoglobin Concentrations**

KEGG:gga00860:Porphyrin and chlorophyll metabolism (L60: 2 up)

**Reduction in Albumin (>50%), Globulin, and Total Protein**

Albumin (ALB, NP_990592) (L60: up; L10 down)
KEGG:gga04120: Ubiquitin mediated proteolysis (L60: 4 down; L10: 3 up, 4 down)
GO:0006519: amino acid and derivative metabolic process (L60: 4 up, 8 down)
GO:0006520: amino acid metabolic process (L60: 4 up, 7 down)
GO:0009063: amino acid catabolic process (L60: 1 up, 3 down)
GO:0009308: amine metabolic process (L60: 8 up, 5 down)
GO:0009310: amine catabolic process (L60: 1 up, 3 down)
GO:0016504: protease activator activity (L10: 1 up, 1 down)

**Reduction in Aspartate Amino Transferase Concetrations**

Aspartate aminotransferase (also known as Glutamic-oxaloacetic transaminase, GOT) (L60: 1 down)

**Increase in Uric Acid**

Aspartate aminotransferase (also known as Glutamic-oxaloacetic transaminase, GOT) (L60: 1 down)
KEGG:gga00860:Porphyrin and chlorophyll metabolism (L60: 2 up)
KEGG:gga00220: Urea cycle and metabolism of amino groups (L60: 2 up, 1 down)
KEGG:gga00910: Nitrogen metabolism (L 60: 2 down; L10, 1 down)
GO:0044270: nitrogen compound catabolic process (L60: 1 up, 3 down)

| Reduction in Na+ and K+ Ion Concentration |
|---|
| GO:0030001: metal ion transport (L60: 11 up, 3 down) |
| GO:0006812: cation transport (L60: 13 up, 3 down) |
| GO:0015672: monovalent inorganic cation transport (L60: 12 up, 0 down) |
| **Genomics-Directed Observations:** |
| **2,6-DNT Metabolism** |
| KEGG:gga00480: Glutathione metabolism (L60: 3 up) |
| KEGG:gga00624:1- and 2-Methylnaphthalene degradation (L60: 1 up; L10; 1 up. 1 common gene and response among doses) |
| KEGG:gga00626:Naphthalene and anthracene degradation (L10: 1 up). |
| KEGG:gga00980: Metabolism of xenobiotics by cytochrome P450 (L60: 1 up) |
| GO:0006725: aromatic compound metabolic process (L60: 0 up, 6 down) |
| GO:0006807: nitrogen compound metabolic process (L60: 5 up, 9 down) |
| GO:0016775: phosphotransferase activity, nitrogenous group as acceptor (L60: 3 up, 1 down) |
| GO:0016840: carbon-nitrogen lyase activity (L60: 3 down) |
| **Impacts on Lipid Metabolism** |
| KEGG:gga00061: Fatty acid biosynthesis (L60: 2 down; L10: 1 down. 1 common gene and response among doses) |
| KEGG:gga00071: Fatty acid metabolism (L60: 2 down; L10: 1 down. 1 common gene and response among doses) |
| KEGG:gga00564: Glycerophospholipid metabolism (L60: 3 up, 1 down; L10: 2 up. 2 common genes and responses among doses) |
| KEGG:gga00565: Ether lipid metabolism (L60: 1 up, 2 down) |
| KEGG:gga00590: Arachidonic acid metabolism (L60: 1 down) |
| KEGG:gga01040: Polyunsaturated fatty acid biosynthesis (L60: 1 up, 2 down; L10: 1 up) |

*Figure 8.* Gene ontology (GO) and KEGG pathway entries that best described the toxicological phenotypes observed in Northern bobwhite exposed to 2,6-DNT for 60d were used to gain toxicogenomic insights into the mechanisms underlying the toxicological effects. Toxicogenomic assay investigated effects in liver (L) tissues of animals exposed to 2,6 DNT at 10 and 60 mg/kg/d relative to tissue specific controls. Reference to "up" and "down" in the table refers to either increased or decreased transcript expression relative to controls.

*Gross-Level Effects*

The high dose of 60 mg/kd/d, 2,6-DNT caused edema in the gastrointestinal (GI) tract of most birds with corresponding impacts on genes within the "prostaglandin synthesis and regulation" pathway (Figure 25), which is key to the inflammatory response in mammalian models. Annexins (ANXA) are critical components of the inflammatory pathway (Figure 26) which are recognized to inhibit prostaglandin production in mammalian models (Green *et al*., 1998).

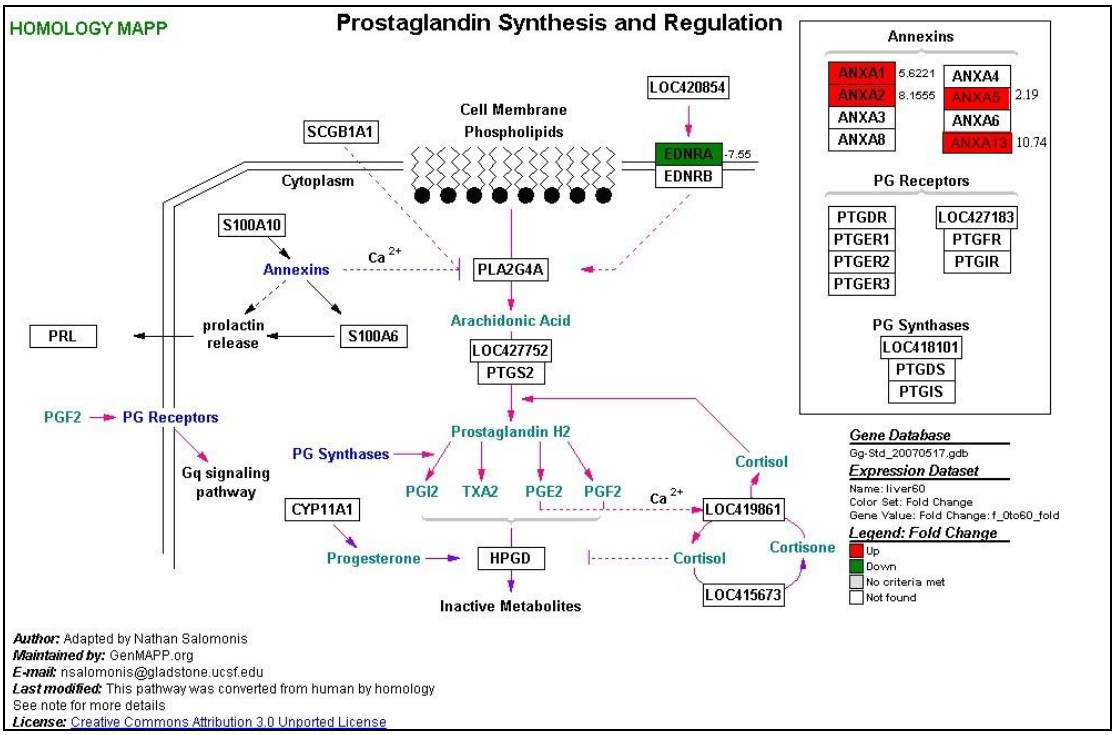


*Figure 9.* Effects of 2,6-DNT exposure on the prostaglandin synthesis and regulation pathway in liver tissue of Northern bobwhite dosed with 60 mg/kg/d, 2,6-DNT in a 60d exposure. Significant impacts on transcript expression relative to controls are represented by red (increased expression) and green (decreased expression).

Anxa1,-2,-5, -13 were over-expressed (Figure 24) at p-value < 0.01 (Anxa7 and -8 were over-expressed at p-value < 0.05) in the L60 treatment. Additionally, expression of the G protein-coupled receptor, endothelin A receptor (EDNRA), which regulates endothelin-1-mediated induction of prostaglandin E2 (Leis *et al*., 1998; Shames *et al*., 2007), was reduced. Although prostaglandins are pro-inflammatory in most organs, they are recognized to have anti-inflammatory effects on GI mucosa (Morteau, 2004). Increased expression of annexins as well as decreased expression of EDNRA can each contribute to reduced prostaglandin production which is consistent with increased inflammation observed in the GI tract and liver of Northern bobwhite.

2,6-DNT exposure resulted in liver lesions and oval cell / biliary hyperplasia, in 11 of 12 males and 10 of 12 females in the L60 treatment group, ultimately contributing to 4 deaths (Quinn, Jr. *et al*., 2007). Oval cell hyperplasia is believed to occur in mammalian species as a part of a hepatic-repair process when liver injury exceeds the proliferation capacity of normal hepatocytes, and the observed biliary hyperplasia indicates that the primary site of injury involved the bile duct epithelium (Greaves, 2007). Investigation of GO terms including "response to wounding," "wound healing" and "hemostasis" indicated differential expression of a number of genes including greatly increased expression (4.35 fold, $\log_2$, Figure 24) of tissue factor (LOC429084) as well as increased expression of interleukin 10 receptor beta (IL10RB) in the L60 group. Increased expression of clotting factors such as (LOC429084) is consistent with initiation of clotting at the site of lesion injury (Gouault-Helimann and Josso, 1979) while the increased expression of Interleukin 10 Receptor, beta (IL10RB) is an indicator of enhanced immune response against wound infection.

Brown pigmentation was observed to accumulate in quail liver Kupffer cells in all high-dose (L60) birds accompanied by dark brown granular pigmentation in spleen, diarrhea and weight loss in the majority of birds. 2,6-DNT and related chemicals such as 2,4,6-trinitrotoluene are known to cause anemia in mammalian species *via* erythrolysis which may in turn affect the primary blood conditioning organs including liver, kidney and spleen (Ferguson and McCain, 1999; Talmage *et al*., 1999). Our results indicated that the gene coproporphyrinogen oxidase (CPOX) and a gene similar to uroporhyrinogen-III synthase (UROS, LOC426223), components of the porphyrin and chlorophyll metabolism pathway and an instrumental facilitator of heme biosynthesis, were over-expressed in the

L60 treatment (Figure 24 and 25). Increased expression of CPOX and UROS genes (Figure 27), if uncoupled from heme synthesis, is consistent with accumulation of porphyrin by-products (uroporphyrinogen III, protoporphyrinogen IX) causing porphyria (Elder, 1998).

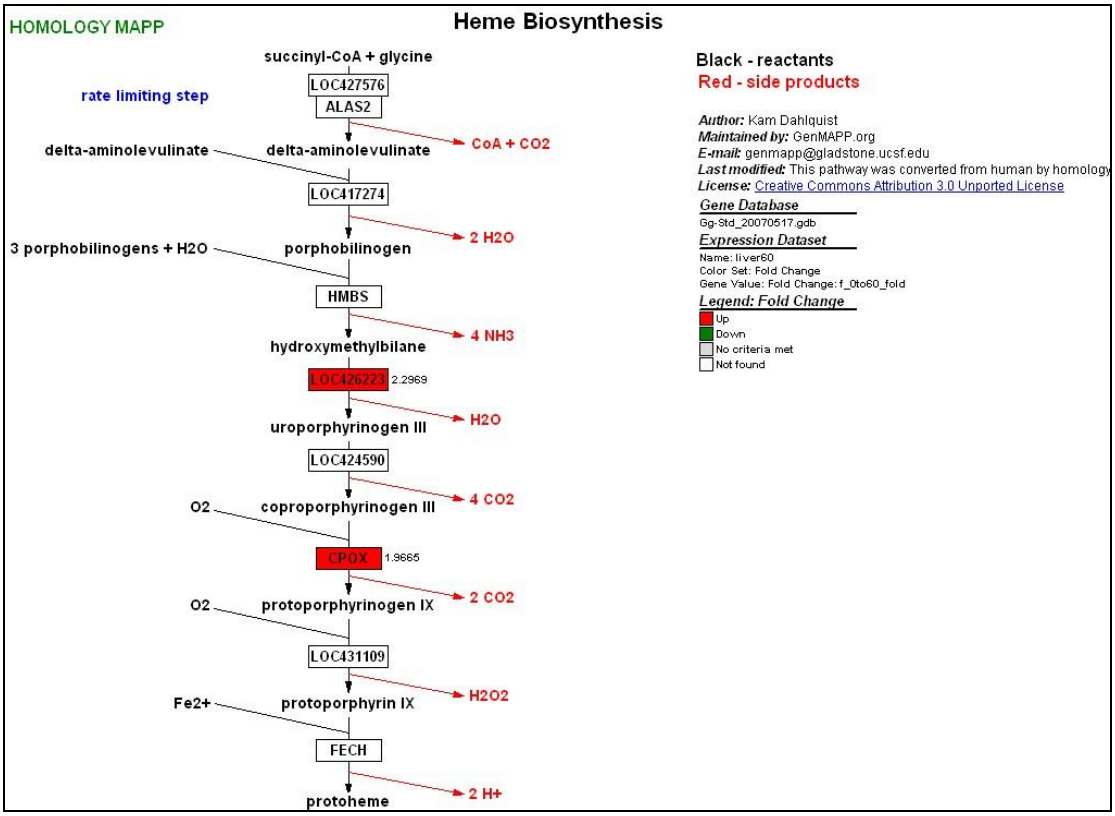


*Figure 10.* Effects of 2,6-DNT exposure on the heme biosynthesis pathway in liver tissue of Northern bobwhite dosed with 60 mg/kg/d, 2,6-DNT in a 60d exposure. Significant impacts on transcript expression relative to controls are represented by red (increased expression) and green (decreased expression).

Porphyrin by-products are indicators of impeded heme cycling and can also result in symptoms similar to those observed in 2,6-DNT exposed quail including diarrhea, loss of sensation, low RBC count and abnormal liver function (http://www.mayoclinic.com/health/). While some gross-level effects such as weight loss

arise from complex impacts on physiology, evidence from transcriptomics provides additional insight into systemic-level responses of the organism. The gene expression results provide plausible mechanisms underlying several gross-level effects of 2,6-DNT.

*Effects on Blood Chemistry*

Changes in blood chemistry parameters were some of the most sensitive indicators of 2,6-DNT exposure with impacts occurring at doses as low as 10 mg/kg/day. A significant decrease was seen in plasma-glucose levels in L10 and L60 birds in absence of 2,6-DNT-related changes in feed consumption (Quinn, Jr. *et al*., 2007). Consistent with reduced glucose levels and weight loss observed in Northern bobwhite, expression of genes represented in 9 pathways and 2 GO categories related to carbohydrate and energy metabolism were affected (Figure 25) with the majority of genes having decreased expression in 2,6-DNT exposures (Appendix A). Investigation of the glycolysis and gluconeogenesis pathway (Figure 28) for the L60 treatment indicated impacts on genes involved in establishing equilibrium between the glycolysis and gluconeogenesis processes.

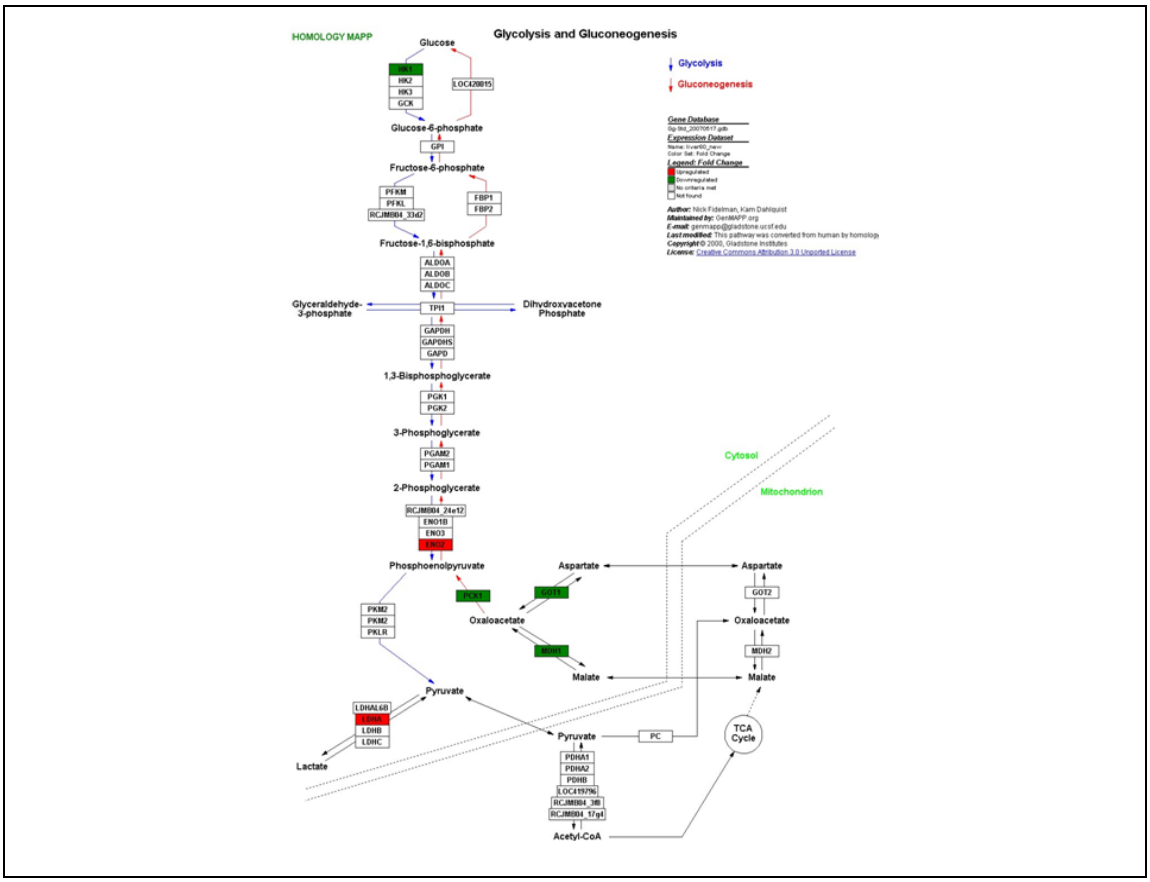


*Figure 11*. Effects of 2,6-DNT exposure on the glycolysis and gluconeogenesis pathway in liver tissue of Northern bobwhite dosed with 60 mg/kg/d, 2,6-DNT in a 60d exposure. Significant impacts on transcript expression relative to controls are represented by red (increased expression) and green (decreased expression) highlights.

Phosphoenolpyruvate carboxykinase 1 (PCK1), recognized as a major regulatory point for gluconeogenesis (Go´mez-Valade´s *et al*., 2008), had -2.7 fold $log_2$ and -3.6 fold $log_2$ decreased expression in L10 and L60 birds respectively (Figure 24). The concerted decreases in expression of (PCK1), aspartate aminotransferase (GOT1), and malate dehydrogenase 1 (MDH1) are consistent with a shift in equilibrium away from gluconeogenesis, therefore limiting glucose synthesis. In contrast, enolase 2 (ENO2), which utilizes the substrate formed in the PCK1-catalyzed reaction (phosphoenolpyruvate),

was up-regulated presumably in response to low substrate concentrations and in response to decreased plasma-glucose levels.

Significant reductions in red blood cell (RBC) concentrations, plasma albumin levels (> 50% reduction), aspartate aminotransferase concentrations, total proteins, globulin concentration, and $Na^+$ and $K^+$ ion concentrations were also seen in L60 birds whereas uric acid concentrations nearly doubled in L60 birds (Quinn, Jr. *et al*., 2007). As described above, the KEGG pathway "porphyrin and chlorophyll metabolism" involved in the "heme biosynthesis pathway" (Figure 27) was affected in the L60 treatments including over-expression of CPOX and UROS (Figure 24, Appendix A). Increased expression of these genes is consistent with a compensatory effort to return hemoglobin and RBCs to normal levels.

A significant increase in albumin transcript copy number (ALB, 3.58 fold, $\log_2$, Figure 24) is consistent with a response to compensate for marked reductions in plasma-albumin levels observed by Quinn *et al* (Quinn, Jr. *et al*., 2007). Reductions in plasma-globulin and total proteins correspond with decreased expression of the majority of genes associated with the "ubiquitin mediated proteolysis" pathway (Appendix A) as well as the GO categories related to amine and amino acid catabolism (Figure 25, Appendix B) indicating a potential inhibition of protein cycling initiated by 2,6-DNT exposure.

Reduced blood concentrations of aspartate aminotransferase, also known as glutamic-oxaloacetic transaminase (GOT1), at 60 mg/kg/day 2,6-DNT, corresponded with reduced GOT1 transcript expression in L60 (Figure 24). In addition to the role of GOT1 in glycolysis and gluconeogenesis described above, it and number of additional genes are components of the metabolic pathways "nitrogen cycle," "urea cycle and metabolism of

amino acid groups," and "porphyrin and chlorophyll metabolism" as well as members of the GO category "nitrogen compound catabolic process" which were impacted in the L60 treatment (Figure 25, Appendixes A and B). Additionally, genes including, agmatine ureohydrolase (AGMAT) and ornithine decarboxylase 1 (ODC1), which are involved in the "urea cycle and metabolism of amino acid groups" pathway (Appendix A), had increased expression in L60 birds (Figure 24). These overall impacts, and most specifically the increased expression of components of the "urea cycle and metabolism of amino acid groups" pathway are consistent with elevated blood-urea concentrations observed in birds exposed to high doses of 2,6-DNT (Quinn, Jr. *et al.*, 2007).

Finally, observed reductions in Na+ and K+ ion concentrations were accompanied by a predominant increase in transcript expression within the GO categories: "metal ion transport," "cation transport" and "monovalent inorganic cation transport" (Figure 25). Cation transport is fundamental to a variety of physiological processes and finding the ultimate cause of depletion is a difficult task. A potential result of depletion is manifested in the increased expression of the Na+/glucose contraporter 1 (SLC5A1, Figure 24) a probable response to increase absorption of glucose from dietary sources (Turk *et al.*, 1994) to supplement already marginalized glucose supply to cells.

*Genomics-Directed Observations*

In addition to having toxicological phenotypes guide our investigation of mechanisms of action, we utilized genomic results to identify impacts beyond the results provided in the toxicological bioassays to improve assessment of 2,6-DNT pharmacology and provide connectivity among systemic effects. Regarding 2,6-DNT pharmacology, pathways involved in phase I and phase II xenobiotic metabolism ("metabolism of

xenobiotics by cytochrome P450" and "glutathione metabolism," respectively) had increased expression in the L60 treatment (Figure 25 and Appendix A). Specifically, glutathione-S tranferase (GSTA) and cytochrome P450 (CYP7A1) had increased expression within these pathways (Figure 24). TNT and related compounds (including 2,6-DNT) can induce these mechanisms (Ekman D R *et al*., 2003; Johnson *et al*., 2000; Reddy G *et al*., 2000) which may ultimately influence toxicity (Sims and Steevens J A, 2008). Additionally, the numerous results indicating that pathways involved in nitrogen and urea metabolism (described above) are being impacted by 2,6-DNT suggest the potential for denitrification of 2,6-DNT and processing of resultant nitrogenous metabolites.

An additional observation provides a systemic link to the impacts on energy metabolism discussed above, in this case, involving perturbations caused by 2,6-DNT on transcript expression of genes and pathways involved in lipid metabolism. An isomer of 2,6-DNT, 2,4-dinitrotoluene (2,4-DNT), has been observed to impact lipid metabolism in the fish model, fathead minnow (*Pimephales promelas*), resulting in decreased expression of a number of genes that regulate fatty acid synthesis and causing phospholipid accumulation in the liver (Wintz *et al*., 2006). Similarly, 2,6-DNT caused decreased expression of several genes involved in lipid metabolism in Northern bobwhite liver (L60) including fatty acid synthase (FASN), a gene similar to long-chain acyl-CoA synthetase 3 (LOC424810), 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR), a gene similar to delta7-sterol reductase (Loc422982), a gene similar to acyl-CoA synthetase long-chain family member (ACSL1), acetyl-CoA carboxylase alpha (ACACA), stearoyl CoA desaturase (SCD) apolipoprotein B (APOB), and acyl-CoA dehydrogenase family 11 (ACAD11). In the fathead minnow model, impacts of 2,4-DNT appeared to be modulated

*via* the peroxisome proliferative activated receptor α (PPARA1) and peroxisome proliferative activated receptor γ (PPARG1) pathways. No significant effect was observed on PPARA1 in Northern bobwhite liver, however PPARG was down regulated ($p = 0.03$, -1.0 fold $log_2$ in L10) in addition to peroxisome proliferative activated receptor γ coactivator 1 (PPARGC1A, -1.1 fold $log_2$ in L10, -1.1 fold, $log_2$ in L60). These impacts on components of the lipid metabolism pathway are an additional indicator of 2,6-DNT-induced perturbation of energy metabolism in Northern bobwhite which ultimately may have forced the stress leading to many of the observed toxicological impacts.

*Impacts on Energy Metabolism is Central to 2,6-DNT Toxicity*

Consumption of 2,6-DNT impacted expression of genes that facilitate energy metabolism including both gluconeogenesis (Figure 28) and lipid metabolism (Figure 29) (Rawat *et al*., 2010c).

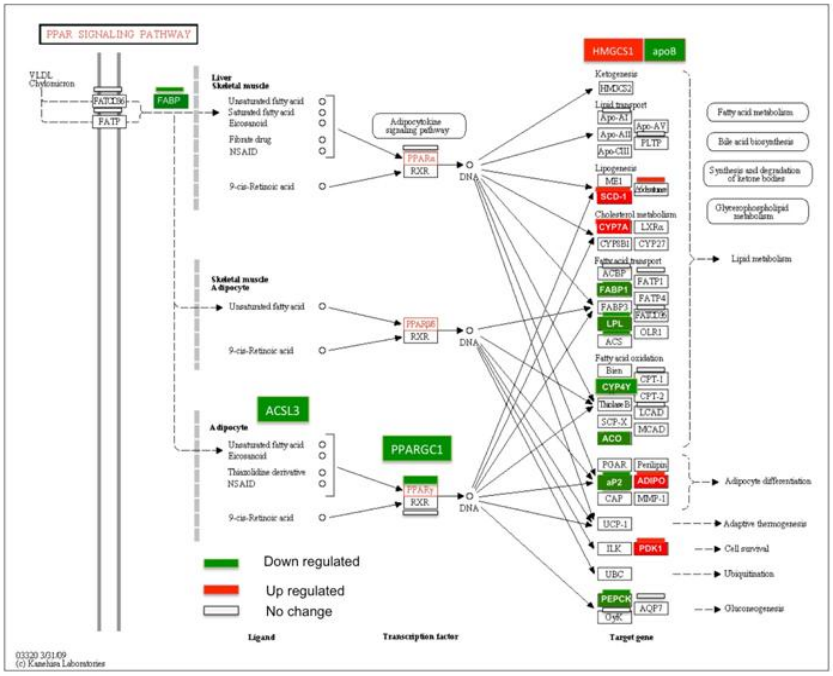


*Figure 12.* Effects of 2,6-DNT exposure on the peroxisome proliferative activated receptor (PPAR) pathway in liver tissue of Northern bobwhite dosed with 60 mg/kg/d, 2,6-DNT in a 60d exposure. Significant impacts on transcript expression relative to controls are represented by red (increased expression) and green (decreased expression) highlights.

Gluconeogenesis is the process by which non-carbohydrate, carbon substrates are converted to glucose, the paramount short-term energy storage molecule. Marked reductions in PCK1 expression, in addition to reduced expression of GOT1 and MDH1 indicate impaired potential for glucose generation (Go´mez-Valade´s *et al*., 2008) which corresponds with reduced glucose levels in blood in the absence of feeding rate changes in Northern bobwhite. From the perspective of energy metabolism, lipids represent medium- to long-term storage molecules. Expression of PPARG1 and PPARGC1A, genes which represent a key control points for initiation of fatty acid metabolism (Figure 29) and

mitochondrial biogenesis (Liang and Ward, 2006), were reduced as was the case for a variety of additional genes involved in lipid metabolic pathways. Reduced expression of these genes is indicative of impairment of lipid catabolism which can ultimately reduce availability of lipid as a substrate for production of cellular energy in times of energetic debt. A study investigating PPARA1 knockout mice indicated increased concentrations of long-chain fatty acids in plasma in addition to impairments in citric acid cycle flux, enhanced urea cycle activity, and increased amino acid catabolism (Makowski *et al*., 2009). If PPAR impacts are conserved in bird species, these results indicate that reduced blood-protein levels and increased blood-urea levels may have resulted from enhanced catabolism of amino acids. Further, these results suggest that 2,6-DNT-induced impairment of glucose and lipid metabolism in Northern bobwhite may have caused reduced availability of carbohydrate- and lipid-based substrates for energy generation, elevating amino acid-based materials as a predominant substrate for this process. This would additionally contribute to a principle gross-level effect observed in 2,6-DNT exposure, weight loss / wasting.

## Knowledgebase Design and Development

Quail Genomics knowledgebase hosts the genomic data for Northern bobwhite which includes nucleotide and protein sequences, meta-data properties and microarray expression data. The knowledgebase provides a central repository for storage, data management and access through a web interface.

### *Annotation Search*

Users can access sequence information (Figure 30) through single query searches (i.e. unigene ID) and batch search (i.e. by blast hit cutoff or microarray experiment). We cross referenced the various entities of data with internal ID that allow comprehensive

annotation search with gene name, protein and regular expression search (i.e. cytochrome p450) that might be of interest to specific users.

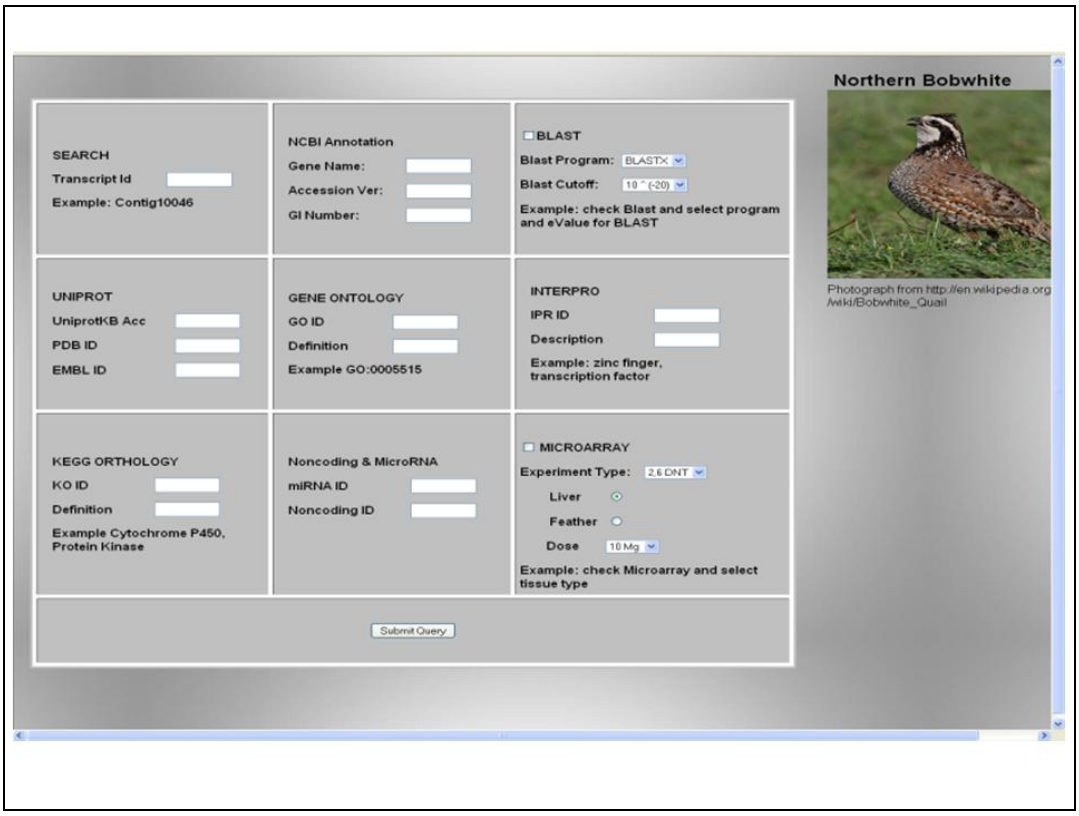


*Figure 13.* Web browser results of query search options for the Quail Genomics.

While browsing through any search (i.e. differentially expressed gene for an experiment), the user can click hyperlinked unigene ID to see the detailed report. The output is provided in tabular form with assembling, sequence, structural properties and metadata (Figure 31).

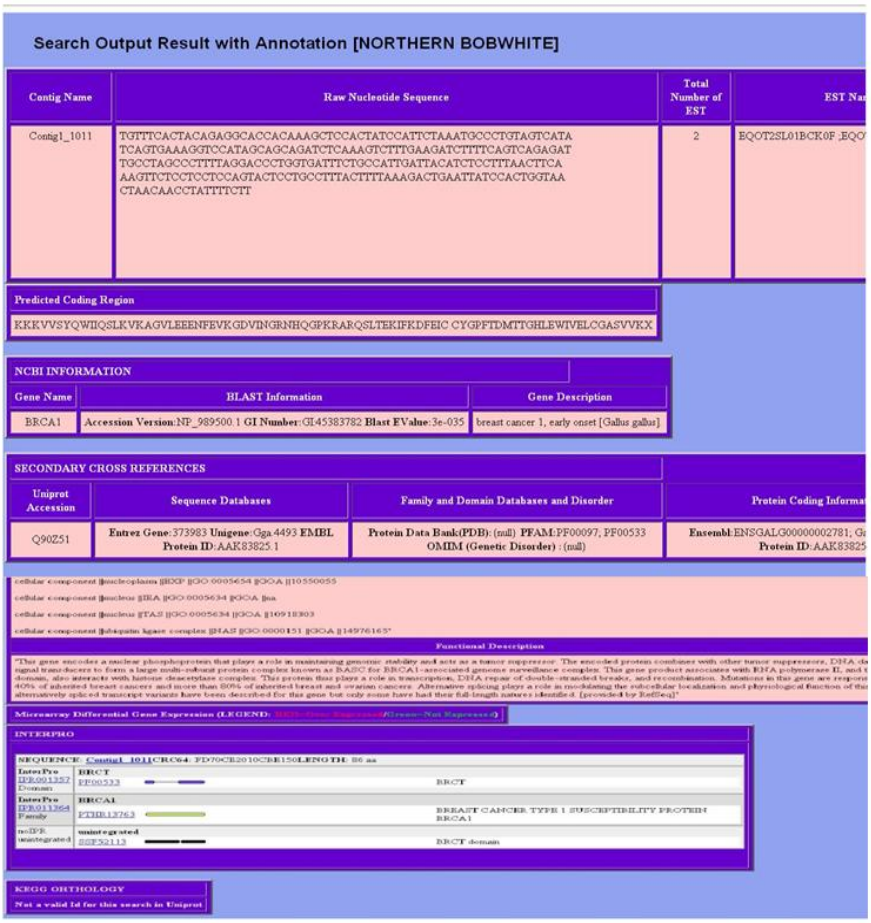


*Figure 14.* Results of the output in the browser executed after performing a parameter search.

*Additional Searches*

Beside annotation search, additional searches have been integrated in this platform. Users can search ESTs that assemble as contigs and visualize the overlap and direction against the assembly. Users can also input their sequence and perform blast search (BLASTN, TBLASTX) against the indexed nucleotide sequences of Northern bobwhite. The expression data is stored based on experiment and dose and output can be viewed for microarray probes sorted on *p*-value. The microarray probes which had a statistically significant increase or decrease in expression (*p*-value<.05) are highlighted in red or green, respectively.

*GO Browser*

The Northern bobwhite orthologs are functionally annotated under the Inferred from the electronic annotation (IEA) evidence level. The GO categories of these orthologs can be browsed for biological processes, cellular components, and molecular functions through the GO Tree Browser implemented from Amigo (Harris *et al*., 2006a) (Figure 32). With guidelines as defined by GO consortium (http://www.geneontology.org/GO.evidence.shtml), these orthologs are candidates that might be considered for update to Inferred from Sequence Orthology (ISO) evidence level.

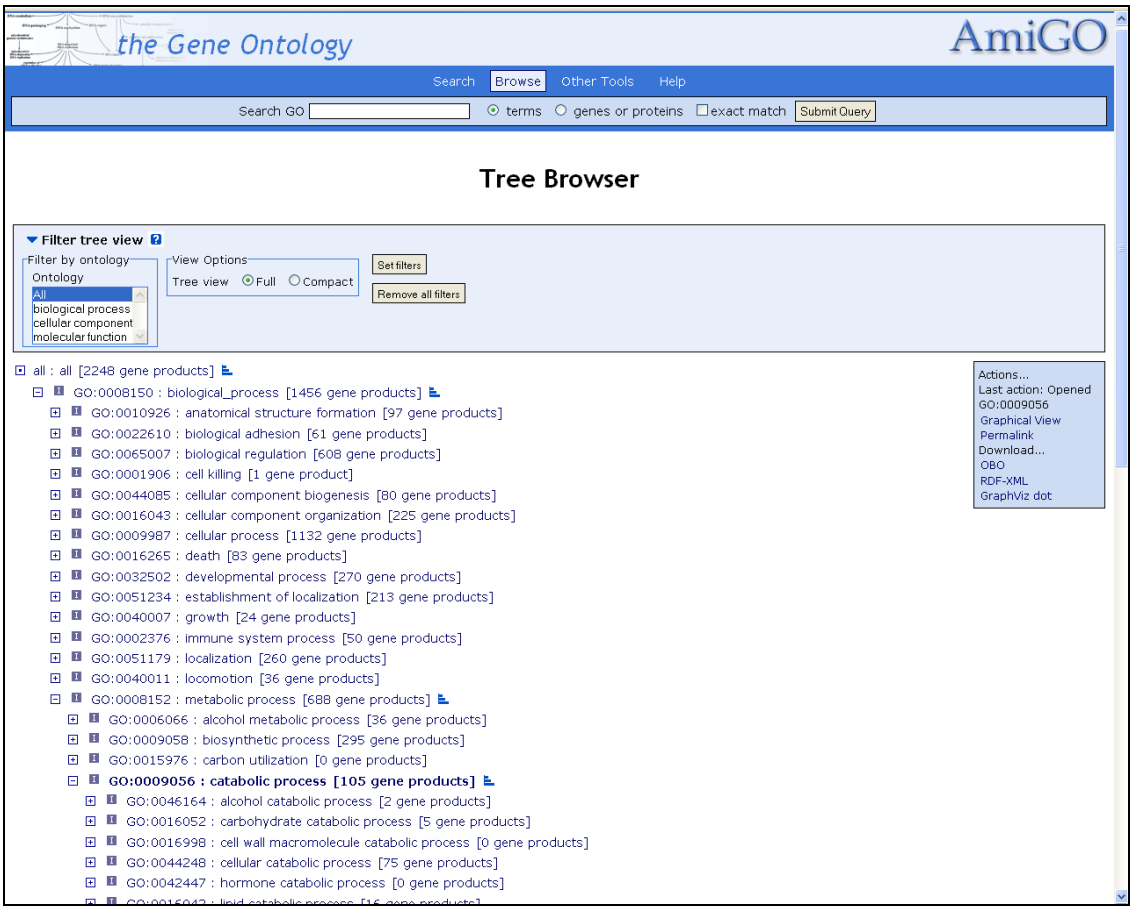


*Figure 15.* GO tree browser locally installed at Quail genomics.

*Genomic Scaffolds*

The high number of unigenes (after assembly) identified for an organism is an over representation of the total protein-coding genes that are expected to make up its genome. Frequently, due to missing EST sequences, the EST's from a single gene may not overlap to assemble to a contiguous sequence resulting in non-overlapping contigs and singletons or splits in genes. Also, stringent parameters during assembly of EST's into contigs might lead to unassembled sequences especially when the sequences have low genome coverage. These issues of missing sequences or fragmentation might lead to partial representation of a protein coding sequence. Many of the sequence fragments might actually represent the same protein leading to redundancy in the assembled sequences.

A pipeline is built that generates scaffolds consisting of multiple-unigenes by aligning partial sequence fragments against specific coding regions of a gene. The scaffolds can be built by querying "gene of interest" which fetches similar unigenes from the database. These are aligned against the indexed protein database of chicken and visualized in a web browser. It might be of interest to further study these scaffolds to understand redundancy or potential alternate spliced elements. In conjunction with Parallel Blast output stored as persistent data and the sequence information stored in relational database management system (RDBMS), our pipeline allows users to interact and visualize results "on the fly."

The user can select parameter BLASTP for predicted proteins from ESTScan and BLASTX for nucleotide sequences and *e-value* cutoff to visualize scaffolds for the Northern bobwhite unigenes. All the unigenes that comprise more than six fragments are listed in the web page with annotation. Clicking on the gene of interest will show frame

direction along with sequence alignment against the indexed *Gallus gallus* Refseq protein sequence data (downloaded February 2008). Northern bobwhite unigenes alignment against the chicken protein sequence for gene F5 is shown for raw nucleotide and predicted protein (Figure 33).

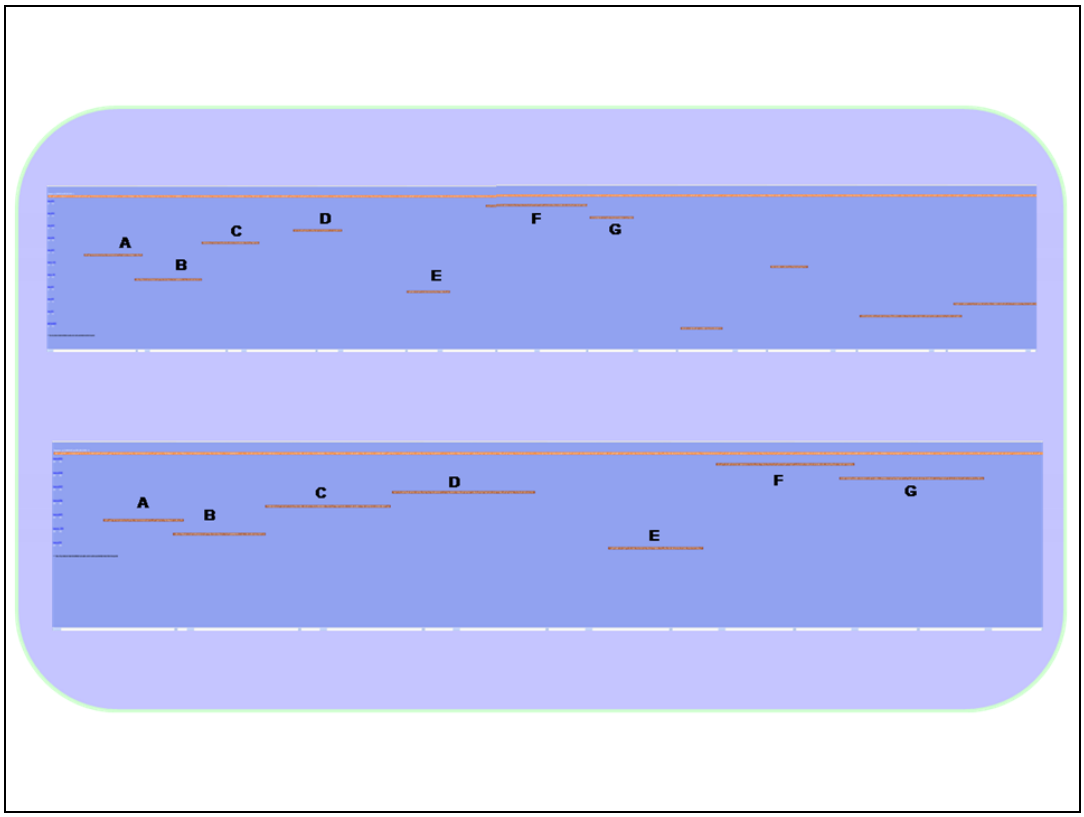


*Figure 16.* The figure depicting the non-overlapping sequences lead to split among the unigenes representing same gene. The upper panel shows Northern *bobwhite* unigenes alignment against the chicken protein sequence for gene F5. The lower panel is alignment against same gene with the Northern *bobwhite* predicted protein. The labels in both panels represent same unigenes for raw nucleotide and predicted protein and roughly expanse the same region of protein-coding chicken sequence. ***** The magnification of the two panels is different.

This example explains the redundancy in the assembled sequences, stemming from the fragmentation of unigenes due to non-overlapping contigs and singletons. The unigenes representing the same protein-coding sequence of chicken fragmented due to either absence

of overlapping sequences or insufficient assembling parameter threshold. This information can also be utilized not only to quality control sequencing coverage for an individual gene coding region but also to study alternate splicing mechanism. Also, more meaningful full-length protein-coding sequence can be built either by tuning *e-value* cutoff or cross matching alignment of raw nucleotide with the predicted protein output. One other advantage of using the scaffold visualized in web browser is that it allows multiple users to access a central system without separate installation of dependencies and software(s) and local databases.

*Past Applications and Results*

The data represented in Quail Genomics knowledgebase have provided insights into the metabolic perturbations underlying several observed toxicological phenotypes in a 2,6-DNT-exposure case study investigating Northern bobwhite. The comprehensive metadata attributes helped to identify RT-qPCR validated impacts.

## Sequence Assembly and CAPRG

Most of the transcriptome projects for non-model organisms focus on maximizing the number of genes found often termed as "gene hunting" and minimizing the redundant contigs (Papanicolaou *et al*., 2009). To broadly address above, the researcher first broadly classified assembly method as follows, PAVE, VELVET, MIRA using de novo strategies like OLC and graphs and CAPRG that uses reference genome to build initial clusters. The performance comparisons were done for assemblies generated from these methods. Assembler parameterization has been shown as important step in determining the output of ESTs project (Papanicolaou *et al*., 2009). The output of graph based method like Velvet is highly dependent on K-mer size while the OLC assemblers like CAP3 are affected by

identity percent. The parameter space for the assembling output with different methods into consideration is taken and compared with the output of different assemblers against CAPRG. The measurement of the assembly output can be done by the size and accuracy of their contigs (Miller *et al*., 2010). The assembly output is benchmarked based on first, the attributes of the assembling output like number of contigs and average length instead of N50 as N50 statistics for different assemblies are not comparable (Miller *et al*., 2010). Secondly, the annotation based on homology and more stringent reciprocal blast hit (RBH) is used and redundancy factor of the contigs generated by number of unique homologs is evaluated. Same input file for each individual organism and same cutoffs were used for BLAST for assemblies comparison to evaluate the performance of different assembly methods and parameter space.

One of the observations in transcriptome sequencing with NGS technologies is that full length transcripts are generally not sequenced though the transcripts are produced from whole mRNA (Papanicolaou *et al*., 2009). This might be due to failure of an assembler to provide sufficient evidence of alignment due to high number of mismatches specially due to non-coding regions non conservation, alternative splicing, multiple SNPs (Papanicolaou *et al*., 2009). As a result, the sequencing results in missing EST sequences and the EST's may not overlap to assemble to a contiguous sequence giving rise to non-overlapping contigs and singletons or splits in gene (Potter *et al*., 2004; Rawat *et al*., 2010b). This issue of fragmentation not only leads to partial representation of a protein coding sequence but also redundancy in the assembled sequences where many of the contigs might actually represent the same protein. Therefore this results in redundancy factor that is introduced in assembly when different contigs originates from same locus (Papanicolaou *et al*., 2009).

The graph algorithms for assembly like VELVET uses K-mer and reads with high number of shared K-mer are considered to have sequence similarity (Miller *et al*., 2010). The computational cost is sufficiently reduced due to faster detection of shared K-mer as compared to all-against-all pair wise sequence alignment (Miller *et al*., 2010). However this approach leads to lower sensitivity and therefore leads to missing true overlaps. It is observed that CAPRG, MIRA and PAVE outperformed VELVET and closely competed with each other for highest number of genes detected for reference chicken proteome and non redundant database for Northern bobwhite (Figure 34) and Japanese quail (Figure 35).

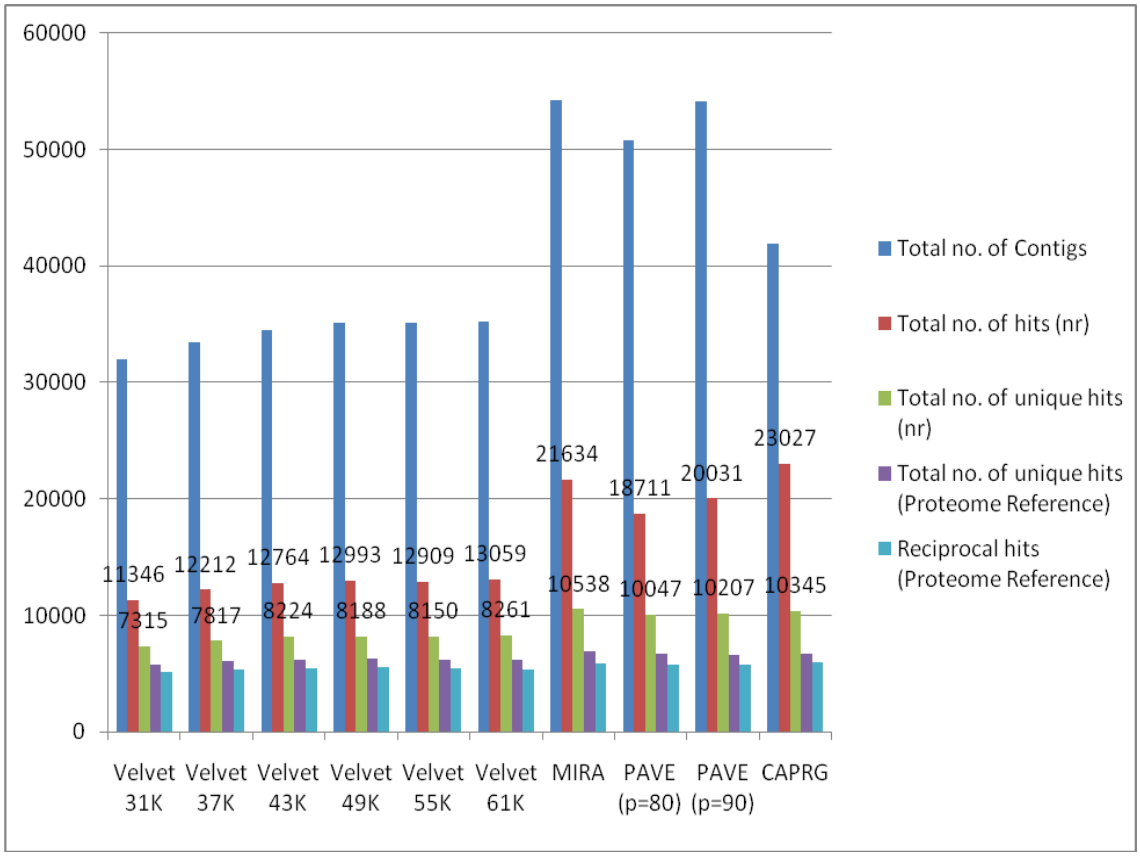


*Figure 17.* Assembling comparison for N. bobwhite with different assemblers and parameter space.

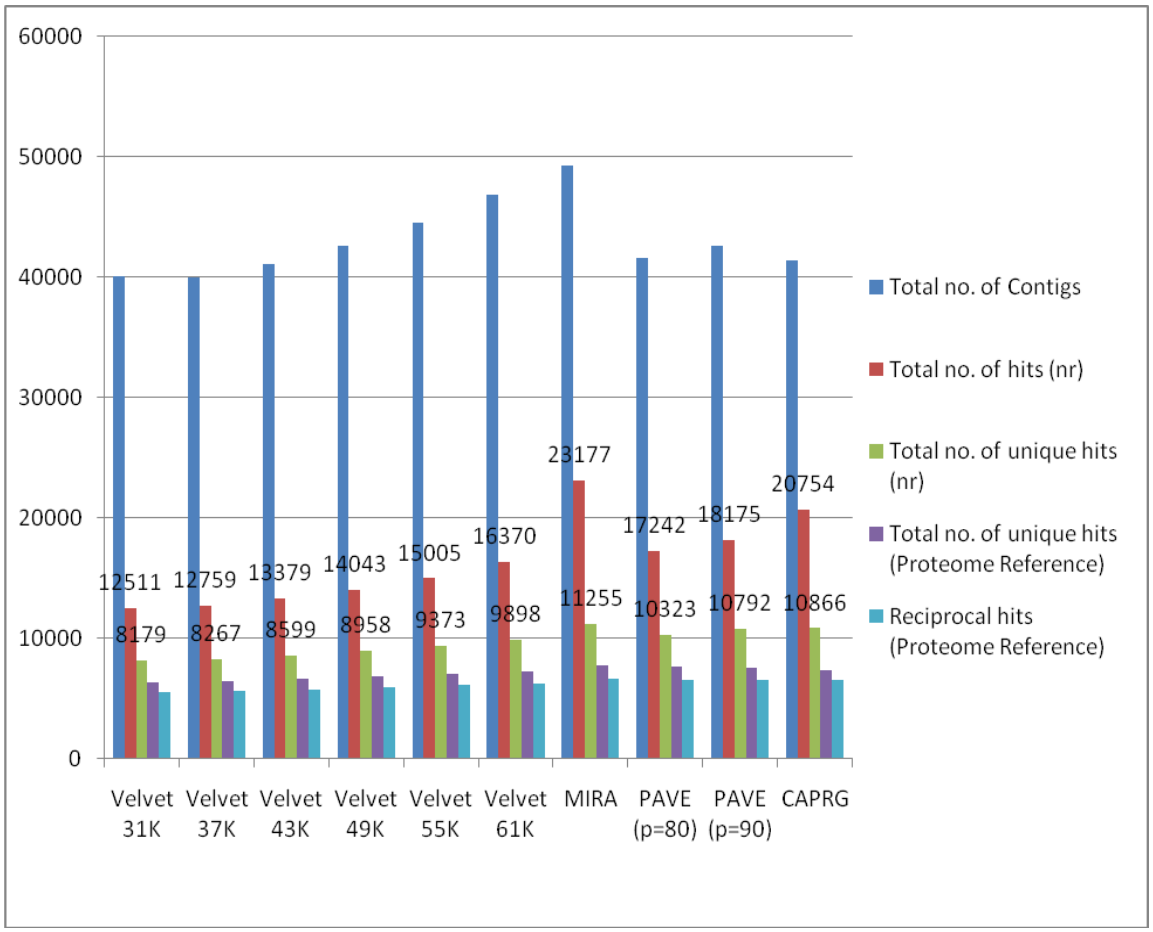


*Figure 18.* Assembling comparison for J. quail with different assemblers and parameter space.

The redundancy index can be seen in two different perspectives, the assembler is not sensitive to provide joins between reads leading to split in assembly or assembler is able to recognize putative regions in same locus/identify SNPs that are not identified by other assemblers. If the average length of contigs is not low as compared to other assembler, the assembler might actually be recognizing putative regions in same locus rather than disjoining to form shorter fragments. The redundancy index is calculated by dividing the total hits from non redundant database divided by total unique hits that will give number of contigs that belong to same locus for each organism. The redundancy index of MIRA and CAPRG was highest followed by PAVE. The average length of contigs for

MIRA, PAVE, CAPRG were nearly equal in Japanese quail, it was highest for MIRA followed by CAPRG for Northern bobwhite (Figure 36). Therefore MIRA and CAPRG performance in detecting putative regions might be higher in same locus without sacrificing the sensitivity to form join to form longer contigs.

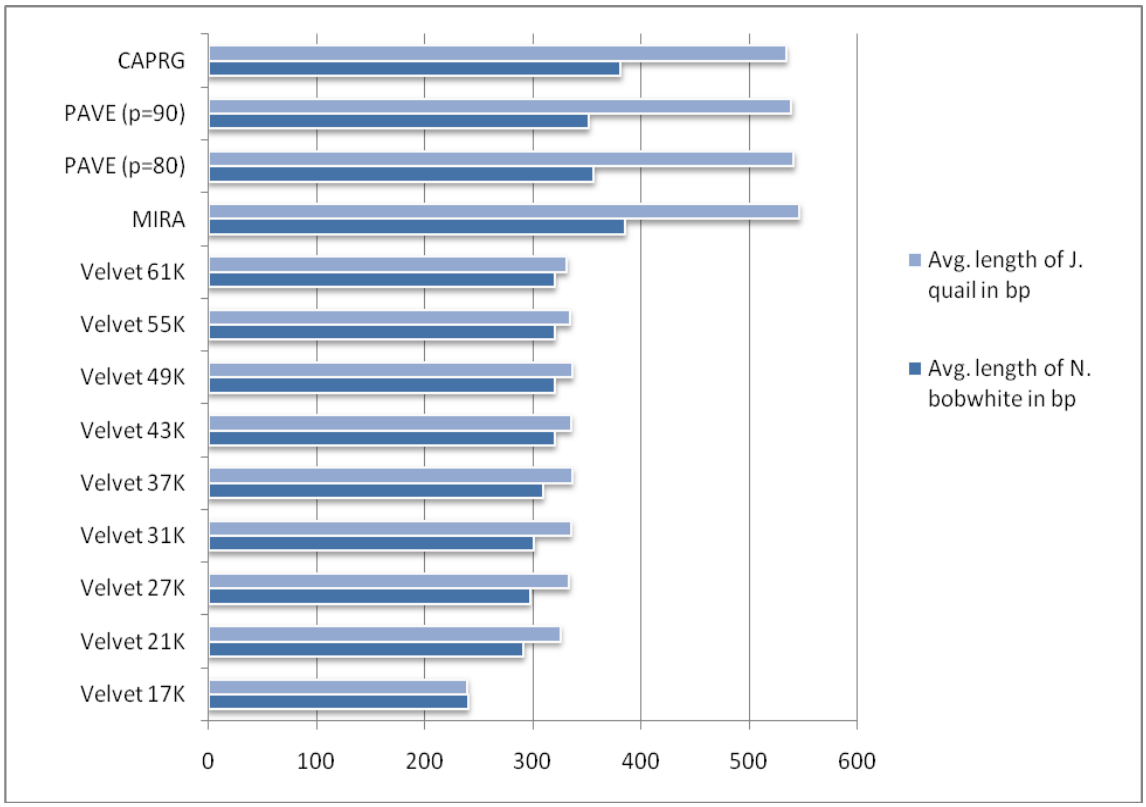


*Figure 19.* Comparison of average length of contigs of different assemblies.

The parameter space might be important in detecting higher number of putative genes as demonstrated in earlier work (Papanicolaou *et al*., 2009). However, with the two transcriptomics datasets, it is found that intra-assembling comparison with different parameters does not lead to higher number of diverse genes (Figure 37A).

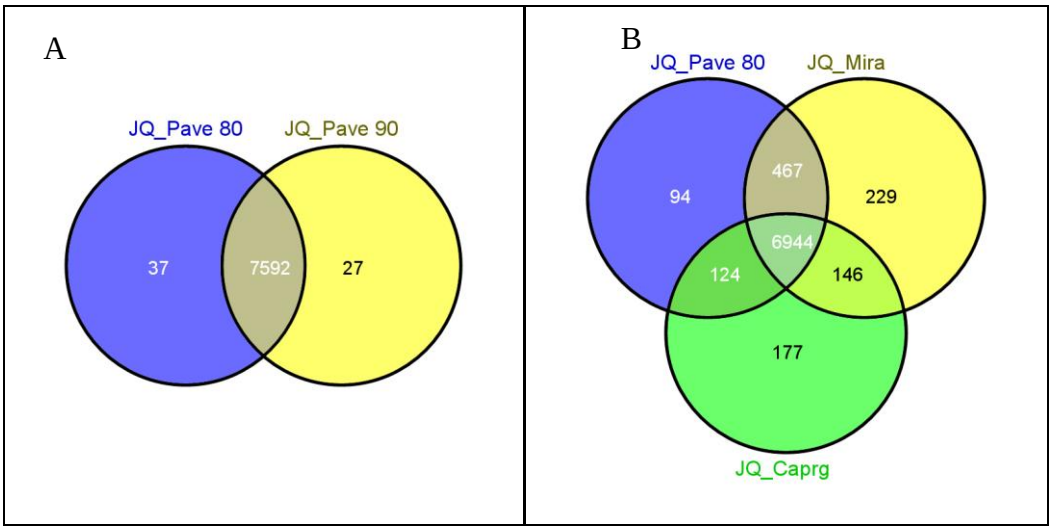


*Figure 20.* Comparison of overlapping protein coding of Japanese quail datasets against chicken proteome. (A) Intra-assembling parameterization of PAVE at identity 80% and 90%. (B) Inter-assembling comparison among PAVE, MIRA and CAPRG.

In fact, most of the coding regions detected by same assembly with different parameters overlapped significantly. However, when the total number of genes is compared that can be identified with different assembling methods, it is observed that high number of putative genes were identified uniquely by them (Figure 37B). Similar trends were seen for the Northern bobwhite data (Figure 38A, 38B).

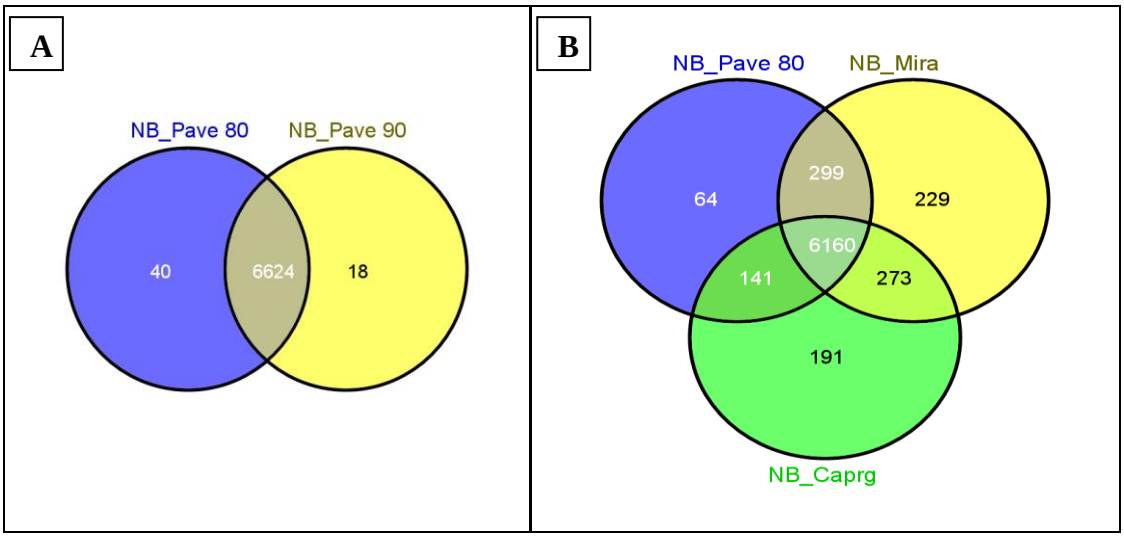


*Figure 21.* Comparison of overlapping protein coding of Northern Bobwhite datasets against chicken proteome. (A) Intra-assembling parameterization of PAVE at identity 80% and 90%. (B) Inter-assembling comparison among PAVE, MIRA and CAPRG.

Keeping these results in perspective, the assemblies generated from intra assembly parameterization might be useful. However, it would be more advantageous if two or more assembling methods are used for transcriptomics study that might help in generating higher number of putative genes, also discussed in earlier study (Papanicolaou *et al*., 2009).

The presence of repeats, relaxed assembling parameter can result in false positive joins that could result in chimeric contigs (Miller *et al*., 2010). One of the advantages of CAPRG is that fewer ESTs are exposed to other ESTs for alignment with limited window size strategy, as compared to all-against-all, pair wise and K-mer approach, leading to lesser chance of chimera for Roche/454. This also has an overall effect in the reduction of contigs inflation. The primary reason for contigs inflation can be attributed to non-coding DNA sequenced from multiple haplotypes that are heterozygous due to lesser selective constraint (Papanicolaou *et al*., 2009). The total number of contigs does not reflect higher number of genes sequenced as shown (Figure 34 and Figure 35). Overall the CAPRG produces lesser number of superfluous contigs and therefore completes assembling at

fraction of runtime as compared to other methods (Figure 39) except VELVET (that finishes in ~ 20 mins). It also puts lesser strain on the computing resources for further analysis with lesser contigs, without compromising sensitivity in detecting putative genes.

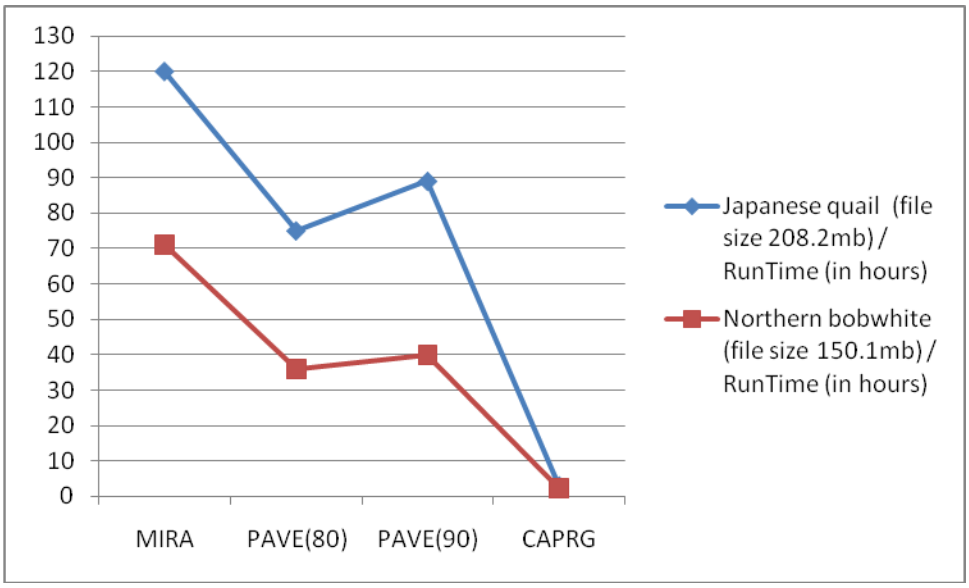


*Figure 22.* Runtime of assembling performed by MIRA, PAVE and CAPRG for Northern bobwhite and Japanese quail.

The overall quality of assembly can also be represented by the coverage of each contigs; the number of reads per contigs (Papanicolaou *et al*., 2009). The distribution of number of ESTs per contigs between PAVE that is established assembly pipeline and CAPRG is compared for Northern bobwhite (Table 7) and Japanese quail (Table 8).

Table 1

*The Distribution of Number of Reads Per Contigs for CAPRG, PAVE p=80, 90 for Northern Bobwhite*

| ESTs in Contig | Number of Contigs | | |
|---|---|---|---|
| | CAPRG | PAVE (P=80) | PAVE (p=90) |
| 2 | 14626 (34.89%) | 20839 (41.06%) | 21808 (40.26%) |
| 3-5 | 14799 (35.30%) | 16415 (32.34%) | 17737 (32.75%) |
| 6-10 | 6767 (16.14%) | 7253 (14.29%) | 7841 (14.48%) |
| 11-20 | 3524 (8.40%) | 3496 (6.89%) | 3909 (7.22%) |
| 21-50 | 1734 (4.13%) | 1808 (3.56%) | 2041 (3.77%) |
| 51-100 | 457 (1.09%) | 577 (1.14%) | 561 (1.04%) |
| >100 | 12 (0.03%) | 370 (0.73%) | 270 (.50%) |
| Total | 41919 (100%) | 50758 (100%) | 54167 (100%) |

Table 2

*The Distribution of Number of Reads Per Contigs for CAPRG, PAVE p=80, 90 for Japanese Quail*

| ESTs in Contig | Number of Contigs | | |
|---|---|---|---|
| | CAPRG | PAVE (P=80) | PAVE (p=90) |
| 2 | 20076 (48.14%) | 22668 (51.07%) | 23132 (50.81%) |
| 3-5 | 13724 (32.91%) | 13574 (30.58%) | 13945 (30.63%) |
| 6-10 | 4144 (9.94%) | 3829 (8.63%) | 3923 (8.62%) |
| 11-20 | 1984 (4.76%) | 1966 (4.43%) | 2092 (4.60%) |
| 21-50 | 1192 (2.86%) | 1335 (3.01%) | 1396 (3.07) |
| 51-100 | 555 (1.33%) | 536 (1.21%) | 573 (1.26%) |
| >100 | 57 (.14%) | 480 (1.08%) | 463 (1.02%) |
| Total | 41702 (100%) | 44388 (100%) | 45524 (100%) |

The distribution of number of ESTs per contigs can be considered as the coverage of the sequencing. The CAPRG produced higher percent of contigs in most of the categories representing higher coverage and sensitivity. It also seems to be tolerant to the shallow coverage ("two ESTs per contigs," the minimum requirement to form a contigs) that might have higher chances to form chimeric joins.

To further validate the new strategy adopted by CAPRG, it might be interesting to look at the percent of the ESTs that are binned in each window in expectation to form contigs, how many actually assembled to generate contigs is calculated. It is found that

high percentage 95.4% (206157/216073) and 97% (249745/257566) of reads that were identified/binned for each window successfully assembled as contigs. Finally the distribution of all the contigs that were binned across the expanse of chromosome can be visualized with this approach (Figure 40).

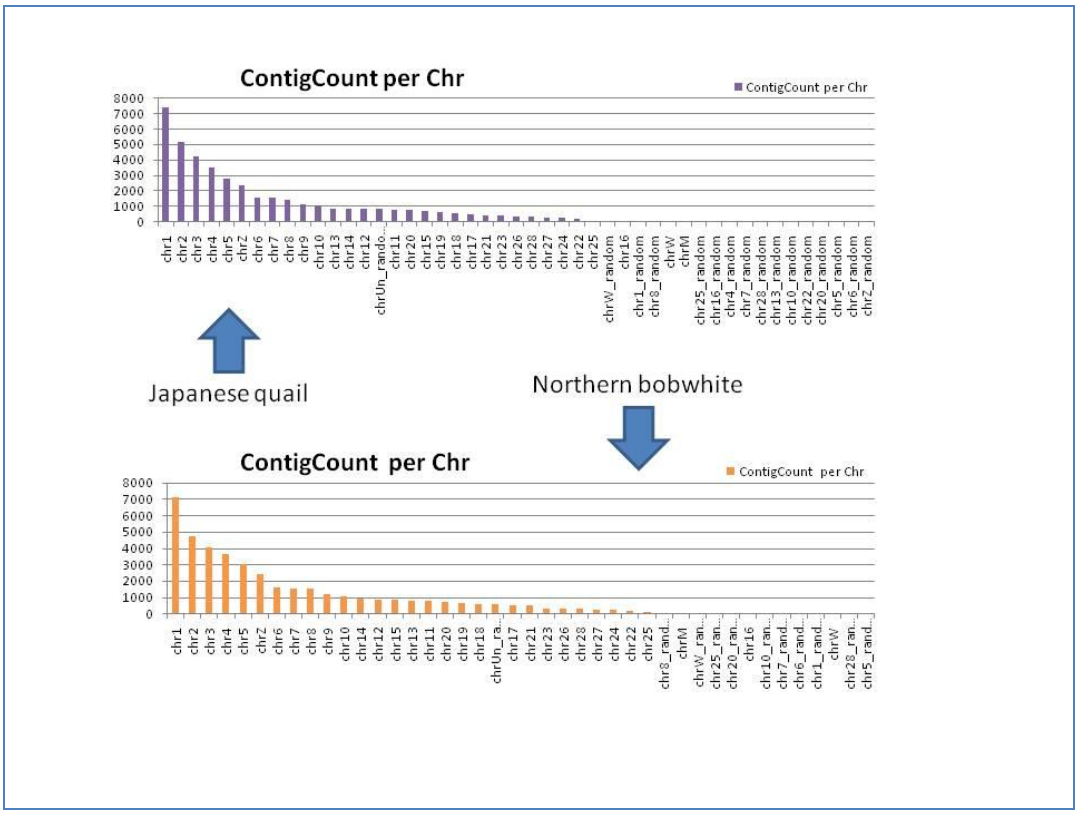


*Figure 23.* Distribution of the number of contigs per chromosome for Japanese quail and Northern bobwhite against chicken reference genome.

The distribution of both transcriptome against chicken genome had similar distribution with chromosome 1, chromosome 2, chromosome 3, chromosome 4, and chromosome 5 representing the categories with highest contigs. This information can be utilized to further analyze contigs that are considered "unknown" and are generally ignored in most studies besides giving a broader picture of the entire transcriptome of a non-model organism.

Genomic Comparisons between Avian Species

The comparison among toxicological model species, the Northern bobwhite and Japanese quail and passerine bird, the zebra finch to the MEC compounds 1,3,5-trinitro-1,3,5-triazacyclohexane (RDX) and 4-amino-2,6-dinitrotoluene (4A-DNT) of avian responses will be conducted. The goal of this research is to assess the similarity in responses of the avian species to RDX and 4A-DNT using clinical toxicological methods, analytical chemistry and genomic inquiry.

The sequences from TIGR gene index and Unigene were compared by searching each set against the NCBI non-redundant protein databases. The TIGR gene index resulted in 10,622/14,384 (74%) hits and Unigene in 5379/14,432 (37%) hits at $E \leq 10^{-5}$. Based on these results, the TIGR gene index for zebrafinch is selected for further genomic comparisons.

The overlap between the three species for comparison was done after comparing against the chicken reference proteome from Refseq (Figure 41). The overlap among the three species was substantial and microarray probes will be designed for zebrafinch.

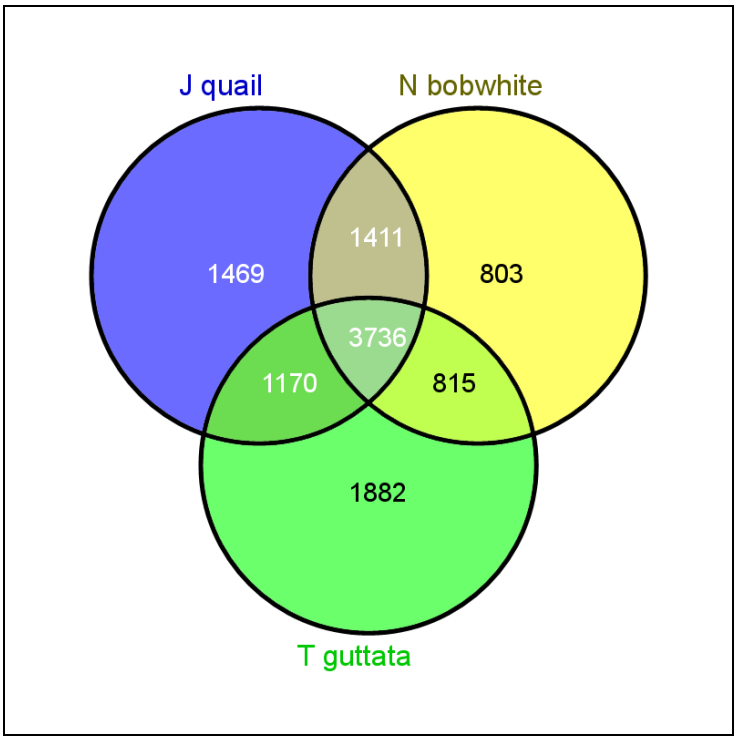


*Figure 24.* The genomic comparison between the three avian species. The total number of overlapped protein coding sequences between Japanese Quail, Northern Bobwhite and zebrafinch.

## Conclusions

In the project, the entire life cycle of development of an annotated microarray from sequencing, annotation is presented, and application to understanding sublethal impacts of 2,6-DNT dosing in Northern bobwhite. The comparative genome analyses with model organisms is performed and gene ontology annotation for 8,825 unique orthologous Northern bobwhite genes. Further, Northern bobwhite and Japanese quail is annotated along with high-throughput annotation and transitioned this knowledge to develop a custom microarray for Northern bobwhite. Microarray analysis of liver tissue from Northern bobwhite exposed to 2,6-DNT indicated a variety of molecular impacts resulting from 2,6-DNT exposures that were consistent with overt toxicological symptoms. Key molecular-to-phenotype impacts included: prostaglandin pathway-mediated inflammation,

increased expression of heme synthesis pathway in response to anemia, a shift in energy metabolism toward protein catabolism *via* inhibition of control points for glucose and lipid metabolic pathways, PCK1 and PPARGC1, respectively as well as a variety of additional insights. The conservation of genomic responses among species including parallel impacts of a 2,6-DNT isomer on gene expression observed in fathead minnow (Wintz *et al*., 2006) indicates the potential for commonality of mechanisms of action across distantly-related phylogenetic relatives. Overall, our observations not only illustrate the utility of the genomic infrastructure developed for Northern bobwhite for assessment of systemic perturbations, but also the potential for generalization of genomic responses among phyla.

The result of my effort is a web-accessible knowledgebase for Northern bobwhite which includes user friendly navigation tools and provides EST assembling information, sequence and structural properties and complex search utilities, bundled with an alternative method to generate sequence scaffolds to "stitch" transcripts against a reference genome.

A new pipeline "Contigs Assembly Pipeline using Reference Genome" (CAPRG) is built to assemble long reads for non model organisms that have available reference genome available for close phylogenetic neighbor. It is found that CAPRG performance is near equivalent or better in the two transcriptomic datasets based on different benchmarks but also completes the assembly in fraction of time as compared to assemblers that yield competitive results. Based on the study, it is observed that applying two or more assembling methods can help to enhance the putative genes coverage for any transcriptomics study.

## Future Work

The information from different stressors for Japanese quail and Northern bobwhite is curated and integrated in the knowledgebase. The insights can be developed describing association between stressors, gene/proteins and effects/diseases (Davis A.P. *et al*., 2008). Also, the use of cross-species comparative studies and different animal models has been found critical to understand the physiological mechanism and gene-protein functions (Mattingly *et al*., 2006). The cross-species comparative studies of these genes will help identify interactions between chemical and genes (Mattingly *et al*., 2006). Network inference is another challenging area where a Gaussian graphical model can be implemented to study transcriptional changes for the genes (Kiyosawa *et al*., 2010). Also, the enrichment of differentially expressed genes from datasets to discover gene signatures that will accurately identify and predict novel associations with tissue and stressors can be implemented (Patel and Butte, 2010).

CAPRG at present is a single threaded application, and the researcher plans to develop multi-threaded application that will result in considerable speedup. A web-based infrastructure that will allow scientists to quickly assemble the next generation sequencing data will assist in their research.

# APPENDIX A

## SUMMARY OF KEGG PATHWAY ANALYSIS FOR THE 60D, 2,6-DNT EXPOSURE IN NORTHERN BOBWHITE. TREATMENTS INCLUDED BIRDS EXPOSED TO 0, 10, AND 60 MG/KG/D WHERE GENOMIC ANALYSIS WAS CONDUCTED ON LIVER TISSUES.

| Dose | Target ID | p-Value | Fold Change | Fold Change (Log2) | Gene name | KEGG_PATHWAY |
|---|---|---|---|---|---|---|
| Liver_60 | Contig10068 | 0.0002 | -17.52 | -4.13 | LOC420337 | gga00190:Oxidative phosphorylation |
| Liver_60 | Contig328 | 0.0049 | -12.00 | -3.58 | PCK1 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | Contig328 | 0.0049 | -12.00 | -3.58 | PCK1 | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | Contig328 | 0.0049 | -12.00 | -3.58 | PCK1 | gga00620:Pyruvate metabolism |
| Liver_60 | Contig2_8467 | 0.0006 | -10.29 | -3.36 | LOC416852 | gga00350:Tyrosine metabolism |
| Liver_60 | Contig2_8467 | 0.0006 | -10.29 | -3.36 | LOC416852 | gga00360:Phenylalanine metabolism |
| Liver_60 | Contig2_15500 | 0.0000 | -8.63 | -3.11 | LOC427391 | gga00272:Cysteine metabolism |
| Liver_60 | Contig2_15500 | 0.0000 | -8.63 | -3.11 | LOC427391 | gga00430:Taurine and hypotaurine metabolism |
| Liver_60 | Contig20398 | 0.0013 | -8.36 | -3.06 | MAT1A | gga00271:Methionine metabolism |
| Liver_60 | Contig20398 | 0.0013 | -8.36 | -3.06 | MAT1A | gga00450:Selenoamino acid metabolism |
| Liver_10 | Contig10068 | 0.0005 | -7.18 | -2.84 | LOC420337 | gga00190:Oxidative phosphorylation |
| Liver_10 | Contig20371 | 0.0012 | -6.56 | -2.71 | PCK1 | gga00020:Citrate cycle (TCA cycle) |
| Liver_10 | Contig20371 | 0.0012 | -6.56 | -2.71 | PCK1 | gga00620:Pyruvate metabolism |
| Liver_60 | Contig20590 | 0.0001 | -5.05 | -2.34 | ACSL1 | gga00071:Fatty acid metabolism |
| Liver_60 | EQOT2SL01BV5XD_R | 0.0001 | -4.92 | -2.30 | LOC424011 | gga02010:ABC transporters - General |
| Liver_60 | Contig4524 | 0.0000 | -4.88 | -2.29 | LOC416728 | gga00630:Glyoxylate and dicarboxylate metabolism |
| Liver_60 | Contig2507 | 0.0007 | -4.03 | -2.01 | ULK2 | gga04140:Regulation of autophagy, |
| Liver_60 | Contig18214 | 0.0037 | -3.84 | -1.94 | LOC423081 | gga03022:Basal transcription factors |
| Liver_10 | Contig20398 | 0.0064 | -3.78 | -1.92 | MAT1A | gga00271:Methionine metabolism |
| Liver_10 | Contig20398 | 0.0064 | -3.78 | -1.92 | MAT1A | gga00450:Selenoamino acid metabolism |
| Liver_60 | Contig13806_R | 0.0024 | -3.78 | -1.92 | CDC16 | gga04120:Ubiquitin mediated proteolysis, |
| Liver_10 | Contig15317_R | 0.0000 | -3.66 | -1.87 | GALNT17 | gga00512:O-Glycan biosynthesis |
| Liver_10 | Contig15317_R | 0.0000 | -3.66 | -1.87 | GALNT17 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_60 | EQOT2SL01B0F5U | 0.0071 | -3.55 | -1.83 | FASN | gga00061:Fatty acid biosynthesis |
| Liver_60 | EQOT2SL01B0F5U | 0.0071 | -3.55 | -1.83 | FASN | gga01040:Polyunsaturated fatty acid biosynthesis |
| Liver_60 | EQOT2SL02HCJSM | 0.0055 | -3.48 | -1.80 | NT5C2 | gga00230:Purine metabolism |
| Liver_60 | EQOT2SL02HCJSM | 0.0055 | -3.48 | -1.80 | NT5C2 | gga00240:Pyrimidine metabolism |
| Liver_60 | EQOT2SL02HCJSM | 0.0055 | -3.48 | -1.80 | NT5C2 | gga00760:Nicotinate and nicotinamide metabolism |
| Liver_60 | EQOT2SL01BWOS1 | 0.0003 | -3.40 | -1.76 | ACACA | gga00061:Fatty acid biosynthesis |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | EQOT2SL01BWOS1 | 0.0003 | -3.40 | -1.76 | ACACA | gga00620:Pyruvate metabolism |
| Liver_60 | EQOT2SL01BWOS1 | 0.0003 | -3.40 | -1.76 | ACACA | gga00640:Propanoate metabolism |
| Liver_60 | Contig4390 | 0.0020 | -3.30 | -1.72 | PAH | gga00400:Phenylalanine |
| Liver_10 | Contig10636 | 0.0073 | -3.20 | -1.68 | COL11A1 | gga01430:Cell Communication, |
| Liver_60 | Contig2766 | 0.0080 | -3.16 | -1.66 | LOC423767 | gga02010:ABC transporters - General |
| Liver_60 | ERPU0F301EFF8G | 0.0006 | -3.15 | -1.65 | DPYD | gga00240:Pyrimidine metabolism |
| Liver_60 | ERPU0F301EFF8G | 0.0006 | -3.15 | -1.65 | DPYD | gga00410:beta-Alanine metabolism |
| Liver_60 | ERPU0F301EFF8G | 0.0006 | -3.15 | -1.65 | DPYD | gga00770:Pantothenate and CoA biosynthesis |
| Liver_10 | EQOT2SL02HCJSM | 0.0075 | -3.14 | -1.65 | NT5C2 | gga00230:Purine metabolism |
| Liver_10 | EQOT2SL02HCJSM | 0.0075 | -3.14 | -1.65 | NT5C2 | gga00240:Pyrimidine metabolism |
| Liver_10 | EQOT2SL02HCJSM | 0.0075 | -3.14 | -1.65 | NT5C2 | gga00760:Nicotinate and nicotinamide metabolism |
| Liver_60 | Contig8394 | 0.0055 | -3.09 | -1.63 | XDH | gga00230:Purine metabolism |
| Liver_60 | Contig8394 | 0.0055 | -3.09 | -1.63 | XDH | gga00232:Caffeine metabolism |
| Liver_60 | ERPU0F302F2V6B | 0.0052 | -3.07 | -1.62 | LOC421577 | gga04120:Ubiquitin mediated proteolysis, |
| Liver_60 | ERPU0F302FO7U7 | 0.0004 | -3.01 | -1.59 | ACLY | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | ERPU0F302FO7U7 | 0.0004 | -3.01 | -1.59 | ACLY | gga00720:Reductive carboxylate cycle (CO2 fixation) |
| Liver_60 | Contig2_21624 | 0.0002 | -2.94 | -1.56 | LOC424135 | gga00565:Ether lipid metabolism |
| Liver_60 | Contig2_24452 | 0.0004 | -2.92 | -1.54 | LOC418775 | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | EQOT2SL02I8A79_R | 0.0003 | -2.91 | -1.54 | ITCH | gga04120:Ubiquitin mediated proteolysis, |
| Liver_60 | Contig1_9727_R | 0.0082 | -2.85 | -1.51 | LOC423679 | gga00600:Sphingolipid metabolism |
| Liver_10 | ERPU0F301DSDKY | 0.0000 | -2.82 | -1.50 | NDST1 | gga00534:Heparan sulfate biosynthesis |
| Liver_10 | ERPU0F301DSDKY | 0.0000 | -2.82 | -1.50 | NDST1 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_60 | EQOT2SL02FZF9T_R | 0.0010 | -2.81 | -1.49 | ITPR1 | gga04070:Phosphatidylinositol signaling system, |
| Liver_60 | ERPU0F302HAWFG | 0.0006 | -2.80 | -1.48 | LOC427964 | gga00500:Starch and sucrose metabolism |
| Liver_60 | Contig20388 | 0.0089 | -2.76 | -1.47 | LOC416425 | gga00564:Glycerophospholipid metabolism |
| Liver_60 | Contig20388 | 0.0089 | -2.76 | -1.47 | LOC416425 | gga00565:Ether lipid metabolism |
| Liver_60 | Contig20388 | 0.0089 | -2.76 | -1.47 | LOC416425 | gga00590:Arachidonic acid metabolism |
| Liver_60 | Contig20388 | 0.0089 | -2.76 | -1.47 | LOC416425 | gga00591:Linoleic acid metabolism |
| Liver_60 | Contig20388 | 0.0089 | -2.76 | -1.47 | LOC416425 | gga00592:alpha-Linolenic acid metabolism |
| Liver_10 | ERPU0F302FT003 | 0.0001 | -2.71 | -1.44 | LOC415993 | gga00260:Glycine |
| Liver_10 | ERPU0F302FT003 | 0.0001 | -2.71 | -1.44 | LOC415993 | serine and threonine metabolism |
| Liver_60 | Contig2_476 | 0.0018 | -2.67 | -1.42 | SULT1C | gga00272:Cysteine metabolism |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00051:Fructose and mannose metabolism |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00052:Galactose metabolism |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00500:Starch and sucrose metabolism |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00521:Streptomycin biosynthesis |
| Liver_60 | EQOT2SL01BB8IU | 0.0031 | -2.67 | -1.41 | HK1 | gga00530:Aminosugars metabolism |
| Liver_60 | Contig2725 | 0.0006 | -2.66 | -1.41 | PRPS2 | gga00030:Pentose phosphate pathway |
| Liver_60 | Contig2725 | 0.0006 | -2.66 | -1.41 | PRPS2 | gga00230:Purine metabolism |
| Liver_60 | Contig7800 | 0.0075 | -2.61 | -1.38 | LOC423618 | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | Contig7800 | 0.0075 | -2.61 | -1.38 | LOC423618 | gga00310:Lysine degradation |
| Liver_60 | Contig7800 | 0.0075 | -2.61 | -1.38 | LOC423618 | gga00380:Tryptophan metabolism |
| Liver_60 | EQOT2SL02JAURZ_R | 0.0042 | -2.60 | -1.38 | LOC421790 | gga00500:Starch and sucrose metabolism |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | EQOT2SL02JAURZ_R | 0.0042 | -2.60 | -1.38 | LOC421790 | gga00790:Folate biosynthesis |
| Liver_10 | Contig2507 | 0.0059 | -2.53 | -1.34 | ULK2 | gga04140:Regulation of autophagy, |
| Liver_10 | Contig2_24 | 0.0080 | -2.52 | -1.34 | DCTD | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig2594 | 0.0006 | -2.46 | -1.30 | HNMT | gga00340:Histidine metabolism |
| Liver_60 | Contig11188 | 0.0032 | -2.45 | -1.29 | HMGCR | gga00100:Biosynthesis of steroids |
| Liver_60 | Contig2042 | 0.0031 | -2.44 | -1.29 | LOC420534 | gga02010:ABC transporters - General |
| Liver_10 | Contig1_17222_R | 0.0099 | -2.43 | -1.28 | GLS | gga00251:Glutamate metabolism |
| Liver_10 | Contig1_17222_R | 0.0099 | -2.43 | -1.28 | GLS | gga00471:D-Glutamine and D-glutamate metabolism |
| Liver_10 | Contig1_17222_R | 0.0099 | -2.43 | -1.28 | GLS | gga00910:Nitrogen metabolism |
| Liver_10 | ERPU0F302HZMC8 | 0.0007 | -2.40 | -1.26 | MID1 | gga04120:Ubiquitin mediated proteolysis, |
| Liver_60 | Contig4040 | 0.0019 | -2.39 | -1.26 | UPB1 | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig4040 | 0.0019 | -2.39 | -1.26 | UPB1 | gga00410:beta-Alanine metabolism |
| Liver_60 | Contig4040 | 0.0019 | -2.39 | -1.26 | UPB1 | gga00770:Pantothenate and CoA biosynthesis |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00620:Pyruvate metabolism |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00630:Glyoxylate and dicarboxylate metabolism |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00710:Carbon fixation |
| Liver_60 | Contig80 | 0.0070 | -2.33 | -1.22 | MDH1 | gga00720:Reductive carboxylate cycle (CO2 fixation) |
| Liver_60 | Contig21856_R | 0.0023 | -2.31 | -1.21 | P4HA1 | gga00330:Arginine and proline metabolism |
| Liver_60 | EQOT2SL01CP1G7 | 0.0028 | -2.29 | -1.19 | AADAC | gga00650:Butanoate metabolism |
| Liver_60 | EQOT2SL01CP1G7 | 0.0028 | -2.29 | -1.19 | AADAC | gga00960:Alkaloid biosynthesis II |
| Liver_60 | EQOT2SL01BAAGD | 0.0050 | -2.29 | -1.19 | LOC415805 | gga00562:Inositol phosphate metabolism |
| Liver_60 | EQOT2SL01BAAGD | 0.0050 | -2.29 | -1.19 | LOC415805 | gga04070:Phosphatidylinositol signaling system |
| Liver_60 | EQOT2SL01BAAGD | 0.0050 | -2.29 | -1.19 | LOC415805 | gga04662:B cell receptor signaling pathway |
| Liver_10 | Contig2_1181 | 0.0016 | -2.23 | -1.15 | LOC428622 | gga03030:DNA polymerase, |
| Liver_60 | Contig7826_R | 0.0086 | -2.22 | -1.15 | LOC424810 | gga00071:Fatty acid metabolism |
| Liver_60 | Contig9902_R | 0.0093 | -2.19 | -1.13 | LOC418728 | gga00500:Starch and sucrose metabolism |
| Liver_60 | Contig9902_R | 0.0093 | -2.19 | -1.13 | LOC418728 | gga00520:Nucleotide sugars metabolism |
| Liver_60 | Contig21413 | 0.0078 | -2.17 | -1.12 | GATM | gga00220:Urea cycle and metabolism of amino groups |
| Liver_60 | Contig21413 | 0.0078 | -2.17 | -1.12 | GATM | gga00260:Glycine |
| Liver_60 | Contig21413 | 0.0078 | -2.17 | -1.12 | GATM | gga00330:Arginine and proline metabolism |
| Liver_60 | Contig21413 | 0.0078 | -2.17 | -1.12 | GATM | serine and threonine metabolism |
| Liver_10 | Contig7826_R | 0.0028 | -2.13 | -1.09 | LOC424810 | gga00071:Fatty acid metabolism |
| Liver_60 | Contig2086 | 0.0035 | -2.11 | -1.08 | MCEE | gga00280:Valine |
| Liver_60 | Contig2086 | 0.0035 | -2.11 | -1.08 | MCEE | gga00640:Propanoate metabolism |
| Liver_60 | Contig3728 | 0.0058 | -2.10 | -1.07 | GPLD1 | gga00563:Glycosylphosphatidylinositol(GPI)-anchor biosynthesis |
| Liver_10 | Contig21856_R | 0.0039 | -2.10 | -1.07 | P4HA1 | gga00330:Arginine and proline metabolism |
| Liver_60 | Contig3784_R | 0.0072 | -2.09 | -1.07 | HERC1 | gga04120:Ubiquitin mediated proteolysis, |
| Liver_60 | Contig622 | 0.0017 | -2.09 | -1.06 | LOC427257 | gga00760:Nicotinate and nicotinamide metabolism |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | Contig22540 | 0.0035 | -2.08 | -1.06 | LOC423122 | gga00592:alpha-Linolenic acid metabolism |
| Liver_60 | Contig22540 | 0.0035 | -2.08 | -1.06 | LOC423122 | gga01040:Polyunsaturated fatty acid biosynthesis |
| Liver_60 | Contig2_12296 | 0.0048 | -2.07 | -1.05 | LOC419158 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | Contig2_12296 | 0.0048 | -2.07 | -1.05 | LOC419158 | gga00620:Pyruvate metabolism |
| Liver_60 | Contig2_12296 | 0.0048 | -2.07 | -1.05 | LOC419158 | gga00640:Propanoate metabolism |
| Liver_60 | Contig2_12296 | 0.0048 | -2.07 | -1.05 | LOC419158 | gga00720:Reductive carboxylate cycle (CO2 fixation) |
| Liver_10 | Contig7267_R | 0.0067 | -2.05 | -1.04 | LOC423632 | gga00532:Chondroitin sulfate biosynthesis |
| Liver_10 | Contig7267_R | 0.0067 | -2.05 | -1.04 | LOC423632 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_10 | Contig11284 | 0.0043 | -2.01 | -1.01 | XDH | gga00230:Purine metabolism |
| Liver_10 | Contig11284 | 0.0043 | -2.01 | -1.01 | XDH | gga00232:Caffeine metabolism |
| Liver_60 | Contig1247 | 0.0049 | -1.98 | -0.99 | ST6GAL1 | gga00510:N-Glycan biosynthesis |
| Liver_60 | Contig1247 | 0.0049 | -1.98 | -0.99 | ST6GAL1 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_60 | Contig8643 | 0.0024 | -1.97 | -0.98 | NDUFS1 | gga00190:Oxidative phosphorylation |
| Liver_60 | Contig5995 | 0.0081 | -1.96 | -0.97 | LOC424302 | gga00380:Tryptophan metabolism |
| Liver_60 | Contig7900 | 0.0033 | -1.91 | -0.93 | UGDH | gga00040:Pentose and glucuronate interconversions |
| Liver_60 | Contig7900 | 0.0033 | -1.91 | -0.93 | UGDH | gga00053:Ascorbate and aldarate metabolism |
| Liver_60 | Contig7900 | 0.0033 | -1.91 | -0.93 | UGDH | gga00500:Starch and sucrose metabolism |
| Liver_60 | Contig7900 | 0.0033 | -1.91 | -0.93 | UGDH | gga00520:Nucleotide sugars metabolism |
| Liver_60 | Contig11769 | 0.0059 | -1.87 | -0.91 | BCKDHB | gga00280:Valine |
| Liver_60 | Contig11769 | 0.0059 | -1.87 | -0.91 | BCKDHB | leucine and isoleucine degradation |
| Liver_60 | Contig15171 | 0.0084 | -1.87 | -0.90 | LOC422982 | gga00100:Biosynthesis of steroids |
| Liver_10 | Contig14261 | 0.0054 | -1.78 | -0.83 | ACACA | gga00061:Fatty acid biosynthesis |
| Liver_10 | Contig14261 | 0.0054 | -1.78 | -0.83 | ACACA | gga00620:Pyruvate metabolism |
| Liver_10 | Contig14261 | 0.0054 | -1.78 | -0.83 | ACACA | gga00640:Propanoate metabolism |
| Liver_10 | Contig20582_R | 0.0042 | -1.75 | -0.81 | BDH1 | gga00072:Synthesis and degradation of ketone bodies |
| Liver_10 | Contig20582_R | 0.0042 | -1.75 | -0.81 | BDH1 | gga00650:Butanoate metabolism |
| Liver 60 | Contig6022_R | 0.0177 | -1.74 | -0.80 | GOT1 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | Contig9824 | 0.0083 | -1.73 | -0.79 | HMGCL | gga00072:Synthesis and degradation of ketone bodies |
| Liver_60 | Contig9824 | 0.0083 | -1.73 | -0.79 | HMGCL | gga00280:Valine |
| Liver_60 | Contig9824 | 0.0083 | -1.73 | -0.79 | HMGCL | gga00650:Butanoate metabolism |
| Liver_60 | Contig9824 | 0.0083 | -1.73 | -0.79 | HMGCL | leucine and isoleucine degradation |
| Liver_60 | Contig2322 | 0.0054 | -1.70 | -0.77 | PANK1 | gga00770:Pantothenate and CoA biosynthesis |
| Liver_10 | Contig7900 | 0.0077 | -1.69 | -0.76 | UGDH | gga00040:Pentose and glucuronate interconversions |
| Liver_10 | Contig7900 | 0.0077 | -1.69 | -0.76 | UGDH | gga00053:Ascorbate and aldarate metabolism |
| Liver_10 | Contig7900 | 0.0077 | -1.69 | -0.76 | UGDH | gga00500:Starch and sucrose metabolism |
| Liver_10 | Contig7900 | 0.0077 | -1.69 | -0.76 | UGDH | gga00520:Nucleotide sugars metabolism |
| Liver_60 | Contig7208 | 0.0021 | -1.63 | -0.71 | GLUL | gga00251:Glutamate metabolism |
| Liver_60 | Contig7208 | 0.0021 | -1.63 | -0.71 | GLUL | gga00550:Peptidoglycan biosynthesis |
| Liver_60 | Contig7208 | 0.0021 | -1.63 | -0.71 | GLUL | gga00910:Nitrogen metabolism |
| Liver_60 | Contig5460 | 0.0051 | -1.58 | -0.66 | HAL | gga00340:Histidine metabolism |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | Contig5460 | 0.0051 | -1.58 | -0.66 | HAL | gga00910:Nitrogen metabolism |
| Liver_60 | Contig1327_R | 0.0077 | 1.79 | 0.84 | TAF12 | gga03022:Basal transcription factors |
| Liver_10 | ERPU0F302F04OE | 0.0062 | 1.81 | 0.85 | HAAO | gga00380:Tryptophan metabolism |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00350:Tyrosine metabolism |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00360:Phenylalanine metabolism |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00564:Glycerophospholipid metabolism |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00624:1- and 2-Methylnaphthalene degradation |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00632:Benzoate degradation via CoA ligation |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00903:Limonene and pinene degradation |
| Liver_10 | Contig5830 | 0.0064 | 1.81 | 0.86 | SH3GLB1 | gga00960:Alkaloid biosynthesis II |
| Liver_60 | Contig4232_R | 0.0064 | 1.87 | 0.90 | LOC416730 | gga00562:Inositol phosphate metabolism |
| Liver_60 | Contig4232_R | 0.0064 | 1.87 | 0.90 | LOC416730 | gga04070:Phosphatidylinositol signaling system |
| Liver_60 | Contig1522_R | 0.0058 | 1.91 | 0.93 | IARS | gga00290:Valine |
| Liver_60 | Contig1522_R | 0.0058 | 1.91 | 0.93 | IARS | gga00970:Aminoacyl-tRNA biosynthesis |
| Liver_60 | Contig1522_R | 0.0058 | 1.91 | 0.93 | IARS | leucine and isoleucine biosynthesis |
| Liver_60 | Contig14222 | 0.0075 | 1.91 | 0.94 | GCLM | gga00251:Glutamate metabolism |
| Liver_60 | Contig14222 | 0.0075 | 1.91 | 0.94 | GCLM | gga00480:Glutathione metabolism |
| Liver_60 | Contig7072 | 0.0009 | 1.93 | 0.95 | RRAS2 | gga04660:T cell receptor signaling pathway |
| Liver_60 | Contig7072 | 0.0009 | 1.93 | 0.95 | RRAS2 | gga04662:B cell receptor signaling pathway |
| Liver_10 | Contig14824 | 0.0047 | 1.94 | 0.96 | PMM2 | gga00051:Fructose and mannose metabolism |
| Liver_10 | EQOT2SL02IEBET | 0.0096 | 1.95 | 0.96 | UBE1C | gga04120:Ubiquitin mediated proteolysis, |
| Liver_10 | Contig17615 | 0.0068 | 1.95 | 0.97 | LOC429096 | gga04070:Phosphatidylinositol signaling system, |
| Liver_60 | Contig4655 | 0.0081 | 1.96 | 0.97 | LOC418131 | gga00400:Phenylalanine |
| Liver_60 | Contig1_10437 | 0.0061 | 1.97 | 0.98 | CPOX | gga00860:Porphyrin and chlorophyll metabolism |
| Liver_60 | Contig1_19531 | 0.0074 | 1.97 | 0.98 | LOC422022 | gga00534:Heparan sulfate biosynthesis |
| Liver_60 | Contig1_19531 | 0.0074 | 1.97 | 0.98 | LOC422022 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_10 | Contig13638 | 0.0069 | 1.99 | 0.99 | SCD | gga01040:Polyunsaturated fatty acid biosynthesis, |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00350:Tyrosine metabolism |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00360:Phenylalanine metabolism |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00564:Glycerophospholipid metabolism |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00624:1- and 2-Methylnaphthalene degradation |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00632:Benzoate degradation via CoA ligation |
| Liver_60 | Contig5830 | 0.0053 | 2.00 | 1.00 | SH3GLB1 | gga00903:Limonene and pinene degradation |
| Liver_10 | Contig21102 | 0.0044 | 2.03 | 1.02 | ADAL | gga00230:Purine metabolism |
| Liver_60 | Contig35 | 0.0061 | 2.04 | 1.03 | ATP12A | gga00190:Oxidative phosphorylation |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00150:Androgen and estrogen metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00340:Histidine metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00350:Tyrosine metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00380:Tryptophan metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00440:Aminophosphonate metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00450:Selenoamino acid metabolism |
| Liver_10 | Contig7804 | 0.0037 | 2.04 | 1.03 | METTL6 | gga00626:Naphthalene and anthracene degradation |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_10 | Contig1_19230 | 0.0056 | 2.07 | 1.05 | TMEM23 | gga00600:Sphingolipid metabolism, |
| Liver_10 | Contig1_13402_R | 0.0067 | 2.07 | 1.05 | LOC422364 | gga04120:Ubiquitin mediated proteolysis, |
| Liver_10 | EQOT2SL02JMWD0_R | 0.0024 | 2.07 | 1.05 | LOC422910 | gga00230:Purine metabolism |
| Liver_10 | EQOT2SL02JMWD0_R | 0.0024 | 2.07 | 1.05 | LOC422910 | gga00240:Pyrimidine metabolism |
| Liver_10 | EQOT2SL02JMWD0_R | 0.0024 | 2.07 | 1.05 | LOC422910 | gga03020:RNA polymerase |
| Liver_60 | Contig21511 | 0.0086 | 2.08 | 1.06 | POLR2B | gga00230:Purine metabolism |
| Liver_60 | Contig21511 | 0.0086 | 2.08 | 1.06 | POLR2B | gga00240:Pyrimidine metabolism |
| Liver_10 | Contig13034_R | 0.0019 | 2.10 | 1.07 | UBE3C | gga04120:Ubiquitin mediated proteolysis, |
| Liver_10 | Contig14610 | 0.0011 | 2.13 | 1.09 | LIPT1 | gga00785:Lipoic acid metabolism, |
| Liver_60 | Contig2_15107 | 0.0082 | 2.13 | 1.09 | BST1 | gga00760:Nicotinate and nicotinamide metabolism |
| Liver_60 | Contig20331 | 0.0097 | 2.14 | 1.10 | LDHA | gga00010:Glycolysis / Gluconeogenesis |
| Liver_60 | Contig20331 | 0.0097 | 2.14 | 1.10 | LDHA | gga00272:Cysteine metabolism |
| Liver_60 | Contig20331 | 0.0097 | 2.14 | 1.10 | LDHA | gga00620:Pyruvate metabolism |
| Liver_60 | Contig20331 | 0.0097 | 2.14 | 1.10 | LDHA | gga00640:Propanoate metabolism |
| Liver_10 | Contig2701_R | 0.0037 | 2.17 | 1.12 | LOC424915 | gga00440:Aminophosphonate metabolism |
| Liver_10 | Contig2701_R | 0.0037 | 2.17 | 1.12 | LOC424915 | gga00564:Glycerophospholipid metabolism |
| Liver_60 | Contig2_23366 | 0.0095 | 2.21 | 1.14 | CKB | gga00330:Arginine and proline metabolism |
| Liver_60 | Contig14610 | 0.0087 | 2.23 | 1.16 | LIPT1 | gga00785:Lipoic acid metabolism |
| Liver_60 | Contig15441 | 0.0060 | 2.28 | 1.19 | STX7 | gga04130:SNARE interactions in vesicular transport, |
| Liver_60 | Contig5990 | 0.0035 | 2.30 | 1.20 | LOC426223 | gga00860:Porphyrin and chlorophyll metabolism |
| Liver_10 | ERPU0F301EC27S | 0.0020 | 2.31 | 1.21 | MINPP1 | gga00562:Inositol phosphate metabolism |
| Liver_10 | Contig2_3201 | 0.0036 | 2.32 | 1.21 | DDX47 | gga00500:Starch and sucrose metabolism |
| Liver_10 | Contig2_3201 | 0.0036 | 2.32 | 1.21 | DDX47 | gga00790:Folate biosynthesis |
| Liver_60 | Contig1262 | 0.0035 | 2.37 | 1.24 | ABCB4 | gga02010:ABC transporters - General |
| Liver_60 | ERPU0F302HB9YP | 0.0043 | 2.37 | 1.25 | LOC431449 | gga00740:Riboflavin metabolism |
| Liver_60 | Contig2563 | 0.0027 | 2.38 | 1.25 | CTPS2 | gga00240:Pyrimidine metabolism |
| Liver_10 | Contig18875 | 0.0068 | 2.46 | 1.30 | LOC418654 | gga00150:Androgen and estrogen metabolism |
| Liver_10 | Contig3796_R | 0.0020 | 2.46 | 1.30 | AKR1B10 | gga00040:Pentose and glucuronate interconversions |
| Liver_10 | Contig3796_R | 0.0020 | 2.46 | 1.30 | AKR1B10 | gga00051:Fructose and mannose metabolism |
| Liver_10 | Contig3796_R | 0.0020 | 2.46 | 1.30 | AKR1B10 | gga00052:Galactose metabolism |
| Liver_10 | Contig3796_R | 0.0020 | 2.46 | 1.30 | AKR1B10 | gga00561:Glycerolipid metabolism |
| Liver_10 | Contig3796_R | 0.0020 | 2.46 | 1.30 | AKR1B10 | gga00620:Pyruvate metabolism |
| Liver_60 | Contig593 | 0.0040 | 2.56 | 1.35 | LOC424491 | gga01040:Polyunsaturated fatty acid biosynthesis |
| Liver_60 | Contig2_1789 | 0.0091 | 2.60 | 1.38 | LOC420253 | gga00230:Purine metabolism |
| Liver_60 | Contig2_1789 | 0.0091 | 2.60 | 1.38 | LOC420253 | gga00240:Pyrimidine metabolism |
| Liver_60 | EQOT2SL01CZKTL | 0.0070 | 2.63 | 1.39 | ATP6V1G3 | gga00190:Oxidative phosphorylation |
| Liver_60 | Contig10405_R | 0.0001 | 2.64 | 1.40 | LOC428769 | gga00534:Heparan sulfate biosynthesis |
| Liver_60 | Contig10405_R | 0.0001 | 2.64 | 1.40 | LOC428769 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_60 | Contig2701_R | 0.0048 | 2.65 | 1.41 | LOC424915 | gga00440:Aminophosphonate metabolism |
| Liver_60 | Contig2701_R | 0.0048 | 2.65 | 1.41 | LOC424915 | gga00564:Glycerophospholipid metabolism |
| Liver_60 | Contig2_9856 | 0.0025 | 2.66 | 1.41 | LOC421894 | gga00251:Glutamate metabolism |
| Liver_60 | Contig2_9856 | 0.0025 | 2.66 | 1.41 | LOC421894 | gga00480:Glutathione metabolism |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | Contig5710 | 0.0004 | 2.67 | 1.42 | NT5C3 | gga00230:Purine metabolism |
| Liver_60 | Contig5710 | 0.0004 | 2.67 | 1.42 | NT5C3 | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig5710 | 0.0004 | 2.67 | 1.42 | NT5C3 | gga00760:Nicotinate and nicotinamide metabolism |
| Liver_60 | Contig17615 | 0.0048 | 2.71 | 1.44 | LOC429096 | gga04070:Phosphatidylinositol signaling system, |
| Liver_60 | Contig2_11780 | 0.0001 | 2.75 | 1.46 | LOC418594 | gga00440:Aminophosphonate metabolism |
| Liver_60 | Contig2_11780 | 0.0001 | 2.75 | 1.46 | LOC418594 | gga00564:Glycerophospholipid metabolism |
| Liver_60 | ERPU0F301AFCX9 | 0.0045 | 2.76 | 1.46 | POLK | gga03030:DNA polymerase, |
| Liver_60 | Contig19621 | 0.0087 | 2.81 | 1.49 | GUCY1B3 | gga00230:Purine metabolism |
| Liver_60 | Contig19280 | 0.0070 | 2.85 | 1.51 | AGMAT | gga00220:Urea cycle and metabolism of amino groups |
| Liver_60 | Contig1_15107 | 0.0000 | 2.87 | 1.52 | GSTA | gga00480:Glutathione metabolism |
| Liver_60 | Contig1_15107 | 0.0000 | 2.87 | 1.52 | GSTA | gga00980:Metabolism of xenobiotics by cytochrome P450 |
| Liver_10 | Contig2_1789 | 0.0094 | 2.88 | 1.53 | LOC420253 | gga00230:Purine metabolism |
| Liver_10 | Contig2_1789 | 0.0094 | 2.88 | 1.53 | LOC420253 | gga00240:Pyrimidine metabolism |
| Liver_10 | Contig9128_R | 0.0004 | 2.91 | 1.54 | LOC420606 | gga02010:ABC transporters - General, |
| Liver_60 | Contig15317 | 0.0001 | 2.97 | 1.57 | GALNT17 | gga00512:O-Glycan biosynthesis |
| Liver_60 | Contig15317 | 0.0001 | 2.97 | 1.57 | GALNT17 | gga01030:Glycan structures - biosynthesis 1 |
| Liver_60 | Contig1_2893 | 0.0067 | 3.05 | 1.61 | COL6A2 | gga01430:Cell Communication |
| Liver_60 | Contig6166 | 0.0046 | 3.12 | 1.64 | LOC422551 | gga00565:Ether lipid metabolism |
| Liver_10 | EQOT2SL01C4BUX | 0.0002 | 3.34 | 1.74 | DCK | gga00230:Purine metabolism |
| Liver_10 | EQOT2SL01C4BUX | 0.0002 | 3.34 | 1.74 | DCK | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig22786 | 0.0053 | 3.45 | 1.79 | NT5C1B | gga00230:Purine metabolism |
| Liver_60 | Contig22786 | 0.0053 | 3.45 | 1.79 | NT5C1B | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig22786 | 0.0053 | 3.45 | 1.79 | NT5C1B | gga00760:Nicotinate and nicotinamide metabolism |
| Liver_60 | Contig21986 | 0.0037 | 3.45 | 1.79 | GJB6 | gga01430:Cell Communication |
| Liver_60 | Contig2_23244 | 0.0011 | 3.48 | 1.80 | ODC1 | gga00220:Urea cycle and metabolism of amino groups |
| Liver_60 | Contig3796_R | 0.0000 | 3.74 | 1.90 | AKR1B10 | gga00040:Pentose and glucuronate interconversions |
| Liver_60 | Contig3796_R | 0.0000 | 3.74 | 1.90 | AKR1B10 | gga00051:Fructose and mannose metabolism |
| Liver_60 | Contig3796_R | 0.0000 | 3.74 | 1.90 | AKR1B10 | gga00052:Galactose metabolism |
| Liver_60 | Contig3796_R | 0.0000 | 3.74 | 1.90 | AKR1B10 | gga00561:Glycerolipid metabolism |
| Liver_60 | Contig3796_R | 0.0000 | 3.74 | 1.90 | AKR1B10 | gga00620:Pyruvate metabolism |
| Liver_60 | EQOT2SL01C4BUX | 0.0010 | 3.90 | 1.97 | DCK | gga00230:Purine metabolism |
| Liver_60 | EQOT2SL01C4BUX | 0.0010 | 3.90 | 1.97 | DCK | gga00240:Pyrimidine metabolism |
| Liver_60 | Contig20411 | 0.0001 | 4.31 | 2.11 | ST3GAL6 | gga00602:Glycosphingolipid biosynthesis - neo-lactoseries |
| Liver_60 | Contig20411 | 0.0001 | 4.31 | 2.11 | ST3GAL6 | gga01031:Glycan structures - biosynthesis 2 |
| Liver_60 | Contig15395 | 0.0008 | 4.48 | 2.16 | TEC | gga04660:T cell receptor signaling pathway, |
| Liver_60 | EQOT2SL02HWRWH | 0.0084 | 4.58 | 2.19 | B3GALNT1 | gga00603:Glycosphingolipid biosynthesis - globoseries |
| Liver_60 | EQOT2SL02HWRWH | 0.0084 | 4.58 | 2.19 | B3GALNT1 | gga01031:Glycan structures - biosynthesis 2 |
| Liver_10 | ERPU0F301DVT0V | 0.0002 | 4.70 | 2.23 | LOC424288 | gga00380:Tryptophan metabolism |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver_60 | Contig4620_R | 0.0079 | 4.95 | 2.31 | CYP3A80 | gga00361:gamma-Hexachlorocyclohexane degradation |
| Liver_60 | Contig4620_R | 0.0079 | 4.95 | 2.31 | CYP3A80 | gga00591:Linoleic acid metabolism |
| Liver_60 | Contig1909 | 0.0004 | 5.00 | 2.32 | SUCLG2 | gga00020:Citrate cycle (TCA cycle) |
| Liver_60 | Contig1909 | 0.0004 | 5.00 | 2.32 | SUCLG2 | gga00640:Propanoate metabolism |
| Liver_10 | Contig21749 | 0.0000 | 5.07 | 2.34 | CYP7A1 | gga00120:Bile acid biosynthesis |
| Liver_60 | Contig7048 | 0.0025 | 5.15 | 2.36 | ATP6V0D2 | gga00190:Oxidative phosphorylation |
| Liver_60 | Contig16296 | 0.0018 | 5.28 | 2.40 | ICA1 | gga04940:Type I diabetes mellitus, |
| Liver_60 | Contig5886 | 0.0050 | 6.58 | 2.72 | ENO2 | gga00010:Glycolysis / Gluconeogenesis |
| Liver_10 | Contig1531 | 0.0000 | 7.34 | 2.88 | LOC420045 | gga01430:Cell Communication, |
| Liver_60 | Contig2_18846 | 0.0007 | 8.17 | 3.03 | LOC415619 | gga00600:Sphingolipid metabolism |
| Liver_60 | Contig7501 | 0.0078 | 9.25 | 3.21 | GUCY1A3 | gga00230:Purine metabolism |
| Liver_60 | Contig1531 | 0.0097 | 13.70 | 3.78 | LOC420045 | gga01430:Cell Communication |

# Appendix B

## SUMMARY OF GENE ONTOLOGY (GO) ANALYSIS FOR THE 60D, 2,6-DNT EXPOSURE IN NORTHERN BOBWHITE. TREATMENTS INCLUDED BIRDS EXPOSED TO 0, 10, AND 60 MG/KG/D WHERE GENOMIC ANALYSIS WAS CONDUCTED ON LIVER TISSUES.

| Treatment | Category | GO ID | GO Term | Count |
|---|---|---|---|---|
| Liver_60 | Molecular Function | GO:0008253 | 5'-nucleotidase activity | 3 |
| Liver_60 | Molecular Function | GO:0003779 | actin binding | 12 |
| Liver_10 | Molecular Function | GO:0030554 | adenyl nucleotide binding | 29 |
| Liver_60 | Cellular Component | GO:0005912 | adherens junction | 3 |
| Liver_60 | Biological Process | GO:0009310 | amine catabolic process | 4 |
| Liver_60 | Biological Process | GO:0009308 | amine metabolic process | 14 |
| Liver_60 | Biological Process | GO:0006519 | amino acid and derivative metabolic process | 13 |
| Liver_60 | Biological Process | GO:0009063 | amino acid catabolic process | 4 |
| Liver_60 | Biological Process | GO:0006520 | amino acid metabolic process | 11 |
| Liver_60 | Molecular Function | GO:0016841 | ammonia-lyase activity | 2 |
| Liver_10 | Biological Process | GO:0009653 | anatomical structure morphogenesis | 9 |
| Liver_60 | Biological Process | GO:0006725 | aromatic compound metabolic process | 6 |
| Liver_10 | Molecular Function | GO:0005524 | ATP binding | 27 |
| Liver_10 | Molecular Function | GO:0005488 | binding | 138 |
| Liver_60 | Molecular Function | GO:0005488 | binding | 274 |
| Liver_60 | Biological Process | GO:0009058 | biosynthetic process | 35 |
| Liver_60 | Biological Process | GO:0007596 | blood coagulation | 5 |
| Liver_10 | Biological Process | GO:0030282 | bone mineralization | 2 |
| Liver_60 | Molecular Function | GO:0005509 | calcium ion binding | 39 |
| Liver_60 | Molecular Function | GO:0005544 | calcium-dependent phospholipid binding | 4 |
| Liver_60 | Molecular Function | GO:0016840 | carbon-nitrogen lyase activity | 3 |
| Liver_10 | Biological Process | GO:0019752 | carboxylic acid metabolic process | 8 |
| Liver_60 | Biological Process | GO:0019752 | carboxylic acid metabolic process | 16 |
| Liver_10 | Molecular Function | GO:0003824 | catalytic activity | 86 |
| Liver_60 | Molecular Function | GO:0003824 | catalytic activity | 160 |
| Liver_60 | Biological Process | GO:0006812 | cation transport | 16 |
| Liver_10 | Biological Process | GO:0048468 | cell development | 9 |

| | | | | |
|---|---|---|---|---|
| Liver_10 | Biological Process | GO:0030154 | cell differentiation | 13 |
| Liver_10 | Cellular Component | GO:0000267 | cell fraction | 5 |
| Liver_10 | Biological Process | GO:0048869 | cellular developmental process | 13 |
| Liver_60 | Biological Process | GO:0044260 | cellular macromolecule metabolic process | 40 |
| Liver_10 | Biological Process | GO:0044237 | cellular metabolic process | 75 |
| Liver_10 | Biological Process | GO:0044267 | cellular protein metabolic process | 40 |
| Liver_60 | Biological Process | GO:0050817 | coagulation | 6 |
| Liver_60 | Molecular Function | GO:0050662 | coenzyme binding | 11 |
| Liver_60 | Biological Process | GO:0009108 | coenzyme biosynthetic process | 6 |
| Liver_60 | Biological Process | GO:0006732 | coenzyme metabolic process | 9 |
| Liver_60 | Molecular Function | GO:0048037 | cofactor binding | 15 |
| Liver_60 | Biological Process | GO:0051188 | cofactor biosynthetic process | 7 |
| Liver_60 | Biological Process | GO:0051186 | cofactor metabolic process | 10 |
| Liver_10 | Molecular Function | GO:0004197 | cysteine-type endopeptidase activity | 4 |
| Liver_10 | Cellular Component | GO:0005737 | cytoplasm | 39 |
| Liver_60 | Cellular Component | GO:0005737 | cytoplasm | 70 |
| Liver_10 | Cellular Component | GO:0044444 | cytoplasmic part | 23 |
| Liver_60 | Molecular Function | GO:0008092 | cytoskeletal protein binding | 14 |
| Liver_10 | Biological Process | GO:0032502 | developmental process | 21 |
| Liver_60 | Biological Process | GO:0006118 | electron transport | 17 |
| Liver_10 | Molecular Function | GO:0016251 | general RNA polymerase II transcription factor activity | 2 |
| Liver_60 | Biological Process | GO:0006091 | generation of precursor metabolites and energy | 22 |
| Liver_60 | Biological Process | GO:0007599 | hemostasis | 5 |
| Liver_60 | Molecular Function | GO:0043167 | ion binding | 98 |
| Liver_10 | Molecular Function | GO:0005506 | iron ion binding | 8 |
| Liver_60 | Molecular Function | GO:0055102 | lipase inhibitor activity | 3 |
| Liver_60 | Molecular Function | GO:0016829 | lyase activity | 12 |
| Liver_10 | Cellular Component | GO:0000119 | mediator complex | 2 |
| Liver_60 | Cellular Component | GO:0000119 | mediator complex | 2 |
| Liver_10 | Cellular Component | GO:0005624 | membrane fraction | 4 |
| Liver_10 | Cellular Component | GO:0031974 | membrane-enclosed lumen | 8 |
| Liver_10 | Biological Process | GO:0008152 | metabolic process | 91 |
| Liver_60 | Molecular Function | GO:0046872 | metal ion binding | 97 |
| Liver_60 | Biological Process | GO:0030001 | metal ion transport | 14 |

| | | | | |
|---|---|---|---|---|
| Liver_60 | Biological Process | GO:0015672 | monovalent inorganic cation transport | 12 |
| Liver_10 | Biological Process | GO:0002009 | morphogenesis of an epithelium | 3 |
| Liver_10 | Biological Process | GO:0007275 | multicellular organismal development | 16 |
| Liver_10 | Biological Process | GO:0032501 | multicellular organismal process | 20 |
| Liver_10 | Biological Process | GO:0008285 | negative regulation of cell proliferation | 3 |
| Liver_60 | Biological Process | GO:0044270 | nitrogen compound catabolic process | 4 |
| Liver_60 | Biological Process | GO:0006807 | nitrogen compound metabolic process | 14 |
| Liver_60 | Biological Process | GO:0055086 | nucleobase, nucleoside and nucleotide metabolic process | 11 |
| Liver_60 | Molecular Function | GO:0008252 | nucleotidase activity | 3 |
| Liver_10 | Molecular Function | GO:0000166 | nucleotide binding | 42 |
| Liver_60 | Biological Process | GO:0009117 | nucleotide metabolic process | 10 |
| Liver_10 | Cellular Component | GO:0043233 | organelle lumen | 8 |
| Liver_10 | Biological Process | GO:0006082 | organic acid metabolic process | 8 |
| Liver_60 | Biological Process | GO:0006082 | organic acid metabolic process | 16 |
| Liver_10 | Molecular Function | GO:0016491 | oxidoreductase activity | 18 |
| Liver_60 | Molecular Function | GO:0016491 | oxidoreductase activity | 39 |
| Liver_10 | Molecular Function | GO:0016725 | oxidoreductase activity, acting on CH or CH2 groups | 2 |
| Liver_60 | Molecular Function | GO:0016614 | oxidoreductase activity, acting on CH-OH group of donors | 6 |
| Liver_60 | Molecular Function | GO:0016627 | oxidoreductase activity, acting on the CH-CH group of donors | 7 |
| Liver_60 | Molecular Function | GO:0016616 | oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor | 6 |
| Liver_60 | Molecular Function | GO:0004859 | phospholipase inhibitor activity | 3 |
| Liver_60 | Molecular Function | GO:0016775 | phosphotransferase activity, nitrogenous group as acceptor | 4 |
| Liver_10 | Biological Process | GO:0043687 | post-translational protein modification | 21 |
| Liver_10 | Molecular Function | GO:0016504 | protease activator activity | 2 |
| Liver_60 | Molecular Function | GO:0030674 | protein binding, bridging | 3 |
| Liver_60 | Molecular Function | GO:0004673 | protein histidine kinase activity | 3 |
| Liver_10 | Biological Process | GO:0006464 | protein modification process | 23 |
| Liver_60 | Biological Process | GO:0051258 | protein polymerization | 5 |
| Liver_10 | Biological Process | GO:0022618 | protein-RNA complex assembly | 4 |

| | | | | |
|---|---|---|---|---|
| Liver_10 | Molecular Function | GO:0017076 | purine nucleotide binding | 34 |
| Liver_10 | Molecular Function | GO:0032555 | purine ribonucleotide binding | 32 |
| Liver_60 | Molecular Function | GO:0005097 | Rab GTPase activator activity | 5 |
| Liver_60 | Biological Process | GO:0032482 | Rab protein signal transduction | 5 |
| Liver_60 | Molecular Function | GO:0005099 | Ras GTPase activator activity | 5 |
| Liver_60 | Biological Process | GO:0065008 | regulation of biological quality | 13 |
| Liver_60 | Biological Process | GO:0050878 | regulation of body fluid levels | 5 |
| Liver_60 | Biological Process | GO:0032313 | regulation of Rab GTPase activity | 5 |
| Liver_60 | Biological Process | GO:0032483 | regulation of Rab protein signal transduction | 5 |
| Liver_10 | Biological Process | GO:0006950 | response to stress | 10 |
| Liver_60 | Biological Process | GO:0009611 | response to wounding | 7 |
| Liver_60 | Biological Process | GO:0022613 | ribonucleoprotein complex biogenesis and assembly | 7 |
| Liver_10 | Molecular Function | GO:0032553 | ribonucleotide binding | 32 |
| Liver_10 | Molecular Function | GO:0016455 | RNA polymerase II transcription mediator activity | 2 |
| Liver_10 | Molecular Function | GO:0004714 | transmembrane receptor protein tyrosine kinase activity | 4 |
| Liver_60 | Molecular Function | GO:0019842 | vitamin binding | 9 |
| Liver_60 | Biological Process | GO:0042060 | wound healing | 5 |

# Appendix C

## SUMMARY TABLE OF GENES INVESTIGATED USING RT-qPCR

| Gene Symbol | Gene Name | Forward primer sequence | Reverse primer sequence |
|---|---|---|---|
| HMGCS1 | 3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 | CATCCCAAGCCCTGCTAAGA | TGCAACAGTAACGCCTTCAGA |
| ACACA | acetyl-Coenzyme A carboxylase alpha | CCCGAGGTTGCCATGGA | GCTCGTTGGGTGGGTGATAT |
| AGMAT | agmatine ureohydrolase | TGGACTGCAGACGTGTGGTT | GTCCCGGCAGTACTTGTTACG |
| AKR1B10 | aldo-keto reductase family 1, member B10 | TCCAGATCCAGAGGAATGTGATT | AAAGACCTTGAAGTTCTCCACAATG |
| ALB | albumin | TCAGATAAGCCAGAAATTCCCTAAA | CACGCATACACTCCAGCATATCTC |
| ANXA13 | annexin A13 | GCAAAAATGGGCAGTTCACA | TCTTAGCATCTCTGTCTGCATCAAA |
| ANXA2 | annexin A2 | CGAGTCGCCATGCAAACC | TGGGAAGCACACAGCATCAG |
| BMP2 | bone morphogenetic protein 2 | CAAAAGAGAAAAGCGTCAAGTGAA | CGGATGCCTTTTGCAACTGT |
| CDC16 | cell division cycle 16 homolog | CTGTACGTTGCTAGGGTTCTATTATCC | GGGAGATGTCACAGTCTTCTATCAAG |
| CPOX | coproporphyrinogen oxidase | GCCACCGAGGACGGAAA | CATGAGGATTCTTTGGGTGGAT |
| CYP7A1 | cytochrome P450, family 7, subfamily A, polypeptide 1 | GATGACATGGAGAAAGCGAAGA | CTCCAAAAAGTGGCAGGAATG |
| ENO2 | enolase 2 (gamma, neuronal) | GGGATTCAGATTGTGGGAGATG | TTCTCTTCAACAGCTCGCTCAA |
| FASN | fatty acid synthase | GTTGGCACAGTGGCTAATTGAG | CGCCTTCCATTCTCTAACACACTT |
| GOT1 | aspartate aminotransferase | GGGAGCGCGCATTGTG | CTTCACGTTGTCCTTCCATTCA |
| GSTA | glutathione S-transferase class-alpha | CTCGTTGGCAACAAGCTAAGC | GAGGGAATGCAGACAGTATATCAGACT |
| HMG | 3-hydroxy-3-methylglutaryl-CoA reductase | GCTGCACAATGCCATCTATAGAA | TTGAACCCCTAACATCTGCAAA |
| IARS | isoleucyl-tRNA synthetase | TCAGCCAGTTCCCAAAACG | TGGTGGAGAAGCTACTGGAGAAT |
| LDHA | lactate dehydrogenase A | CACACTCCAAGCTGGTCATTGT | GGACCAAGTTAAGACGGCTTTCT |
| LOC418187 | organic anion transporting polypeptide 1c1 | GAAATCTCTTGCAGTTGGTCTATATACAC | CAATCAGCGCACCAAAATAAAC |
| LOC427984 | monovalent inorganic cation transport | TGGGATTTCTTTGTTTATGATCTTAGC | GGTGTGTCTGCTTTTAGCATTGG |
| LOC429084 | coagulation factor III (thromboplastin, tissue | CCATACATTTGAAATAAGACGTTGACA | CAGTGGTACAAAGCACCATGCT |

| | | | |
|---|---|---|---|
| | factor) | | |
| MDH1 | malate dehydrogenase 1, NAD | GGGAATCCAGCAAACACTAACTG | AGCGAGTCAAGCAGCTGAAGT |
| ABCB4 | ATP-binding cassette, sub-family B (MDR/TAP), member 4 | GGCCAGAAGCAAAGAATAGCA | TGTTGCCTCATCAAGCAGAAGA |
| NDUFB5 | NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa | TCTAGAGCACTGGGAATACTACAAACA | CCATCTCTTTCTCATAGTGCTTTTCA |
| ODC1 | ornithine decarboxylase 1 | GGTGACTGGATGCTGTTTGAAA | ATGTATTGTTGGCCTCTGGAATC |
| PAH | phenylalanine hydroxylase | TTCCCCAGCTTGAAGATGTTTC | GGAGAGCAAGCCTGCAACAG |
| PCK1 | phosphoenolpyruvate carboxykinase 1 | CAAGTGAGGGAGGTTCATTGAAA | TCTTCCTCTGAACCATCACAGATATG |
| PPARGC1A | peroxisome proliferator-activated receptor gamma, coactivator 1 alpha | TCAGCACACAAAACCATGCA | CTCCACGAATTCTCAGTCTTAACAAC |
| RYR3 | ryanodine receptor 3 | TGGGATTCAAGACATTACGAACTATACT | TACTGCCAATAATCCCACTGTTAAGA |
| SLC5A1 | sodium/glucose cotransporter | GCTCTTTGATTACATTCAGTCAGTTACC | GACGCCGAGCAGGAAGAC |
| TRIP12 | thyroid hormone receptor interactor 12 | ATGGGTTGCTTTCATCATCAGTAG | CTGTGTATCAAGCAGTTAAGCTGTACAG |
| UGDH | UDP-glucose dehydrogenase | CAGATCGAAACAATTGGGAAGAA | CGGAAATCCTTCAGTACCTATTAGACA |
| UROS | uroporphyrinogen III synthase | CAGGGAGTTCCAGCAAGCA | GGATATGCGGGAGGCAAAA |
| XDH | xanthine dehydrogenase | CCCAACCAGCTGCATCTGT | GATGGTCTGACATGCATTATGGA |

# REFERENCES


Altschul,S.F. *et al*. (1995) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215,** 403-410.

Ankley,G.T. *et al*. (2006) Toxicogenomics in regulatory ecotoxicology. *Environmental Science & Technology*, **40,** 4055-4065.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25,** 25-29.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17,** 509-519.

Balthazart,J., Tlemcani,O. and Ball,G.F. (1996) Do sex differences in the brain explain sex differences in the hormonal induction of reproductive behavior? What 25 years of research on the Japanese quail tells us. *Hormones and Behavior*, **30,** 627-661.

Bork,P. *et al*. (1998) Predicting function: From genes to genomes and back. *Journal of Molecular Biology*, **283,** 707-725.

Boyle,E.I. *et al*. (2004) GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20,** 3710-3715.

Bragg,L.M. and Stone,G. (2009) k-link EST clustering: evaluating error introduced by chimeric sequences under different degrees of linkage. *Bioinformatics*, **25,** 2302-2308.

Brazma,A. and Vilo,J. (2001) Gene expression data analysis. *Microbes and Infection*, **3,** 823-829.

Causton,H.C., Quackenbush,J. and Brazma,A. (2003) A Beginner's Guide Microarray Gene Expression Data Analysis. Blackwell Publishing.

Chevreux,B. *et al*. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14,** 1147-1159.

Chou,H.H. and Holmes,M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17,** 1093-1104.

Cogburn,L.A. *et al*. (2007) Functional genomics of the chicken - A model organism. *Poultry Science*, **86,** 2059-2094.

Conte,M.G. *et al*. (2008) Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics*, 9.

Dahlquist,K.D. *et al*. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, **31,** 19-20.

Dalca,A.V. and Brudno,M. (2010) Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics*, **11,** 3-14.

Darling,A., Carey,L. and Feng,W. (2003) The Design, Implementation, and Evaluation of mpiBLAST.

Davis A.P. *et al*. (2008) The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. *BMC Med Genomics*, 9.

Dennis,G. *et al*. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4.

Dutilh,B.E., Huynen,M.A. and Strous,M. (2009) Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics*, **25,** 2878-2881.

Ekman D R *et al*. (2003) SAGE Analysis of Transcriptome Responses in Arabidopsis Roots Exposed to 2,4,6-Trinitrotoluene1. *Plant Physiology*, **133,** 1397-1406.

Elder,G.H. (1998) Genetic defects in the porphyrias: Types and significance. *Clinics in Dermatology*, **16,** 225-233.

Ellegren,H. (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17,** 1629-1631.

Engelhardt,B.E. *et al*. (2005) Protein Molecular Function Prediction by Bayesian Phylogenomics. *PLOS Computational Biology*, 1.

Ferguson,J.W. and McCain,W.C. (1999) Toxicological Study No. 6955-31-97-05-02, 14-Day Feeding Study of Hexahydro-1,3,5, Trinitro-1,3,5-Triazine (RDX) in the White-Footed Mouse, Peromyscus leucopus. U.S. Army Center for Health Promotion and Preventive Medicine (USACHPPM).

Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nat Genet.*, **3,** 266-272.

Gouault-Helimann,M. and Josso,F. (1979) Initiation in vivo of blood coagulation. The role of white blood cells and tissue factor. pp. 3249-3253.

Go´mez-Valade´s,A.G. *et al*. (2008) *Pck1* Gene Silencing in the Liver Improves Glycemia Control, Insulin Sensitivity, and Dyslipidemia in *db/db* Mice. pp. 2199-2210.

Greaves,P. (2007), *Histopathology of Preclinical Toxicity Studies: Interpretation and Relevance*, p. 508.

Green,P.G., STRAUSBAUGH,H.G. and Levine,J.D. (1998) Annexin I Is a Local Mediator in Neural-Endocrine Feedback Control of Inflammation. pp. 3120-3126.

Gust,K.A. *et al*. (2009) Neurotoxicogenomic Investigations to Assess Mechanisms of Action of the Munitions Constituents RDX and 2,6-DNT in Northern Bobwhite (Colinus virginianus). *Toxicological Sciences*, **110,** 168-180.

Harris,M.A. *et al*. (2006a) The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, **34,** D322-D326.

Harris,M.A. *et al*. (2006b) The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, **34,** D322-D326.

Heng Li and Nils Homer. (2010) A survey of sequence alignment algorithms for next generation sequencing. *Briefings in Bioinformatics*.

Heng Li and R Durbin. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26,** 589-595.

Hillier,L.W. *et al*. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432,** 695-716.

Hu,W. *et al*. (2003) Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource. *Nature Genetics*, **35,** 139-147.

Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4,** 44-57.

Huang,X.Q. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Research*, **9,** 868-877.

Hudek,A.K. *et al*. (2003) Genescript: DNA sequence annotation pipeline. *Bioinformatics*, **19,** 1177-1178.

Irizarry,R.A. *et al*. (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31.

Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequence. *Proc. Int. Conf. Intell. Syst. Mol. Biol*, 138-147.

Jenkins,T.F. *et al*. (2006) Identity and distribution of residues of energetic compounds at army live-fire training ranges. *Chemosphere*, **63,** 1280-1290.

Johnson,M.S. *et al*. (2005) Influence of oral 2,4-dinitrotoluene exposure to the Northern Bobwhite (Colinus virginianus). *International Journal of Toxicology*, **24,** 265-274.

Johnson,M.S. *et al*. (2007) Subacute toxicity of oral 2,6-dinitrotoluene and 1,3,5-trinitro-1,3,5-triazine (RDX) exposure to the northern bobwhite (Colinus virginianus). *Environmental Toxicology and Chemistry*, **26,** 1481-1487.

Johnson,M.S. *et al*. (2000) Fate and the biochemical effects of 2,4,6-trinitrotoluene exposure to tiger salamanders (Ambystoma tigrinum). *Ecotoxicol. Environ. Saf*, **46,** 186-191.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, **34,** D354-D357.

Kent,W.J. (2002) BLAT - The BLAST-like alignment tool. *Genome Research*, **12,** 656-664.

Kitano,H. (2002) Systems biology: A brief overview. *Science*, **295,** 1662-1664.

Kiyosawa,N. *et al*. (2010) Gene set-level network analysis using a toxicogenomics database. *Genomics*, **96,** 39-49.

Lalli,E. and SassoneCorsi,P. (1994) Signal-Transduction and Gene-Regulation - the Nuclear Response to Camp. *Journal of Biological Chemistry*, **269,** 17359-17362.

Lamb,J. *et al*. (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, **313,** 1929-1935.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10.

Leis,H.J. *et al*. (1998) Prostaglandin Endoperoxide Synthase-2 Contributes to the Endothelin/Sarafotoxin-Induced Prostaglandin E2 Synthesis in Mouse Osteoblastic Cells MC3T3-E1): Evidence for a Protein Tyrosine Kinase-Signaling Pathway and Involvement of Protein Kinase C. pp. 1268-1277.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98,** 31-36.

Li,H. *et al*. (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25,** 2078-2079.

Li,H., Ruan,J. and Durbin,R. (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18,** 1851-1858.

Li,R.Q. *et al*. (2008b) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24,** 713-714.

Li,R.Q. *et al*. (2009b) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25,** 1966-1967.

Liang,H. and Ward,W.F. (2006) PGC-1alpha: a key regulator of energy metabolism. pp. 145-151.

Lodish,H. *et al*. (2004) *Molecular Cell Biology*. W.H.Freeman and Company.

Makowski,L. *et al*. (2009) Metabolic profiling of PPAR–/– mice reveals defects in carnitine and amino acid homeostasis that are partially reversed by oral carnitine supplementation. pp. 586-604.

Mao,X.Z. *et al*. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21,** 3787-3793.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437,** 376-380.

Mattingly,C.J. *et al*. (2006) The Comparative Toxicogenomics Database (CTD): A resource for comparative toxicological studies. *Journal of Experimental Zoology Part A-Comparative Experimental Biology*, **305A**, 689-692.

Metzker,M.L. (2010) Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. *Nature Reviews Genetics*, **11,** 31-46.

Meyer,E. *et al*. (2009) Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, 10.

Miller,J.R., Koren,S. and Sutton,G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95,** 315-327.

Moreno-Hagelsieb G,L.K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24,** 319-324.

Morteau,O. (2004) Prostaglandins Intestinal Edema. *Oral Tolerance: the response of the intestinal mucosa to dietary antigens*.

Nagaraj,S.H. *et al*. (2007) ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. *Nucleic Acids Research*, **35,** W143-W147.

Ning,Z.M., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: A fast search method for large DNA databases. *Genome Research*, **11,** 1725-1729.

Papanicolaou,A. *et al*. (2009) Next generation transcriptomes for next generation genomes using est2assembly. *Bmc Bioinformatics*, 10.

Patel,C.J. and Butte,A.J. (2010) Predicting environmental chemical factors associated with disease-related gene expression data. *Bmc Medical Genomics*, 3.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci*, **85,** 2444-2448.

Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, **6,** S22-S32.

Perea,M.T. *et al*. (2009) Influence of avian reproduction ecotoxicological endpoints in the assessment of plant protection products. *Journal of Environmental Science and Health Part B-Pesticides Food Contaminants and Agricultural Wastes*, **44,** 106-112.

Pertea,G. *et al*. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19,** 651-652.

Potter,S.C. *et al*. (2004) The ensembl analysis pipeline. *Genome Research*, **14,** 934-941.

Quevillon,E. *et al*. (2005) InterProScan: protein domains identifier. *Nucleic Acids Research*, **33,** W116-W120.

Quinn,M.J., Jr. *et al*. (2007) Effects of subchronic exposure to 2,6-DinitroToluene in the Northern Bobwhite (*COLINUS VIRGINIANUS*). *Environ. Toxicol. Chem.*, **26,** 2202-2207.

Quinn,M.J., Jr. *et al*. (2009) Sublethal effects of subacute exposure to RDX (1,3,5-trinitro-1,3,5-triazine) in the Northern Bobwhite (*COLINUS VIRGINIANUS*). *Environ. Toxicol. Chem.*, **28,** 1266-1270.

Rawat *et al*.,A. (2010a) CAPRG: Sequence alignment pipeline for next generation sequencing for non model organisms. *Manuscript in preperation*.

Rawat,A. *et al*. (2010b) Quail Genomics: a knowledgebase for Northern bobwhite. *Bmc Bioinformatics*, 11.

Rawat,A. *et al*. (2010c) From raw materials to validated system: The construction of a genomic library and microarray to interpret systemic perturbations in Northern bobwhite. *Physiological Genomics*.

Reddy G *et al*. (2000) Toxicity of 2,4,6-TrinitrotoIuene (TNT) in Hispid Cotton Rats (*Sigmodon hispidus*): Hematological, Biochemical, and Pathological Effects. *International Journal of Toxicology*, **19,** 169-177.

Reece,J.B., Campbell,N.A. and Mitchell,L.G. (2002) *Biology*. Benjamin Cummings.

Sabbioni,G. *et al*. (2005) Hemoglobin adducts, urinary metabolites and health effects in 2,4,6-trinitrotoluene exposed workers. *Carcinogenesis*, **26,** 1272-1279.

Salomonis,N. *et al*. (2007) GenMAPP 2: new features and resources for pathway analysis. *Bmc Bioinformatics*, 8.

Sauer,J.R. *et al*. (2005) Using the North American Breeding Bird Survey as a tool for conservation: A critique of BART *et al*. (2004). *Journal of Wildlife Management*, **69,** 1321-1326.

Shames,D.S., Minna,J.D. and Gazdar,A.F. (2007) DNA Methylation in Health, Disease, and Cancer. pp. 85-102.

Sims,J.G. and Steevens J A. (2008) The role of metabolism in the toxicity of 2,4,6-trinitrotoluene and its degradation products to the aquatic amphipod Hyalella azteca. *Ecotoxicol Environ Saf.*, **70,** 38-46.

Smith,T.F. and Waterman,M.S. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, **147,** 195-197.

Soderlund,C. *et al*. (2009) PAVE: Program for assembling and viewing ESTs. *BMC Genomics*, 10.

Stekel,D. (2003) *Microarray Bioinformatics*.

Talmage,S.C. *et al*. (1999) Nitroaromatic munition compounds: environmental effects and screening values. *Rev. Environ. Contam. Toxicol.*, **161,** 1-156.

Tchounwou,P.B. *et al*. (2001) Transcriptional Activation of Stress Genes and Cytotoxicity in Human Liver Carcinoma Cells (HepG)Exposed to 2,4,6-Trinitrotoluene, 2 2,4-Dinitrotoluene, and 2,6-Dinitrotoluene. *Environ. Toxicol.*, **16,** 209-216.

Thomas,P.D., Mi,H.Y. and Lewis,S. (2007) Ontology annotation: mapping genomic regions to biological function. *Current Opinion in Chemical Biology*, **11,** 4-11.

Turk,E., Martin,M.G. and Wright,E.M. (1994) Structure of the Human Na+/Glucose Cotransporter Gen*SeG LTI*. pp. 15204-15209.

Udall,J.A. *et al*. (2006) A global assembly of cotton ESTs. *Genome Research*, **16,** 441-450.

Velculescu,V.E. *et al*. (1997) Characterization of the yeast transcriptome. *Cell*, **88,** 243-251.

Vera,J.C. *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17,** 1636-1647.

Wang,J. *et al*. (2007) An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. *Genome Biology*, 8.

Werner,T. (2008) Bioinformatics applications for pathway analysis of microarray data. *Current Opinion in Biotechnology*, **19,** 50-54.

Wintz,H. *et al*. (2006) Gene Expression Profiles in Fathead Minnow Exposed to 2,4-DNT: Correlation with Toxicity in Mammals. pp. 71-82.

Wu,J.M. *et al*. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Research*, **34,** W720-W724.

Ye,J. *et al*. (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research,* **34,** W293-W297.

Ye,R.W. *et al*. (2001) Applications of DNA microarrays in microbial systems. *Journal of Microbiological Methods*, **47,** 257-272.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics,* **17,** 847-848.

Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18,** 821-829.

Zhang,Z. *et al*. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7,** 203-214.