Functional Annotation of the Bovine and Porcine Genomes

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

John R. Garbe

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Scott Fahrenkrug

August 2015

Abstract

Pigs and cattle are two of the most important sources of animal protein for the human population around the world. Continued increase in production is necessary in order to meet rising demands for animal food products. Large gains in animal production, efficiency, health, and welfare have been achieved through genetic improvement using traditional breeding methods. Commercialized high-throughput genomic technologies are being incorporated into breeding programs to increase the rate of genetic improvement of livestock. The availability of the porcine and bovine genome sequences presents an opportunity to better understand the genetic causes of variation in animal performance. This thesis reports several experiments that identify and characterize this variation. Two high-throughput gene expression assay platforms for use in identifying genes associated with production traits in pigs and cattle are annotated and their performance characterized. Genes whose expression patterns are associated with milk yield in dairy cattle and the efficiency of conversion of feed to muscle in beef cattle are identified. A collection of forty-eight million points of variation in the bovine genome was characterized by location and effect. Better understanding of animal biology also benefits human health through the use of animal models in biomedical research. Towards this aim a resource to aid the development of genetically modified animals was developed from a comprehensive transfer of biomedical annotation from human and model mammalian genomes to the pig genome. These annotations are a resource for the better understanding of genetic causes of variation in animal performance and for developing methods for

applying this information to improve animal performance for both agricultural and biomedical purposes.

Table of Contents

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Justification of the problem

Artificial selection in livestock has been enormously successful, responsible for dramatic increases in milk production in dairy breeds over the second half of the last century. This improvement was based on the selective breeding of high merit bulls, where merit has traditionally been determined by the performance measurement of a bull's female offspring. New developments in cattle genomics now allow more accurate and earlier merit estimation through the incorporation of genetic information. The Illumina BovineSNP50 high-throughput genotyping chip is used to test associations between 50,000 polymorphic locations in the bovine genome with production traits to estimate the genetic merit of an animal [1]. The genomic information obtained by this $85 test has about the same impact on an animal's estimate of meritas adding fifteen daughters to a bull's proof, and is available the day the bull is born [2]. The Illumina BovineHD SNP Chip, which genotypes 770,000 SNPs, provides an order of magnitude more data than the 50K chip. However, these tests return merely a sliver of the information encoded in the bovine genome. Rapidly decreasing sequencing costs will soon enable even the complete genome sequencing of individual animals for merit estimation, producing three billion base-pairs of information about each animal. There are thousands of genes in the bovine genome, but little is known about which genes are associated with important production traits. Millions of locations in the genome differ between any two animals, but it is unknown which of these variations are responsible for the large differences seen in their production performance. Identifying genes associated with production traits and identifying the genomic variation that causes changes in performance is the key to harnessing the power of high-throughput genomic technology and applying it to ensure the continued improvement of livestock.

Understanding a genome is a three part process: sequencing and assembling a genome build to determine the nucleotide sequence of the genome; structurally annotating the genome to identify genomic elements including genes and their components, regulatory motifs, and sequence variations; and functional annotation, attaching biological information to genomic elements such as the expression and biochemical and biological function of genes involved in regulation and interactions, and the biological effects of sequence variation. Each of these steps is more difficult than the one before it, but yields more useful biological information. A usable draft bovine genome build is available, but the current structural annotation is incomplete and the functional annotation is sparse. The draft pig genome sequence is even less complete. Improved structural and functional annotations of these genomes is required in order to identify and understand the genetic components of economically important traits that will facilitate the development of improved methods for the genetic improvement of these animal species. The increasing use of pigs in biomedical research also motivates the deciphering of the swine genome in order to enable the development of large animal models of human diseases.

1.2 Literature review

The bovine genome was sequenced from a Hereford cow by the Bovine Genome Sequencing and Analysis Consortium using a combination of bacterial artificial chromosome and whole-genome shotgun sequencing [3]. The consortium has developed four major iteratively improved assemblies of the genome, with each newer version incorporating more sequence data and improved assembly methods. The current version is Btau4.1 which covers 92% of the estimated 2.87 Gb of the genome [4]. The Salzberg group at the University of Maryland has recently released an alternative assembly, University of Maryland assembly release 3 (UMD3.1), which utilizes additional post-processing algorithms to produce a genome build with more coverage, fewer gaps, fewer inversions, deletions, and translocations, and fewer single-nucleotide errors [5].

The swine genome was sequenced from a Duroc pig by the Swine Genome Sequencing Consortium and the Wellcome Trust Sanger Institute using bacterial artificial chromosome sequencing. The first 4X draft porcine genome build (Sscrofa9.2) covering all chromosomes has been released with build 10.2 nearing completion, although the International Swine Genome Sequencing Consortium has not yet published a paper detailing the results.

The bovine and swine genomes have been structurally annotated by annotation pipelines which use specialized software to identify genes and genomic variation including single nucleotide polymorphisms (SNPs) and repeats. Gene annotation, a difficult problem due to the complex structures of eukaryotic genes, is accomplished using a combination of *ab* *initio* and homology based methods. *Ab initio* gene prediction methods use Hidden Markov Models and an *a priori* model of gene structural components such as codon usage, transcription initiation, and polyA signals to predict gene structure. Homology based methods use alignments of mRNA reference sequences and expressed sequence tags (ESTs) to identify the coding regions of genes. Both methods frequently generate erroneous gene models due to the complex structures of eukaryotic genes and incomplete transcript sequence databases. The National Center for Biotechnology Information, the Ensembl project, and the University of Santa Cruz have all generated automated gene annotations of the pig and bovine genomes. Higher quality gene annotations can be generated by expert human analysis but such manual annotation is labor intensive. Btau4.1 has had over 4,000 genes manually annotated and the pig community has also been organizing manual gene annotation efforts.

Functional annotation of the bovine and porcine genomes has primarily been accomplished through homology based annotation. The sequences of genes of unknown function are aligned against gene sequences from related species for which functional annotation is available. High sequence similarity between two genes implies high functional similarity, and functional annotation is transferred from annotated genes to unannotated genes. The mouse and human genomes have comparatively large amounts of functional annotation that is easily transferred to the bovine and porcine genomes, however, this annotation is aimed at basic biological or biomedical functions, not the traits and phenotypes that are of economic importance in agricultural animals. Porcine and bovine specific functional annotation requires expensive laboratory work to obtain.

Functional annotation can be derived from high-throughput gene expression studies using microarrays or transcriptome sequencing, association studies, or using automated annotation methods that estimate the functional effects of genomic variation. The Animal Quantitative Trait Locus database maintains a collection of all publicly available association data for livestock species including pigs and cattle. The Gene Expression Omnibus is the equivalent database for gene expression studies, but is not limited to just agricultural species. Both array- and sequence-based datasets are catalogued [6]. While a myriad of gene expression and association studies have been published, functional annotation remains sparse and current high-throughput methods offer the opportunity to both validate previous findings and discover new results at higher resolution.

1.3 Specific Aims

The focus of this research was on four specific aims: the development of annotation for the Bovine Oligonucleotide Microarray (BOM) and Swine Protein-Annotated Microarray (SPAM); the application of the BOM microarray and whole-transcriptome sequencing (RNA-Seq) to catalog gene expression patterns in cattle and identify genes involved in economically important traits; annotating high-density bovine genomic variation to identify potentially functional variants in the bovine variome; and developing annotation to guide the development of genetically modified pigs for use in biomedical research.

2 Aim 1: Annotation of BOMC and SPAM arrays

2.1 Introduction

Microarrays previously developed for cattle have primarily utilized amplified cDNA probes or have designed oligonucleotide probes in the absence of the currently available bovine genome sequence [7-9]. Further, oligonucleotide array designs have focused almost exclusively on nucleic acid sequences, without invoking more sophisticated annotation techniques that can differentiate orthologous and paralogous genes [10]. The design of the BOM has foremost relied on the assignment of bovine expressed sequences to phylogenetically defined vertebrate proteins. Consensus sequences were created by clustering the ESTs assigned to protein families and these were aligned to the Btau2.0 draft bovine genome sequence assembly [4] to maximize cardinality and reduce probe redundancy. Probes for the BOM were designed primarily within 3' biased exons predicted to be constitutively expressed and to have approximately constant Tm and unique representation within Btau2.0. Similarly, porcine expressed sequences were assigned to protein families and clustered against vertebrate protein sequences. Advances in the quality of the bovine genome builds since the design of the arrays presents an opportunity to generate updated array annotation, as well as a retrospective bioinformatic characterization of the bovine probe set in terms of redundancy, mismatch stringency, and genome representation using the bovine genome assembly build UMD3.1.

2.2 Methods

Post-facto genome based annotation: Annotation of probes on both microarrays was reassessed to take advantage of the most recent builds of the bovine and porcine genomes. To assign annotation to the 23,580 experimental bovine oligo probes the consensus sequences were compared to the UMD2.0 build of the bovine genome [5] with GMAP [11] and the Decypher system GeneDetective program resulting in alignment of 21,976 consensus sequences with coverage ≥ 50% and identity ≥ 95%. Of these, 20,336 have coverage ≥ 95%. Comparison to the annotated transcriptome revealed that 13,939 consensus sequences mapped to the UMD2.0 cDNA set with coverage ≥ 50% and identity ≥ 95% using the Decypher system BLASTn program. Of these, 8,014 have ≥ 95% coverage. The remaining 9,051 consensus sequences were queried by BLAST against the Ensembl (release 52) bovine cDNA set with 873 consensus sequences mapping with coverage ≥ 50% and identity ≥ 95%. Of these, 264 have ≥ 95% coverage. The remaining 7,888 unassigned sequences were queried by BLAST against the NCBI bovine cDNA set with 315 consensus sequences mapping with coverage ≥50% and identity ≥ 95%. Of these, 180 have coverage ≥95%. The remainder were queried by BLAST against NCBI human cDNA with 3,703 alignments with an e-value < 1e-20. In total, 18,830 consensus sequences align to a cDNA and 4,750 have no gene assignment and may represent unannotated genes. Suspected chimeric consensus sequences were identified as those sequences where the 70bp oligo portion of the consensus sequence lies outside of the portion of the consensus sequence that aligns to the genome. For 127 consensus sequences fitting this criterion, the non-aligning portion of the consensus sequence was aligned to the genome using GMAP producing 27 alignments with coverage ≥ 50% and identity ≥ 95%. In total, 18,207 consensus sequences have both gene assignments and genomic alignments, 3,769 have exclusively genomic alignments, 585 have exclusively gene assignments, and 1,016 have no alignment to either cDNA or the genome. These unannotated sequences correspond to 995 ESTs from a variety of sources and 23 contigs.

To assign annotation to the 19,980 pig oligo probes, the consensus sequences were compared to the Sscr9 build of the swine genome (Ensembl release 56) with GMAP and Decypher resulting in the alignment of 14,919 consensus sequences with coverage ≥ 50% and identity ≥ 95%. Of these, 12,600 have coverage ≥ 95%. Comparison to the annotated transcriptome revealed that 9,329 consensus sequences mapped to the Sscr9 cDNA (Ensembl release 56) with coverage ≥ 50% and identity ≥ 95% using the Decypher system BLASTn program. Of these, 5,227 have ≥ 95% coverage. An additional 89 consensus sequences aligned to NCBI pig cDNA sequences and 8,827 aligned to NCBI human cDNA with an e-value < 1e-20. In total, 13,950 consensus sequences have both gene assignments and genomic alignments, 969 have only genomic alignments and 4,295 have exclusively gene assignments. Only 766 consensus sequences have no annotation, comprised of 58 tentative consensus sequences (TCs), 699 contigs and 9 provisional RefSeqs no longer included in the NCBI porcine annotation.

Array printing, tissue samples and experimental design: The microarrays were printed from 384-well plates in which the synthetic 70-mer single stranded oligodeoxyribonucleotides were dissolved to 20 micromolar in 3X SSC and to a final volume of 15 microliters. Printing was performed using a Genomic Solutions Omnigrid 300 microarray printer, equipped with a Telechem Stealth 48 pin print-head containing SMP3 pins. The print format produces a single array containing 12 metarows and 4 metacolumns, with each subarray containing 25 columns and 21 rows and with an element center-to-center spacing of 170 x 165 micrometers. Microarrays were baked after printing for 2 hr at 80 ^{o}C. Slide rehydration was performed over 50 ^{o}C water, followed by snap drying on a 65 ^{o}C heating block for 5 sec; this process was repeated three times. Slides were UV-cross-linked at 120 mJ, washed in 1% (w/v) SDS for 5 min at room temperature, then in water, and were finally spin dried by centrifugation at 1,000 g for 2 min.

Six bovine tissue samples (small intestine, spleen, liver, adrenal gland, anterior pituitary, and thymus) were collected from each of 6 Angus steers at 14 months of age following approved animal use protocols. The tissue samples were immediately frozen on dry ice and stored at -80 °C prior to RNA extraction. RNA was extracted and cDNA synthesized at the University of Missouri (MU) and aliquots of dye-labeled cDNA samples were used to replicate all hybridizations at the University of Minnesota (UMN) and at MU. At each location, samples were hybridized to 36 microarrays using a loop design with like tissues hybridized to the same array and with duplicate samples labeled with Cy3 and Cy5 as technical replicates.

Reverse transcription and array hybridization: Total RNA was extracted using 3 ml of TRI reagent (Ambion, Austin, TX) per 250-300 mg of tissue from each sample. The extract was treated with 0.02 Units of DNase 1 (Ambion) and cleaned up using phenol:chloroform:isoamyl alcohol (25:24:1), a Phase Lock Gel Heavy tube (Eppendorf, Hamburg, Germany) and a YM-30 microcon tube (Millipore, Billerica, MA). Five micrograms of total RNA was used to synthesize cDNA for each dye using 8 thermal cycles at 52 °C for 10 sec and 44 °C for 15 min. The reaction was stopped with 3.5 µl of 0.5 M NaOH/50 mM EDTA and then heated at 65 °C for 15 min. The solution was neutralized with 5 µl of 1 M Tris-HCl (pH 7.5). Finally, 10 µl of cDNA was purified using a MinElute PCR purification kit (Qiagen, Valencia, CA).

The microarrays were prehybridized for 1 hr with 0.2% I-Block (Tropix, Bedford, MA) in 1X PBS solution at 42 °C, then were washed with distilled water at room temperature for 10 min and finally were again washed with isopropanol at room temperature for 5 min using a rotary shaker. The arrays were dried by centrifugation for 5 min at 1000 g. The cDNA and fluorescence dye hybridization steps were accomplished by a modification to the 3DNA array 350 kit protocol (Genisphere Inc., Hatfield, PA). A total of 20 µl of Cy3 (10 µl) and Cy5 (10 µl) labeled cDNA samples was hybridized to each array at 55 °C in a water bath for 16 hr in a dark humidified chamber. The arrays were then washed for 15 min with 2X SSC/0.2% SDS at 55 °C, for 15 min in 2X SSC at room temperature, and finally washed again for 15 min in 0.2X SSC at room temperature. The arrays were again dried by centrifugation for 5 min at 1000 g. Both Cy3 and Cy5 capture reagents were combined with the hybridization buffer and were hybridized to an array for 4 hr at 55 °C in a water bath. The arrays were rewashed and dried as previously described [12].

Data extraction and normalization: At MU, the arrays were immediately scanned on an Axon Genepix 4000B laser scanner (Axon Instruments, Foster City, CA), while at UMN, the arrays were scanned on a GSI Lumonics ScanArray 5000 laser scanner (GSI Lumonics, Watertown, MA). The image data were extracted using BlueFuse for microarrays (BlueGnome, Cambridge, UK) and spots with a quality score of 0 or with a confidence score of less than 0.1 were removed from the data. The filtered data for each array were next normalized by performing a confidence-weighted LOESS regression for each print-tip and then standardizing log-intensity values to have a zero mean and unit variance using JMP Genomics (SAS Institute, Cary, NC). Intensity data were extracted using customized PERL scripts and data were plotted using R (R Development Core Team, 2007).

2.3 Preliminary Results

Array annotation: The bovine array contains 23,580 probes with 18,830 mapped to cDNA. Considering that some gene products are targeted by more than one probe the 18,830 mapped probes represent 16,341 unique genes. Gene Ontology (GO) annotation [13] for the bovine and porcine genomes from Ensembl BioMart (release 56) [14] was used to assign functional annotation to the genes represented on the arrays. For the 16,341 unique gene transcripts represented on the array, bovine GO annotation was retrieved for 9,446 transcripts. For the remaining transcripts with no bovine GO annotation 2,745 GO terms were transferred from orthologous human cDNAs resulting in a total of 12,191 GO annotated transcripts. Functional coverage of the genes on the array was measured by comparing the bovine GO terms with the available GO annotated human genes (18,110). The treeplot in Figure 1a shows the ontological coverage of the array as compared to the human GO annotation for the three categories of GO annotation:

molecular function, biological process, and cellular component. In Figure 1a, each block represents a GO term with the size of the block being proportional to the number of human genes assigned that GO term. The color of the block indicates whether the term is over- or under-represented on the array in comparison to the human genome. The probes on the bovine array fall into the annotation categories in generally the same proportions as do the genes in the human genome, indicating that the microarray provides a broad and even coverage of bovine gene function. Some exceptions include the biological process categories of sensory perception of smell (olfaction), modification-dependent protein catabolic processes and interspecies interaction between organisms and the molecular function category of olfactory receptor activity (ORA) which are represented at less than 5% of their level in the human genome. The molecular function categories of catalytic activity, protein tyrosine kinase activity, and protein kinase activity are overrepresented on the array by over 10-to-one as compared to the human genome.

The cDNA annotation for the UMD2.0 bovine genome build has 22,447 genes, excluding pseudogenes. There are 16,341 unique genes represented on the bovine array providing a gene representation of about 72%. The microarray's physical coverage of the bovine genome is shown in Figure 2a. The density of the 21,976 probes with assigned genomic coordinates is plotted against the density of all annotated genes across the 29 bovine autosomes and the X chromosome, plus the unassigned contigs (U). The physical distribution of genes represented on the array closely mirrors the distribution of all genes on the array except for a few small regions of high gene density. An under-represented area on chromosome 15 (45-55 Mb) contains 246 genes, 231 with the ORA GO term.

Another under-represented region on chromosome 10 (20-30 Mb) contains 244 genes, 116 with the ORA GO term.

The physical and functional distribution of genes represented on the porcine array was likewise analyzed, although the lower refinement of the swine genome assembly resulted in a corresponding decrease in the total number of probes with GO annotation. The 18,245 probes on the array that map to cDNA represent 15,204 unique gene transcripts. Porcine GO annotation was available for 9,418 and GO terms were transferred from 5,964 orthologous human cDNA for a total of 11,738 GO annotated transcripts. Functional coverage of the array is shown by a treeplot comparison to annotated human genes in Figure 1b. The porcine array is deficient in some of the same categories as the bovine array, including olfactory receptor activity and interspecies interaction between organisms. The physical coverage of the array is shown in Figure 2b. Several gene-rich regions of the genome are under-represented on the array, such as from 20-30 Mb on chromosome 7 containing 377 genes, 116 with the ORA GO term, and the first 10 Mb of chromosome 9 which contains 246 genes, 166 with the ORA GO term.

Both arrays show a deficiency in representation of genes in the olfactory receptor gene family. This can be attributed to low representation in the initial bovine EST set where only 37 ESTs matched just 30 different ORA genes.

The specific location of a probe sequence within a gene is important for interpreting the magnitude of gene expression reported by a microarray. Alternative splicing, co-transcription, and distance from the 3’ end of the transcript all affect signal strength. The strength of the annotation assigned to a probe is also best understood within the context of a gene browser that displays ESTs, cDNA alignments, and other supporting evidence for a gene annotation. To this end, four distributed annotation system (DAS) sources are available from http://gnomix.ansci.umn.edu:9000/das for viewing the alignments of array probes and consensus sequences to genomic reference sequences using the Ensembl genome browser [15]. The UMN_Btau_BOM Consensus and UMN_Btau_BOM Oligo sources provide the alignments of sequences to the Btau4.0 reference sequence, and the UMN_Sscr_SPAM_Consensus and UMN_Sscr_SPAM_Oligo sources provide the alignments of SPAM sequences to the Sscr9 reference sequence. These alignments are helpful for further investigating probes identified as differentially expressed, such as for determining which exon(s) are detected by a probe, and therefore which isoforms of a gene are being detected. An example of this was observed for the consensus sequence corresponding to the *HNF4a* gene (Figure 3), where the consensus sequence aligns to the first three exons of the gene as well as a portion of intron 3. A probe designed against this intron detects expression of this unannotated exon. For those probes that do not align to a cDNA sequence but do align to the genome, this alignment will facilitate the discovery and annotation of new genes.

Microarray performance: A total of 72 hybridizations to the bovine array were performed to assess the specificity, stringency, and repeatability of the platform. To measure the specificity of the hybridization conditions, the expression levels detected by the negative control probes were compared to those for the experimental probes. The expression distributions of the 60 negative control probes are plotted in Figure 4, with the median expression level of all experimental spots shown in blue. The median experimental expression level was 3 times greater than the median negative control expression level indicating strong specific hybridization in the experiment. However, 7 negative control probes detected mean expression levels greater than the mean for the experimental probes. The oligo sequences for these probes had no BLAT hits against the UMD3.1 bovine genome assembly or BLAST hits against the NCBI dbEST database. Despite no sequence-based evidence that these seven negative controls inadvertently target transcripts, due to the high level of mRNA expression detected by these probes, they are not suitable as negative controls. Analysis of the series of 60 mismatch probes also provided a measure of specificity of the hybridization reaction. Figure 5 shows the decline in detected expression relative to the 0 mismatch probe as the number of mismatches in each probe sequence increases. As expected, detected expression decreases as the number of mismatches in the probe sequence increases.

The location of a probe in relation to the 3’ end of a transcript has been shown to be related to the detected expression intensity presumably due to the premature termination of reverse transcription using poly-dT primed reactions [16]. A series of distance controls was also included on the bovine array to allow quantitation of this effect. Twenty one of the 60 mismatch control genes for which the RefSeq sequence was greater than 1800 bases were selected for the design of distance controls. For each sequence, four probes were designed with the first probe located within 500 bases of the 3' end, the second probe within the region 500-1000 bp from the 3’ end, and so on. An additional 1,740 cDNAs have between 2 and 7 probes mapped to them. To determine the effect of probe distance from the 3' end of the transcript, expression levels detected by each of these probes were compared. For each adjacent probe pair, the mean decrease in signal (as a percentage, using the raw data from 144 measurements) between the two probes was calculated. The percent signal decrease was divided by the number of bases separating the probes to normalize for the distance between the two probes. The per-nucleotide percent decrease for all probe pairs was averaged to obtain an estimate of the effect of probe location on signal intensity. Figure 6 shows that the majority of probe pairs lay between 500 and 2500 kb of the 3' end of their transcript, and that signal intensity drops between 6 and 15 percent over that range. Therefore probes far from the 3’ end of transcripts will systematically detect lower levels of expression than will probes lying near the 3’ end of any transcript (Nielsen *et al*. 2003). However, the relatively small decrease in detected intensity should have only a marginal effect on the ability of the array to detect transcripts, and the relative differences in expression between cDNA samples should still be proportional to gene expression. Nevertheless, when conducting follow-up qRT-PCR validation of microarray results, it is important to design primers in the same location as the probe for the most reliable replication of the microarray results, but near the 3’ end for the most accurate measurement of transcript abundance.

The same 36 array experiment was completed at two different labs by different personnel using slightly different techniques allowing the measurement of the overall variability present in expression measurements using the array. Principal component analysis of the data revealed that 46.56% of the total variance was assigned to site and site*dye effects indicating that all sources of variation introduced in an off-site replication of an experiment are significant (Figure 7). However, these effects can be modeled using an appropriate linear model when analyzing the data.

Table 1. BOM probes detecting expression by tissue

Tissue	Probes Detecting Expression	Probes Detecting Tissue Specific Expression
Adrenal Gland	13,462	681
Anterior Pituitary	12,793	587
Liver	11,959	729
Small Intestine	8,240	496
Spleen	12,861	719
Thymus	10,385	441
Any	17,648	
All	6,799	

“Any” refers to the number of non-redundant genes detected as being expressed in at least one tissue. “All” refers to the number of genes detected as being expressed in all tissues with a positive false discovery rate (pFDR) of 0.01. “Tissue specific” refers to genes detected as being expressed in only one tissue with a pFDR of 0.0001.

A total of 17,648 probes (74%) were determined to detect gene expression in at least one of the six tissues as summarized in Table 1. Expressed genes were determined using a two-sample t-test, where data from each unique spot × tissue combination were tested against the mean of the negative control probes. Tissue specific genes were defined as those genes expressed in only one tissue with a pFDR of 0.0001. The observed range of 1-3% of probes which exhibit tissue-specific expression and 15% of probes which detect expression in a specific tissue is consistent with previous studies in human [17, 18]. When data from the two locations were separately analyzed, the probes identified as detecting expression were similar, with a 77% overlap (Figure 8). To further demonstrate the specificity of the array, the 729 probes that detected expression only in liver were selected for network analysis using Ingenuity Pathways Analysis (IPA). IPA identified several networks of interacting genes which included a significant number of genes involved in liver-specific functions. An example is the drug and lipid metabolism network shown in Figure 9.

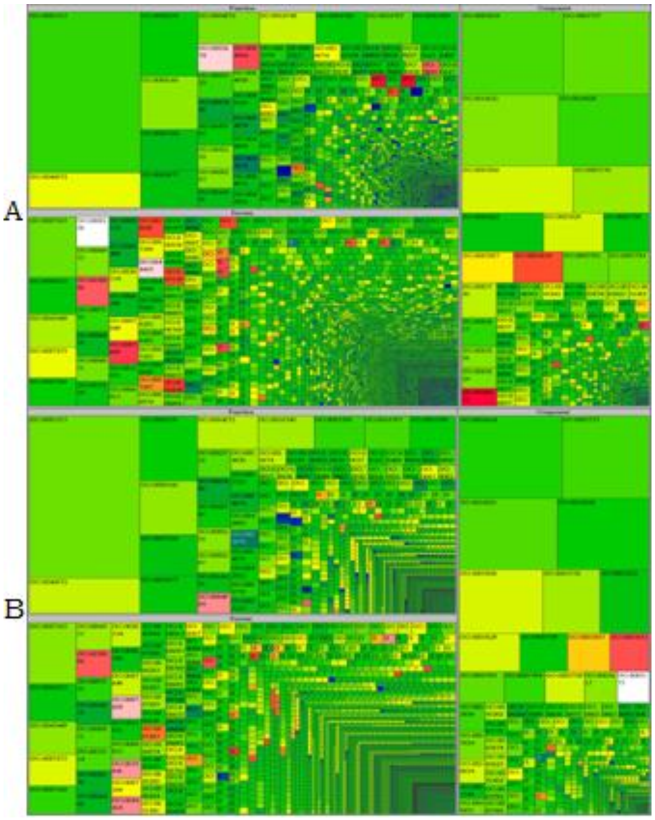


Figure 1: Ontological coverage of: a) the BOM, and b) SPAM oligonucleotide microarrays. The classes of proteins represented by oligonucleotides were analyzed by comparing the proportion of bovine Gene Ontology (GO) terms connected to the BOM and SPAM targets. All bovine GO terms were extracted from NCBI and are displayed using Treemaps in 3 main ontological classes. The number of bovine genes per GO term is proportional to block size. The ratio of BOM or SPAM representation for each GO term is indicated by color; black = no probes, from white (50-fold lower) to blue (10-fold lower) indicates under-representation, green indicates equal representation, from yellow (10-fold higher) to red (≥50-fold higher) indicates over-representation.

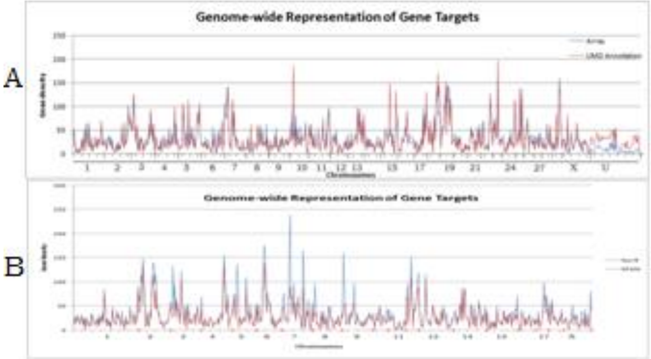
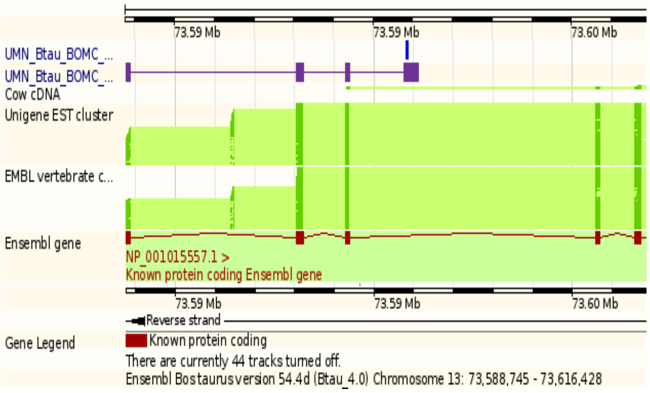


Figure 2: Genome-wide representation of oligonucleotide target sequences on the BOM and SPAM. The positional distribution of targets is plotted relative to: a) bovine (BOM), or b) porcine (SPAM) chromosomes to demonstrate the genome-wide representation of targets on the microarray.



There are currently 44 tracks turned off.

Ensembl Bostaurusversion 54.4d (Btau_4.0) Chromosome 13: 73,588,745 - 73,616,428

Figure 3: A portion of the *HNF4a* gene as displayed by the Ensembl Genome Browser. The consensus sequence for a probe aligns (top track) to three exons and a portion of a predicted intron. The probe sequence aligns (second track) to the intron indicating that the probe detects expression of a previously unannotated splice form of the gene, an observation verified by reverse transcription PCR.

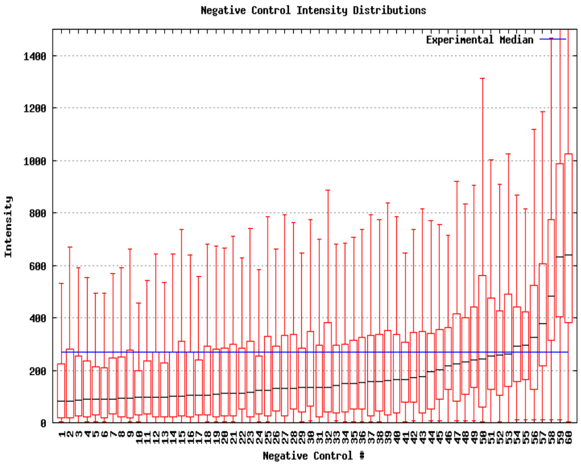


Figure 4: Signal distribution of BOM negative controls. The median signal intensity for negative controls was calculated and compared to the average signal from non-controls (horizontal line at 250). The minimum and maximum intensity values for each negative control (vertical lines), the first and third quartiles (boxes which contain 50% of the values), and the median (the plus sign) are presented. Summary of results from 72 hybridizations representing 6 tissues from 6 animals, with two technical replicates at two locations is presented.

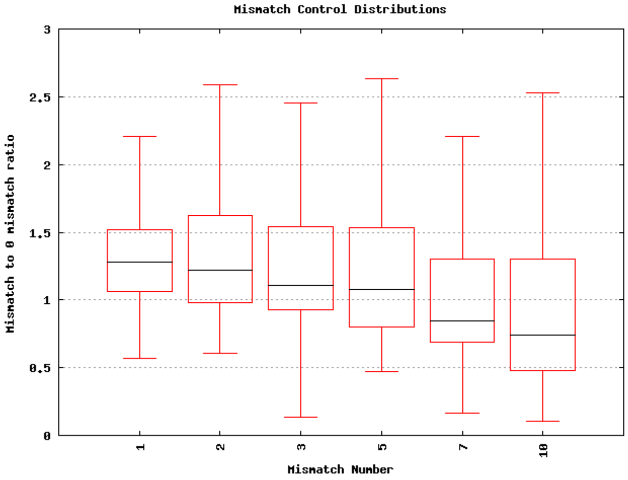


Figure 5: Differential signal detection from mismatch-target oligonucleotides on the BOM array. The normalized signal intensity from target oligonucleotides with 1, 2, 3, 5, 7, and 10 mismatches are presented as a percentage of the log-intensity for their respective perfect match target oligonucleotides.

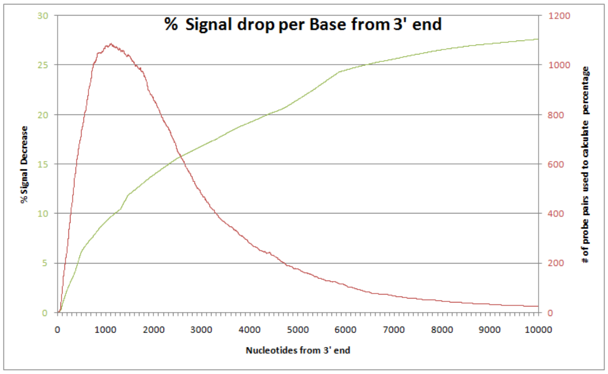


Figure 6: Effect of probe location on intensity. Correlation between log-intensities (light with maximum on the right) from multiple target oligonucleotides predicted to lie within the same gene according to distance of the target from the 3' transcript end. The line with maximum on the left shows the number of probe pairs used to estimate the signal drop.

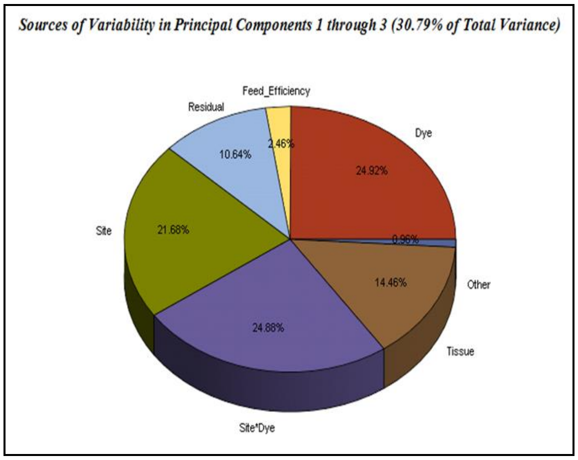


Figure 7: Sources of variation in a 36 slide microarray experiment replicated at two locations.

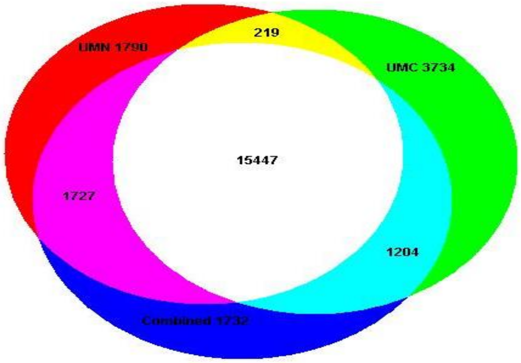


Figure 8: **Overlap between probes detecting gene expression from the analysis of the UMN, MU and combined UMN-MU data.** When the UMN and MU data were separately analyzed 77% of the probes identified as detecting gene expression were common between both data sets. When the data sets were combined an additional 1,732 probes (7%) were identified as detecting gene expression.

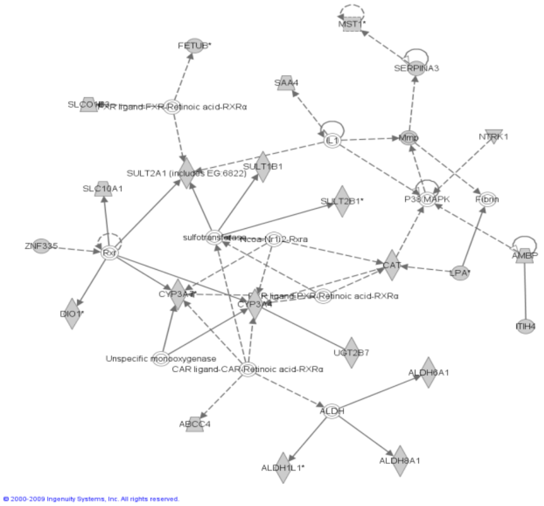


Figure 9: Protein interaction network.

A network identified by Ingenuity Pathway Analysis of 23 liver-specific genes involved in small molecule biochemistry, drug metabolism, and lipid metabolism.

3 Aim 2: Functional annotation of the bovine genome using microarrays and RNA-Seq

3.1 Feed Conversion Efficiency

3.1.1 Introduction

Feed intake and feed efficiency are economically important traits in the beef cattle industry. Feed costs were reported as being at least 60-65% of the total cost incurred in feedlot cattle by Arthur *et al.* [19]. An important measure of feed efficiency is residual feed intake (RFI), defined as the difference in actual feed intake and the expected feed requirements for maintenance of body weight and for weight gain of each animal [20]. Recently, RFI was reported by Herd *et al.* [21] to have great potential as an index of efficiency for beef cattle [21-23] since it is moderately heritable (h^2 = 0.16 to 0.43). Putative QTL influencing RFI were detected on chromosomes 1 (90 cM), 5 (129 cM), 7 (22 cM), 8 (80 cM), 12 (89 cM), 16 (41 cM), 17 (19 cM), and 26 (48 cM) in across-family analyses using microsatellite and SNP genotyping by Nkrumah *et al.* [24]. In the largest and most recent studies Bolormaa *et al.* and Barandse *et al.* used the 10K Affymetrix (Santa Clara, CA) and the 50K Illumina (San Diego, CA) SNP chips to detect RFI associated genomic regions in a variety of beef breeds [25, 26].

The BOMC microarray, comprised of 24,000 long oligonucleotide probes [27], was implemented to identify gene expression profiles in six tissues associated with animals with differing RFI. In addition two RNA-Seq analyses were performed to increase the accuracy in gene expression estimation and determine the concordance in gene expression measurement between microarrays and RNA-Seq experiments.

3.1.2 Methods

Three gene expression experiments were carried out as summarized in Table 2, one using microarrays and two using RNA-Seq.

Microarray Tissue Samples: Six bovine tissue samples (small intestine, spleen, liver, adrenal gland, anterior pituitary, and thymus) were collected from each of 6 Angus steers at 14 months of age that had been selected from among 288 individually fed animals (96 animals in each of three years) for high, intermediate, and low RFI (two animals per group) at the University of Missouri-Columbia (MU). The tissue samples were immediately frozen on dry ice and stored at -80 °C prior to RNA extraction. RNA was extracted and cDNA synthesized at MU and aliquots of dye-labeled cDNA samples were used to replicate all hybridizations at the University of Minnesota (UMN) and at MU. At each location, samples were hybridized to 36 microarrays using a loop design, shown in table 3, with like tissues hybridized to the same array and with duplicate samples labeled with Cy3 and Cy5 as technical replicates. RNA extraction and hybridization was carried out as previously described.

Microarray Data Analysis: The arrays were scanned on an Axon Genepix 4000B laser scanner at 532 nm for Cy3 and at 635 nm for Cy5. The image data were quantified and graded using BlueFuse software, and then filtered (quality =1 and confidence >0.1). The data were analyzed in JMP Genomics 4.0 [28]. The data were normalized with a $\log_2$

transformation, within slide Loess normalization[29], and standardization of each channel to a mean of zero and a standard deviation of one. Differentially expressed (DE) genes were identified using a mixed model with site, feed_efficiency, tissue, dye, site*dye, and feed_efficiency*tissue as fixed effects and array and animal as random effects. The least squares means for feed_efficiency*tissue for each gene were computed and sliced tests for each tissue were performed to determine DE genes for each tissue type. An F-test was used to identify genes with differing expression profiles across the three RFI levels. A false positive discovery rate (q-value) multiple testing correction was employed at a significance level of 0.05 [30].

RNA-Seq 1: RNA samples from the livers of six animals in three feed efficiency groups (the same animals as those used in the microarray experiment) were sequenced with an Illumina Genome Analyzer generating an average of 26 million 32 bp reads per sample. The sequence reads were aligned to theUMD3.1 build of the bovine genome using Bowtie and Tophat [31, 32]. Transcript abundance estimation was performed using Cufflinks [33] with the UMD3.1 NCBI gene build used as the reference transcriptome. A pFDR corrected (p = 0.05) T-test was used to identify differential expression between the three pairs of RFI levels.

RNA-Seq 2: RNA samples extracted from the small intestine, liver, and rib-eye muscle of eight animals in two feed efficiency groups (high and low) were sequenced with an Illumina Genome Analyzer generating an average of 30.9 million 80 bp reads per sample. A pFDR corrected (p = 0.05) T-test was used to identify differential expression between the high and low RFI groups.

Table 2: **RFI expression analysis details**

Platform	**Microarray**	**RNA-Seq 1**	**RNA-Seq 2**
Depth	144 hybridizations	26M 32 bp reads/sample	30M 80 bp reads/sample
Samples	2 High, 2 medium, 2 low RFI	2 High, 2 medium, 2 low RFI	4 High, 4 low RFI
Tissues	small intestine, spleen, liver, adrenal gland, anterior pituitary, thymus	Liver	Liver, small intestine, rib-eye muscle

Table 3: Microarray hybridization scheme

Slide	**Dye**	**Animal**	**Feed Efficiency**	**Tissue**
1	Cy3	2	High	Small intestine
1	Cy5	5	Low	Small intestine
2	Cy3	4	Middle	Liver
2	Cy5	6	Low	Liver
3	Cy3	5	Low	Small intestine
3	Cy5	4	Middle	Small intestine
4	Cy3	6	Low	Liver
4	Cy5	3	Middle	Liver
5	Cy3	3	Middle	Small intestine
5	Cy5	6	Low	Small intestine
6	Cy3	4	Middle	Adrenal gland
6	Cy5	1	High	Adrenal gland
7	Cy3	3	Middle	Adrenal gland
7	Cy5	6	Low	Adrenal gland
8	Cy3	1	High	Liver
8	Cy5	5	Low	Liver
9	Cy3	2	High	Spleen
9	Cy5	4	Middle	Spleen
10	Cy3	5	Low	Liver

10	Cy5	2	High	Liver
11	Cy3	4	Middle	Spleen
11	Cy5	5	Low	Spleen
12	Cy3	1	High	Spleen
12	Cy5	6	Low	Spleen
13	Cy3	6	Low	Pituitary
13	Cy5	1	High	Pituitary
14	Cy3	3	Middle	Pituitary
14	Cy5	2	High	Pituitary
15	Cy3	4	Middle	Thymus
15	Cy5	1	High	Thymus
16	Cy3	5	Low	Thymus
16	Cy5	4	Middle	Thymus
17	Cy3	6	Low	Thymus
17	Cy5	2	High	Thymus
18	Cy3	1	High	Thymus
18	Cy5	3	Middle	Thymus
19	Cy3	6	Low	Small intestine
19	Cy5	2	High	Small intestine
20	Cy3	2	High	Pituitary
20	Cy5	5	Low	Pituitary
21	Cy3	1	High	Adrenal gland
21	Cy5	4	Middle	Adrenal gland
22	Cy3	5	Low	Adrenal gland
22	Cy5	3	Middle	Adrenal gland
23	Cy3	1	High	Small intestine
23	Cy5	3	Middle	Small intestine
24	Cy3	1	High	Pituitary
24	Cy5	4	Middle	Pituitary
25	Cy3	3	Middle	Liver
25	Cy5	1	High	Liver

26	Cy3	3	Middle	Thymus
26	Cy5	6	Low	Thymus
27	Cy3	5	Low	Pituitary
27	Cy5	3	Middle	Pituitary
28	Cy3	4	Middle	Pituitary
28	Cy5	6	Low	Pituitary
29	Cy3	3	Middle	Spleen
29	Cy5	1	High	Spleen
30	Cy3	6	Low	Adrenal gland
30	Cy5	2	High	Adrenal gland
31	Cy3	5	Low	Spleen
31	Cy5	3	Middle	Spleen
32	Cy3	2	High	Adrenal gland
32	Cy5	5	Low	Adrenal gland
33	Cy3	4	Middle	Small intestine
33	Cy5	1	High	Small intestine
34	Cy3	2	High	Liver
34	Cy5	4	Middle	Liver
35	Cy3	2	High	Thymus
35	Cy5	5	Low	Thymus
36	Cy3	6	Low	Spleen
36	Cy5	2	High	Spleen

3.1.3 Results

Microarray: Analysis of the microarray data resulted in a total of 6,251 experimental probes indicating DE in at least one tissue. A total of 869 probes were significant in the adrenal gland (6.45% of the total number probes expressed in the adrenal gland), 1,857 in the liver (15.5% of the total), 642 in the pituitary (5% of the total), 3,465 in the small intestine (42% of the total), 972 in the spleen (7.5% of the total), and 419 in the thymus (4% of the total) while one probe was significant in all six tissues.

RNA-Seq 1: An average of 80.7% of the reads in each sample could be mapped to the genome. A total of 930 genes were DE in at least one of the three pairwise tests. Figures 10a-c show that the distribution of gene expression is consistent between samples and RFI groups. Figure 10d shows the relationship between DE and significance in a volcano plot. A total of 1,098 genes were DE, 232 between high and medium RFI, 393 between high and low RFI, and 895 between medium and low RFI. More DE genes have reduced expression in low RFI animals. Figure 14 shows the gene expression profiles of genes DE between high and low RFI clustered by profile similarity. Pathway analysis of DE genes using Ingenuity Pathways Analysis (IPA) shows the genes are associated with many important liver processes, including liver fibrosis, renal and liver proliferation, and liver hepatitis. A representative interaction network is shown in Figure 15, where all but one gene had elevated expression in the low RFI animals.

RNA-Seq 2: Average read alignment was 80.3% for liver, 77.3% for small intestine, and 81.1% for rib-eye muscle. Figure 10 shows distributions of gene expression by sample in liver, small intestine, and rib-eye muscle, respectively. The distributions show minimal variation between samples. Figures 11 and 12 show the distributions of gene expression by RFI group, with minimal variation in distributions between groups in the same tissue. The volcano plots in Figure 13 show the relationship between degree and significance of DE. A total of 51 DE genes were identified in liver, all but four of which had higher expression in the low RFI group. A total of 24 genes were DE in rib-eye muscle, all of which had lower expression in the low RFI group. No DE genes were detected in small intestine.

RNA-Seq 1 vs RNA-Seq 2: Gene expression in the liver was strongly correlated between the two RNA-Seq datasets with a Spearman correlation of .943 (Figure 16). Comparing the 393 RNA-Seq 1 LVR genes significant for the high-low contrast with the 51 RNA-Seq 2 LVR DE genes produced an overlap of only 4 DE genes, with the direction of fold change the same for all of them (Figure 17).

Microarray vs RNA-Seq: Figures 18a-f show the correlation between microarray gene expression measurements (both before and after normalization) and RNA-Seq measurements for liver and small intestine. The correlation is poor, with little effect due to normalization of the microarray data. Comparing the median raw (not normalized) RNA-Seq 1 liver microarray expression measurements with the median RNA-Seq measurement produced a Pearson correlation of 0.20 and a Spearman correlation of 0.49 (Fig 19). The same comparison using just the microarray probes determined as being expressed in liver reduced the correlations to 0.12 and 0.40, respectively. A total of 110 liver genes were identified as DE in both the microarray and RNA-Seq 1 analysis, however, there was little correlation in expression change ($r^2 = 0.1375$). A total of 7 liver genes were DE in microarray liver and RNA-Seq 2 analyses, but there was no conservation of fold-change direction.

3.1.4 Discussion

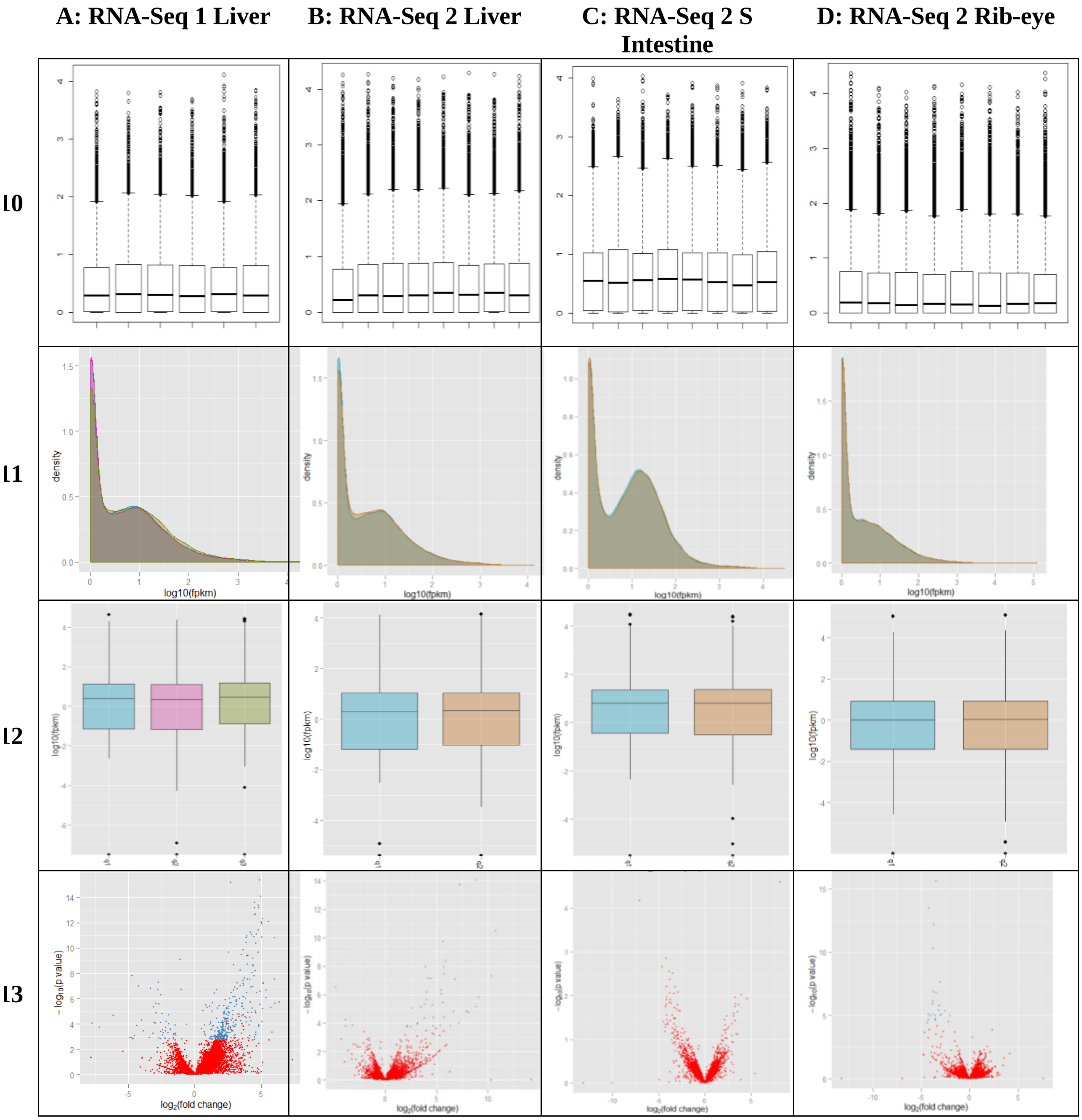
The results show poor correlations between microarray gene expression measurements and RNA-Seq measurements regardless of data normalization, tissue type, or RNA-Seq dataset with correlations no higher than 0.49. In comparison, Marioni *et al.* reported Spearman correlations of between 0.73 and 0.75 when comparing RNA-Seq data generated by Illumina sequencing with Affymetrix microarray data [34]. The two RNA-Seq experiments examined here both have similar percentages of reads mapping to the genome despite the difference in read lengths (32 bp vs 80 bp). The 80 bp reads may have reduced numbers of multi-hit reads, but otherwise the difference in read length has had little effect on gene expression measurement. The longer read length would however be more useful in detecting differential splicing because longer reads are more likely to be successfully aligned across splice boundaries. The two RNA-Seq experiments show high correlation in expression measurements but the identification of DE shows less agreement. The eight animals used in RNA-Seq 2 have a range of feed efficiency, measured as pounds per day, from -2.36 to -1.15 for high RFI animals and 0.99 to 2.56 for the low RFI animals. The animals with more moderate RFI values in each group may be reducing the ability to identify DE between these groups; re-analyzing the data after removing the two animals with the most moderate RFI may increase the number of detected DE genes.

Studies examining hepatic gene expression in relation to feed efficiency have not yet been published so the results presented here cannot be directly compared with previous findings. Genome-wide association studies have been published associationg SNPs with feed efficiency in cattle. Barendse *et al.* identified 161 genomic regions containing 103 genes associated with feed efficiency in a mix of cattle breeds. Comparing these 103 genes with the 5,192 genes identified as DE in three gene expression experiments reveals an overlap of nineteen genes, summarized in Table 4. While the degree of overlap is similar to what would be expected due to random chance, many of the overlapping genes have functions identified by the Swiss-Prot database with relatively direct relationships to growth, nutrient uptake, and health. In the small intestine *ATP1A1* is involved in cross-membrane nutrient transport; *SPA17* is involved in cell-cell adhesion functions such as immune cell migration and metastasis; and *YES1* is a promoter of Neisseria gonorrhoeae in epithelial cells. *HDGF*, which may be involved in the proliferation of smooth –muscle cells, shows DE in three tissues in addition to being associated with RFI by genome-wide association. Together, these genes are promising targets for further investigation, particularly by looking for polymorphisms in the promoter regions of these genes that could be responsible for the observed differences in gene expression.

Table 4: Genes associated with RFI by QTL and gene expression

Gene	Tissue with DE	Description
AFF3	LVR	Lymphoid nuclear protein related to *AF4*
ATP1A1	SI	ATPase, Na+/K+ transporting, alpha 1 polypeptide
HDGF	APIT, SPN, SI	hepatoma-derived growth factor
INPP5D	APIT	inositol polyphosphate-5-phosphatase, 145kDa
KLHDC4	ADR	kelch domain containing 4
LAMC1	SI	laminin, gamma 1
MBNL1	THY, SI	muscleblind-like (Drosophila)
MPPED2	LVR	metallophosphoesterase domain containing 2
NCOA7	LVR	Estrogen nuclear receptor coactivator 1
PSMD13	LVR	proteasome (prosome, macropain) 26S subunit, non-ATPase, 13
RPLP2	RNA-Seq 1 LVR	ribosomal protein, large, P2
SDK1	SPN	sidekick homolog 1, cell adhesion molecule (chicken)
SLC45A2	RNA-Seq 1 LVR	solute carrier family 45, member 2
SPA17	SI	sperm autoantigenic protein 17
SYT9	LVR	synaptotagmin IX
TAX1BP1	SI	Tax1 (human T-cell leukemia virus type I) binding protein 1
UBE2I	APIT, LVR, SI, THY	ubiquitin-conjugating enzyme E2I
WASL	APIT, LVR	Wiskott-Aldrich syndrome-like
YES1	SI	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1

LVR = liver, SI = small intestine, APIT = anterior pituitary, SPN = spleen, THY = thymus, ADR = adrenal gland.



Figures 10-13: Distribution of gene expression by platform, method, and RFI group; volcano plots of differential expression. Figure 10 shows boxplots summarizing the distribution of gene expression measurements ($\log_2$ of FPKM gene expression values as calculated by Cufflinks) by sample. Figure 11 shows kernel density plots of gene expression measurements ($\log_2$ of FPKM+1) by RFI group. Density plots for each group in a tissue should be very similar for robust DE testing. Figure 12 shows boxplots summarizing the distribution of gene expression measurements ($\log_2$ of FPKM) by RFI group. Figure 13 shows volcano plots of DE with significant genes indicated with a lighter shade.

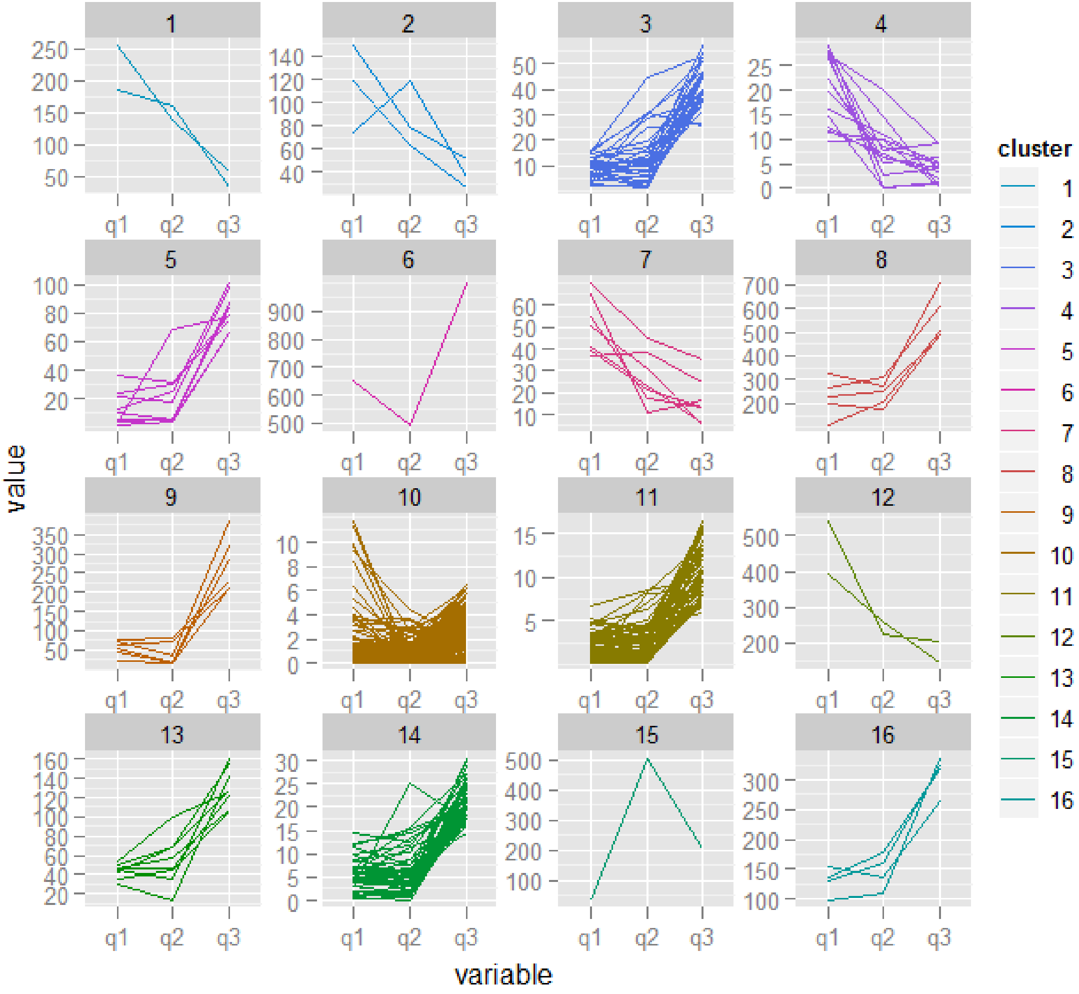


Figure 14: RNA-Seq 1 liver gene expression profiles clustered by similarity. Genes with similar expression changes across RFI Group (q1=high, q2=medium, q3=low) are grouped into 12 clusters. The shown genes were DE between high and low RFI animals.

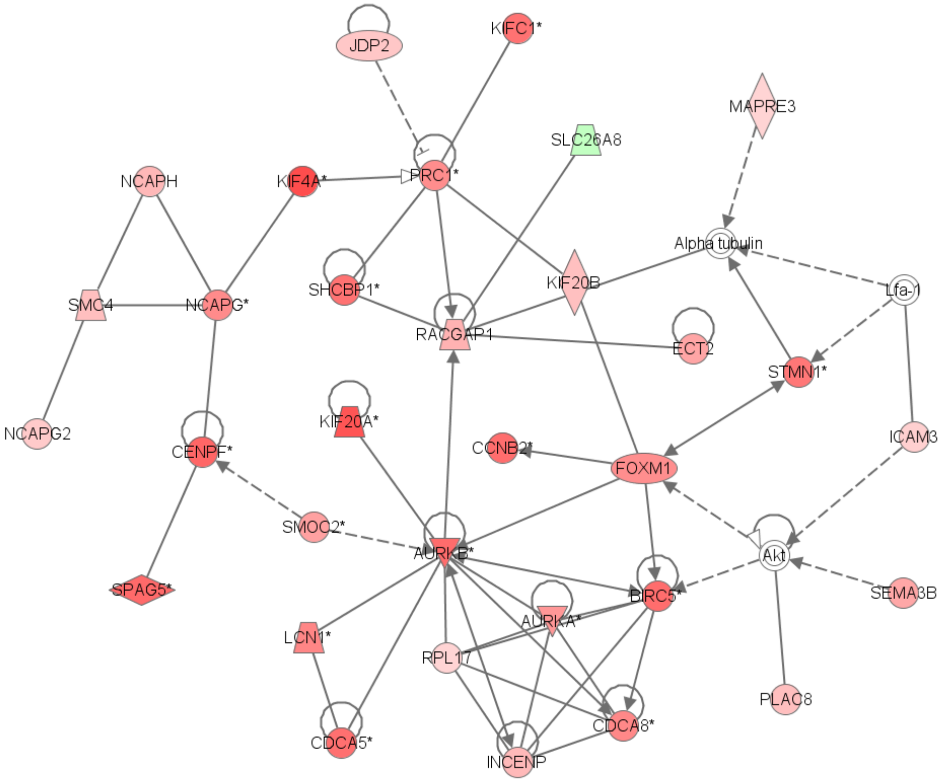


Figure 15: Differentially expressed genes identified by RNA-Seq in a cell cycle network. A functional network involved in the cell cycle, cellular movement, and cellular assembly and organization. Genes with elevated expression in low RFI animals are shaded with the exception of SLC26A8, which has reduced expression.

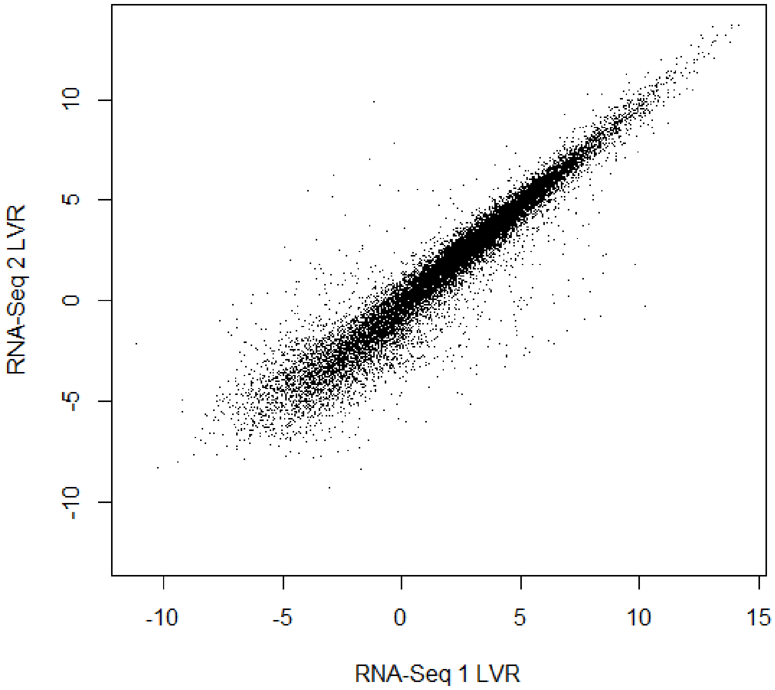


Figure 16: Correlation of expression between liver RNA-Seq experiments. Gene expression estimates are highly correlated (Spearman coefficient=0.943) between two different RNA-Seq experiments measuring gene expression in liver.

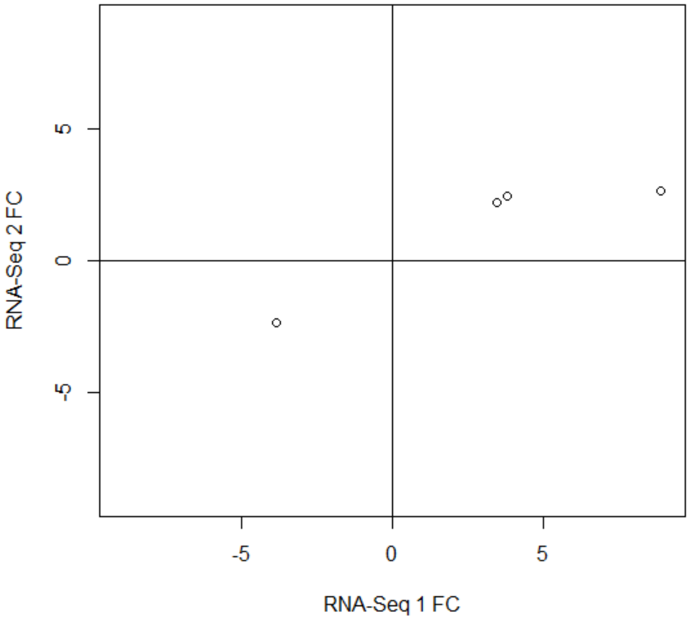


Figure 17: Correlation of fold-change between DE genes in two liver RNA-Seq experiments. Gene expression fold-change direction is conserved in all four genes with DE in two different RNA-Seq experiments.

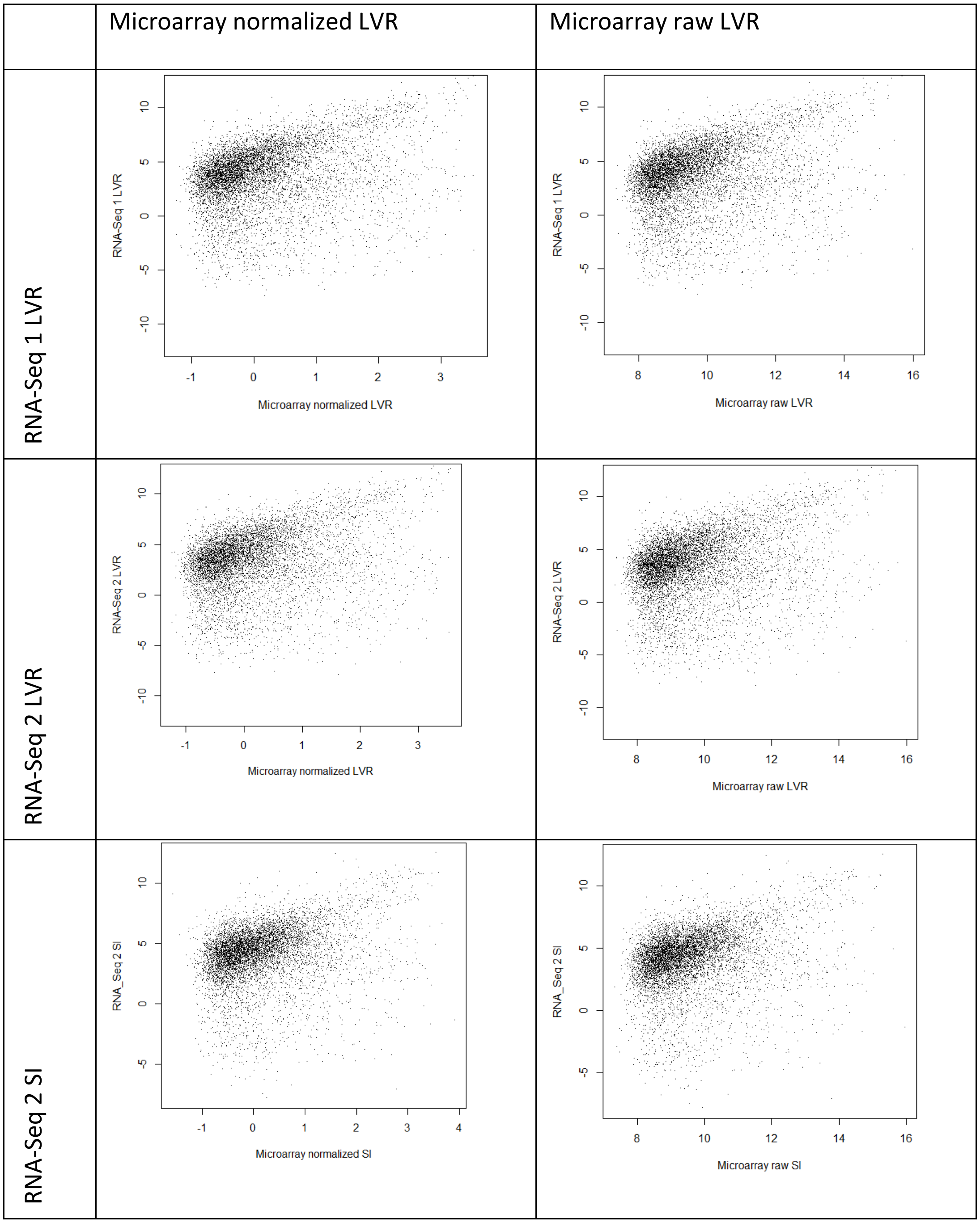


Figure 18: Correlation between microarray and RNA-Seq gene expression measurements. There is a poor correlation between microarray and RNA-Seq measurements regardless of tissue, RNA-Seq experiment, or microarray data normalization method.

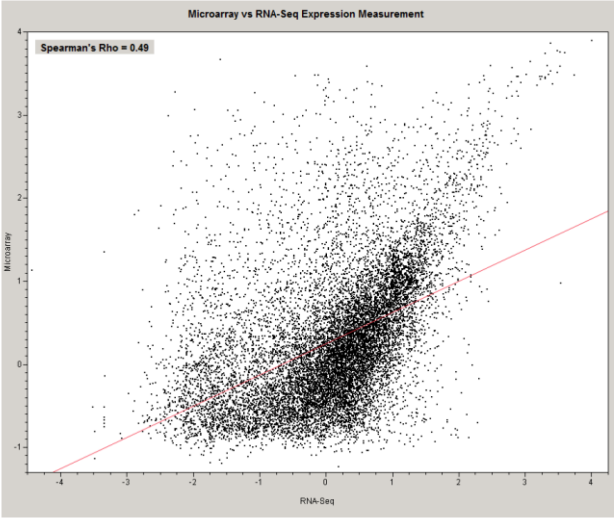


Figure 19: A comparison of microarray and RNA-Seq experimental measurements of gene expression in liver. The expression measurements between the two platforms have a Spearman correlation of 0.49. There is a weak correlation for transcripts with low RNA-Seq coverage, and stronger correlation for transcripts with higher RNA-seq coverage.

3.2 Milk Yield

3.2.1 Introduction

The University of Minnesota maintains a closed herd of Holsteins that has been randomly bred since 1964. This control line (CL) represents dairy cattle genetics as they were in 1964. A contemporary selected line (SL) of Holsteins derived from the same population as the control line has been selectively bred for increased milk yield since 1964. Over 45 years of intensive selection has resulted in cows that produce two and a half times more milk than their ancestors. These cows differ substantially in feed intake and in the partitioning of nutrients among body tissues. High merit cows produce more milk, have greater voluntary feed intakes and use more of their body reserves in early lactation than do cows of low genetic merit for milk production. These control and selected lines provide a powerful resource for identifying the genes responsible for the large physiological differences between the two lines. A previous gene expression study by Loor *et al.* examined the changes in hepatic gene expression over time during periparturition [35] in contemporary cows. The study used a previous generation cDNA microarray with probes targeting the detection of 7,872 expressed sequences. Several association studies have identified genomic regions associated with production traits in high merit Holstein cattle. The most comprehensive of these, by Cole *et al.*, used the Illumina BovineSNP50 SNP chip to identify SNPs associated with thirty-one production, health, reproduction, and body conformation traits [36]. It is hypothesized that profiling the genes expressed in liver tissue during the transition from pregnant and non-lactating to non-pregnant and lactating will reveal genes responsible for differences in metabolism, energy balance and nutrient partitioning.

3.2.2 Methods

Tissue Samples: Liver biopsies from primiparous and multiparous CL and SL cows (n = 32, 8 per linexparity combination) were collected at -9, 21, and 70 days postpartum. Total RNA was extracted from each tissue sample. Within each of the four linexparity combinations, RNA samples from the eight cows were pooled in pairs to create four samples at each time point as shown in Table 5. For the prepartum timepoint, RNA from only one of the two cows was used due to the need to reduce variation in days between sampling and calving. RNA extraction and hybridization was carried out as previously described.

Table 5. Liver RNA Sample pooling

Multi_Select_-14	Cows	**Multi_Select_21**	Cows	**Multi_Select_70**	Cows
MSA-1	9410	MSB-1	9410, 9621	MSC-1	9410, 9621
MSA-2	9504	MSB-2	9504, 9513	MSC-2	9504, 9513
MSA-3	9527	MSB-3	9527, 9349	MSC-3	9527, 9349
MSA-4	9528	MSB-4	9528, 9615	MSC-4	9528, 9615
Multi_Control_-14	Cows	**Multi_Control_21**	Cows	**Multi_Control_70**	Cows
MCA-1	9516	MCB-1	9516, 9557	MCC-1	9516, 9557
MCA-2	9522	MCB-2	9522, 9532	MCC-2	9522, 9532
MCA-3	9607	MCB-3	9607, 9427	MCC-3	9607, 9427
MCA-4	9506	MCB-4	9506, 9626	MCC-4	9506, 9626
Primi_Select_-14	Cows	**Primi_Select_21**	Cows	**Primi_Select_70**	Cows
PSA-1	9705	PSB-1	9705, 9658	PSC-1	9705, 9658
PSA-2	9716	PSB-2	9716, 9724	PSC-2	9716, 9724
PSA-3	9739	PSB-3	9739, 9734	PSC-3	9739, 9734
PSA-4	9741	PSB-4	9741, 9726	PSC-4	9741, 9726
Primi_Control_-14	Cows	**Primi_Control_21**	Cows	**Primi_Control_70**	Cows
PCA-1	9713	PCB-1	9713, 9718	PCC-1	9713, 9718
PCA-2	9715	PCB-2	9715, 9742	PCC-2	9715, 9742
PCA-3	9723	PCB-3	9723, 9657	PCC-3	9723
PCA-4	9738	PCB-4	9738, 9722	PCC-4	9738, 9722

There are four sample pools for each of the twelve time x line x parity combinations. Most sample pools were created from samples from two cows, however all -14 d time-point pools were from one cow, as is pool PCC-3, which is due to the death of cow 9657.

Data Analysis: The microarray image data were quantified and graded using BlueFuse software, and then filtered (quality =1 and confidence > 0.1). Data normalization was accomplished by $\log_2$ transformation, print-tip LOESS normalization, and standardization to a mean of zero and standard deviation of 1. Differentially expressed genes were identified using a mixed model ANOVA. Fixed effects modeled were dye, line, time, parity, and the two- and three-way interactions between time, line, and parity. Variability was adjusted for the random effects of array, batch, and technician. An F-test was used to test for changes in expression across the levels of each effect (time, line, parity, and their interactions). Correction of the P-values for multiple testing was performed using the pFDR method using a p-value cutoff of 0.05.

3.2.3 Results

Numbers of probes detecting differential expression (DE) are summarized in Table 6. Between 926 and 1,276 probes detect DE across Line, Time, and Parity. Fewer probes detect DE for the interaction effects.

Table 6. Differentially expressed genes between lines, parity classes and time points

Effect	**# Significant Probes**
Line	926
Time	991
Parity	1,276
Line*Time	290
Line*Parity	483
Time*Parity	651
Line*Time*Parity	261

3.2.4 Discussion

The analysis did not consider the variable number of animals contributing to each sample pool. Variation in the number of animals in a pool creates variation in the variance of observed expression between samples. Sample pools created from one animal will have more observed expressed variance than sample pools created from two animals, which affects the calculation of the significance of DE. An ANOVA taking into account the heterogenous variation in pool sizes would generate morerobust results and may alter the lists of significant DE genes summarized here. The failure to account for pool sizes most affects the test for significance affects across time as the pool sizes for all time point -14 d samples are different from samples at time points 21 and 70 d.

The use of alternative gene expression measurement methods such as qRT-PCR or RNA-Seq analysis to validate these results would be preferred but was not undertaken due to time and cost constraints. Instead, the results are compared to previously reported results. In a study using a 7,972 probe cDNA microarray Loor *et al.* identified 57 DE genes in the liver across seven time points (–65, –30, –14, +1, +14, +28, and +49 days) relative to parturition [35]. There are two genes, interleukin 27 receptor alpha (*IL27RA*) and neuroblastoma suppression of tumorigenicity 1 (*NBL1*) with DE in both Loor's and this study. The different time points used by Loor *et al.* make it difficult to make direct comparisons between the observed expression profiles for these genes. In a separate study, Cole *et al.* conducted a genome-wide association study to identify SNPs associated with thirty-one dairy traits in contemporary Holstein cows [36]. Comparing the genomic locations of the twenty most significant SNPs for each trait with the locations of genes with DE across line identified five genes containing highly significant SNPs, all with higher expression in SL animals (Table 47). These genes present ideal targets for further study. The reproductive traits of CL cattle are superior to those of SL cattle, therefore some of the genes whose expression profiles at periparturition differ between CL and SL cows could potentially play a role in calving ease. There is, however, no overlap between these genes and regions associated with calving ease identified by Cole *et al*.

Table 7 DE genes associated with production traits by GWAS

Gene	Description	Associated trait
AFF2	Fragile X E mental retardation syndrome protein	strength
GRIA3	Glutamate receptor ionotropic, AMPA 3	daughter pregnancy rate
LAMP2	Lysosome-associated membrane protein 2	protein yield, lifetime net merit
PGLYRP1	peptidoglycan recognition protein 1	fat yield, protein yield, service-sire calving ease, daughter calving ease, lifetime net merit, productive life
RPL37	60S ribosomal protein L37	protein %

4 Aim 3: Bovine SNP classification and functional annotation

4.1 Background

Single nucleotide polymorphisms (SNPs) are single base changes in a nucleotide sequence. Variation within the bovine genome occurs most frequently as SNPs (although structural rearrangements affect a higher percentage of the genome [37]). SNPs occur on average every 285 bases in indicine breeds and 714 bases in taurine breeds, giving the bovine genome an estimated four to ten million SNPs [38]. SNPs are easy to genotype in an automated fashion and occur at a high density in the genome which makes them ideal for use in genome-wide association studies. Following a genome wide association study or gene expression experiment, researchers need to understand the biological mechanisms which cause the association between the observed phenotype and the genetic variation; thus identifying the annotation of the studied SNPs is an important component. Researchers faced with a long list of SNPs or genes statistically associated with a phenotype need to be able to narrow this list and focus on those SNPs or genes which are most likely to be causally related to the phenotype.

Although millions of SNPs exist, the vast majority likely have no functional effect on any phenotype. SNPs located in or near protein coding regions may be more likely to have function effects because they may cause changes in promoter binding sites, amino acid coding, or exon splicing sites. Categorizing SNPs by their location relative to genes identifies this important SNPs. SNP locations fall into eight categories: intronic, 5’UTR, 3’UTR, 5’-upstream, 3’-downstream, splicing site or coding or non-coding. SNPs in coding regions can be further categorized as synonymous or non-synonymous. Regardless of the location of the SNP in question, predicting function is difficult.

The Single Nucleotide Polymorphism Database (dbSNP) is a public database that serves as the primary archive for genetic variation within and across species [39]. Previous bovine SNP discovery efforts have identified over 2.2 million bovine SNPs in dbSNP, 677,000 of them in genes, and almost 7,000 produce a change in the amino acid sequence. A public consortium has recently sequenced the genomes of over one hundred bovine genomes from eleven different breeds, each to a depth of approximately 1X, as part of a SNP discovery project. The indicine breeds sequenced were Brahman, Gir, Nelore, and Sahiwal. The taurine breeds sequenced were Angus, Holstein, Jersey, Limousin, Fleckvieh, Romagnola, and N’Dama. This re-sequencing has identified 48,630,857 biallelic SNPs in the bovine genome. These data have dramatically increased the number of known SNPs and have uncovered a larger percentage of the SNPs likely to possess functional effects on phenotypes.

The current bovine SNP annotation consists of categorization by location and, in coding regions, by whether they are synonymous or non-synonymous. However, the potential for the alleles of these SNPs to have functional effects has not been predicted. Association studies and gene expression experiments highlight short genomic regions or specific genes of interest, but these small segments of DNA still contain large numbers of SNPs. Adding SNP function predictions will highlight which SNPs are most likely to have causative effects on the studied trait. Comparing genes and SNPs associated with productions traits correlated with annotated SNPs will identify candidate SNPs ideal for further study and verification.

4.2 Materials and Methods

Data on SNP location, quality, and breed-frequency for 48,620,857 putative SNPs were obtained from the USDA Bovine Functional Genomics Lab. The SNPs were functionally annotated using ANNOVAR, a genetic variant annotation program that is sufficiently fast to annotate millions of SNPs on a desktop computer [40]. SNPs were annotated in reference to the UMD3.1 build of the bovine genome and the NCBI generated gene build [5]. SNP data provided by the UDSA Bovine Functional Genomics Group is comprised of UMD3.1 genomic coordinates and the reference and alternate nucleotides for 48,630,857 biallelic SNPs. The format of the SNP data was converted to the ANNOVAR input format using a custom Perl script. The UMD3.1 gene annotation was downloaded from ftp://ftp.cbcb.umd.edu/pub/data/Bos_taurus/Bos_taurus_UMD_3.1/annotation/. This annotation was converted from GFF2 format to ensGene format using a custom Perl script. Due to the format conversion, 184 transcript annotations were lost out of a total of 22,760.

4.3 Results and Discussion

To benchmark the SNP annotation produced by ANNOVAR all annotation for SNPs on chromosome 10 was downloaded from dbSNP. These 317,719 SNPs were annotated using ANNOVAR, and the results are shown for the comparison with the annotation from dbSNP in Table 8. The table shows that ANNOVAR generally identifies more SNPs per category than does dbSNP. Manual checking of disagreements between dbSNP and ANNOVAR showed that dbSNP often has no annotation for SNPs that ANNOVAR appeared to have correctly annotated. This comparison provided sufficient evidence that ANNOVAR correctly generated correct SNP annotations to warrant further work, although the precise cause of the discrepancy between the dbSNP and ANNOVAR annotation remains unknown.

Table 9 lists the number of SNPs in each classification category. Note that there are more nonsynonymous SNPs than synonymous, a ratio also found in the dbSNP classifications and in the human 1000 Genomes Project data. In order to examine their potential functional consequences, the amino acid substitutions caused by non-synonymous SNPs (nsSNPs) were scored using the BLOSUM80 scoring matrix (Table 10). The majority of nsSNPs have negative values because the BLOSUM matrix assigns negative scores to most substitutions negative scores. Figure 20 compares the distribution of substitution scores in this bovine dataset with the human 1000 Genomes Project data set [41]. The distributions are very similar with the exception that the bovine data has twice as many -2 scored substitutions as does the human data. To determine specifically which substitutions are over or underrepresented, the frequency of amino acid substitutions (AAS) for each of the 150 possible substitutions was calculated and is displayed in Figures 21 and 22.

Table 8: Summary of SNP annotation produced by ANNOVAR and dbSNP for 317,719 SNPs on chromosome 10.

SNP Class	**ANNOVAR**	**dbSNP**
Intronic	169,807	107,733
Synonymous	2,355	593
UTR3	1,102	690
UTR5	277	209
Upstream	2,060	1863
Downstream	1,993	7,516
Stopgain	28	4
Nonsynonymous	1,253	618
exonic;splicing	35	0
Splicing	51	0
Stoploss	7	0

Table 9: SNP annotation summary for 24.7 million SNPs.

Class	**SNP Count**
Intergenic	14,504,270
Upstream (1K from gene)	122,854
Downstream (1K from gene)	134,021
Upstream; downstream	3,027
UTR3	68,414
UTR5	10,012
UTR5; UTR3	32
Intronic	9,702,509
Exonic	165,635
Splicing	1,451
Exonic; splicing	2,208
Nonsynonymous SNV	74,874
Stopgain SNV	1,523
Stoploss SNV	574
Synonymous SNV	90,856

Categories for SNPs in protein coding regions are shaded.

Table 10: Summary of BLOSUM AAS scores for 219,538 nonsynonymous SNPs.

BLOSUM Score	**SNP count**
-4	5,143
-3	11,664
-2	4,637
-1	11,554
0	16,541
1	17,161
2	3,680
3	4,494

SNPs causing multiple AA changes are only represented once in this chart.

4.3.1 nsSNP Density

The number of nsSNPs per gene was counted and the ten genes with the most nsSNPs are shown in Table 11. These genes typically encode very large proteins and are not necessarily unusually polymorphic. The number of nsSNPs per base of cDNA was calculated to determine the overall density of nsSNPs per gene. Figure 23 plots the distribution of nsSNP density. There is a large difference in nsSNP density between genes, ranging from one nsSNP every 20 bases to one nsSNP every 10,000 bases. Tables 12 and 13 list the 10 genes with the highest and lowest densities of nsSNPs, which could be interpreted as the genes either most and least tolerant to mutation, or most and least susceptible to mutation. The overlap between copy number variations (CNV) in the genome with genes with high nsSNP density was determined to identify genes that may appear to have high nsSNP density due to gene duplication. Table 14 lists the highest density genes not within known CNV. The *H2B* gene is highly conserved across species and is intolerant to mutations, but has many variants in the genome. The presence of *H2B* in this list is likely due to errors in mapping sequence data to the genome resulting in the incorrect identification of nsSNPs. The 1176 genes with a nsSNP density of at least 1:100 were submitted to pathway analysis using Ingenuity Pathway Analysis (IPA). As expected, this gene list is highly enriched for genes associated with recognition of and response to foreign antigens, in particular natural killer cell signaling, EIF2 signaling, graft-versus-host disease signaling, and lipid antigen presentation by CD1. The gene list was also analyzed using DAVID, which showed that the list was enriched for genes associated with olfactory receptor activity, the G-protein coupled receptor protein signaling pathway, cell surface receptor linked signal transduction, and membrane proteins. The 713 genes with a nsSNP density lower than 1:3,000 were also analyzed using IPA and DAVID. This gene list was not as easily categorized as it contained genes with diverse functions, however IPA shows that many are associated with signaling pathways and zinc finger proteins.

Table 11: The ten genes with the most nonsynonymous SNPs.

Gene Name	AAS Count	cDNA Length (bp)	Gene Description
LOC508070	325	7,544	Interferon-induced very large GTPase 1-like
MUC16	170	15,996	mucin 16, cell surface associated
LOC100298796	130	19,924	hypothetical
FREM3	115	6,450	FRAS1 related extracellular matrix 3
LOC789503	114	21,763	mucin protein
TTN	105	101,600	titin
LOC100337244	103	3,423	ATP-binding cassette protein C4-like
LOC100298353	91	11,276	similar to Mucin-5B precursor
LOC751803	89	4,343	WC1 isolate CH149 isoform 2
LOC786060	88	4,665	WC1-like

Table 12 The ten genes with the highest density of nonsynonymous SNPs.

Gene Name	AAS Count	cDNA length	Bases per nsSNP	Description
ICAM2	66	889	13.47	Intercellular adhesion molecule 2
MIR2284P	4	77	19.25	micro RNA
OAS1	53	1033	19.49	2'-5'-oligoadenylate synthetase 1

ABCC4	32	630	19.69	ATP-binding cassette, sub-family C
GIMAP7	74	1475	19.93	GTPase, IMAP family member 7
LOC100298822	40	831	20.78	MHC class I heavy chain-like
LOC508070	325	7544	23.21	Interferon-induced very large GTPase 1-like
LOC614091	47	1133	24.11	MHC class I heavy chain-like
LOC100336101	12	294	24.5	hypothetical protein
MIR2284K	3	74	24.67	micro RNA

Table 13: The ten genes with the lowest density of nonsynonymous SNPs.

Gene Name	**AAS count**	**cDNA length**	**Bases per nsSNP**	**description**
USP9X	1	11,477	11,477	ubiquitin specific peptidase 9, X-linked
ANKRD17	1	10,596	10,596	ankyrin repeat domain 17
LOC100336965	1	9,919	9,919	odz, odd Oz/ten-m homolog 1
ODZ2	1	9,681	9,681	odz, odd Oz/ten-m homolog 2
NBEA	1	9,057	9,057	neurobeachin
CELSR2	1	9,055	9,055	cadherin, EGF LAG seven-pass G-type receptor 2
FBN1	1	8,947	8,947	fibrillin 1
DOCK3	1	8,518	8,518	dedicator of cytokinesis 3
DIP2B	1	8,497	8,497	DIP2 disco-interacting protein 2 homolog B
LOC536240	1	8,441	8,441	HECT domain containing 1

Table 14: The ten genes with the highest density of nonsynonymous SNPs that are not in known CNVs.

Gene Name	Transcript ID	cDNA length	AAS count	Bases per nsSNP
LOC100336101	XM_002695470	294	18	16.33
LOC785910	XM_002697404	927	56	16.55
LOC100337168	XM_002690616	534	26	20.53
LOC783151	XM_002697317	837	40	20.92
LOC100296830	XM_002690614	396	18	22
LOC100336589	XM_002695415	1260	56	22.5
LOC100297304	XM_002695654	249	11	22.63
LOC100336631	XM_002699377	411	18	22.83
LOC784207	XM_002688759	186	8	23.25
H2B	NM_001114854.1	381	16	23.81

4.3.2 Stop gains and stop losses

Polymorphisms that cause the gain or loss of a stop codon in the coding region of a gene are particularly disruptive, causing large changes in the resulting protein that are more likely to have an effect on the function of the protein than do single amino acid changes. The genes containing stop gain and loss SNPs were compared with the OMIM database to determine what phenotypes these SNP could potentially be associated with if homozygous. About one fifth of the genes with stop gains or losses have OMIM annotation. DAVID analysis of genes with stoploss SNPs show enrichment for histone proteins, olfactory receptor genes, Zinc finger proteins, ribosomal proteins, and genes associated with coagulation and phosphoinositide metabolism. The list of genes containing stopgains are enriched for genes associated with olfactory receptor activity and membrane proteins. A summary of nsSNPs that cause stop gains and stop losses and that appear in only one breed are shown in Table 15.

Table 15: Summary of SNPs that cause stop gains or losses and occur in only one breed.

Breed	Sequening Depth	Stopgains	Stoplosses
Sahiwal	10x	21	5
Fleckveih	10x	17	2
Limousin	16x	18	4
Jersey	10.8x	17	4
Holstein	21.9x	41	11
Angus	15.5x	12	0
Nelore	14x	46	7
Gir	2.6x	5	0
Brahmin	13.9x	41	5
N'Dama	10x	0	0
Romanola	3.7x	0	0
Total	128.4x	218	38

4.3.3 nsSNPs by subspecies and breed type

The distribution of nsSNPs can be further subclassified by subspecies and breed. The data contain sequence from four indicine breeds and seven taurine breeds. Figure 24 shows the distribution of BLOSUM scores for four different groups of nsSNPs: nsSNPS unique to taurus breeds (and fixed in indicus breeds), nsSNPs unique to indicine breeds (and fixed in taurus breeds); nsSNPs with both alleles present in both sub species; and nsSNPs with alleles fixed by subspecies. The distributions are very similar except for BLOSUM score

2 where indicine SNPs outnumber taurine SNPs. There are 26 nsSNPs that cause the gain or loss of a stop codon. The genes containing these nsSNPs are listed in Table 16.

The eleven breeds can be grouped into beef (Brahman, Nelore, Angus, Limousin, Romagnola, and N'Dama) and dairy (Gir, Holstein, Jersey, and Sahiwal) breeds, with Romagnola, a dual-purpose breed, not included in either group. Table 17 lists the count of nsSNPs where the SNP is fixed in all breeds in a group. These nsSNPs fixed in beef and dairy breeds are distributed across 17,656 and 16,183 genes, respectively. There are no nsSNPs with one allele fixed in beef breeds and the alternative allele fixed in dairy breeds.

Table 16: The list of genes containing the 26 stopgain or stoploss SNPs whose alleles are fixed by subspecies.

Gene Name	**Stop Type**	**OMIM Phenotype**
ENG	T gain, I loss	Telangiectasia, hereditary hemorrhagic, type 1
FBN3	T gain, I loss	
KIAA1715	T gain, I loss	
LOC512248	T gain, I loss	
LOC618944	T gain, I loss	
LOC786046	T gain, I loss	
LRRTM4	T gain, I loss	
VWA3A	T gain, I loss	
XIST	T gain, I loss	X-inactivation, familial skewed
ZFP36L2	T gain, I loss	
ATG4B	T loss, I gain	
C19H17ORF39	T loss, I gain	
CD5L	T loss, I gain	
EHD1	T loss, I gain	
LIPC	T loss, I gain	Hepatic lipase deficiency; Diabetes mellitus, noninsulin-dependent
LOC100337101	T loss, I gain	
LOC507696	T loss, I gain	
LOC533894	T loss, I gain	
LOC613370	T loss, I gain	
LOC615576	T loss, I gain	
LOC785007	T loss, I gain	
LOC790218	T loss, I gain	
MAPK11	T loss, I gain	
MARK4	T loss, I gain	
SEMA4B	T loss, I gain	
SRGN	T loss, I gain	

Table 17: Summary of nsSNPs that are fixed in beef or dairy breeds.

	Beef	**Dairy**
nsSNP	5,116	9,258
Stopgain	133	220
Stoploss	34	41

4.3.4 CNV and repeats

Based on masking data provided by the USDA 31,116 (14.2%) of nsSNPs are located in repeat regions, and 19,718 (8.7%) are located in CNVs. An alternative CNV dataset from the USDA identified that 34,051 nsSNPs (15.5%) were located in CNVs, where CNVs were identified either by next-generation sequencing or HD SNP chips. This represents a 2-fold enrichment for nsSNPs in CNVs. The genes with high nsSNP density located in CNVs are enriched for genes associated with olfactory receptors, the major histocompatability locus (MHC), cationic amino acid transporters, and interferon.

4.3.5 Effect of SNP quality

The provided SNP data included simple quality scores, from one to four, for each SNP indicating the confidence in the existence of the SNP and its placement in the genome. SNPs in each quality class had different characteristics indicating that incorrect SNP calls may be introducing biases into these results. For example, the transition to transversion ratio of exonic SNPs was 1.16:1, but the ratio for the SNPs with the highest quality score was 3.24:1, a ratio in agreement with previous studies [42, 43]. This indicates that an adjustment in SNP calling thresholds may be warranted.

4.3.6 Conclusion

The SNP functional classifications presented here are a resource for narrowing down the source for the causative polymorphisms affecting traits of interest. These results will require frequent updating as more SNPs are continuously identified by low cost high-throughput sequencing.

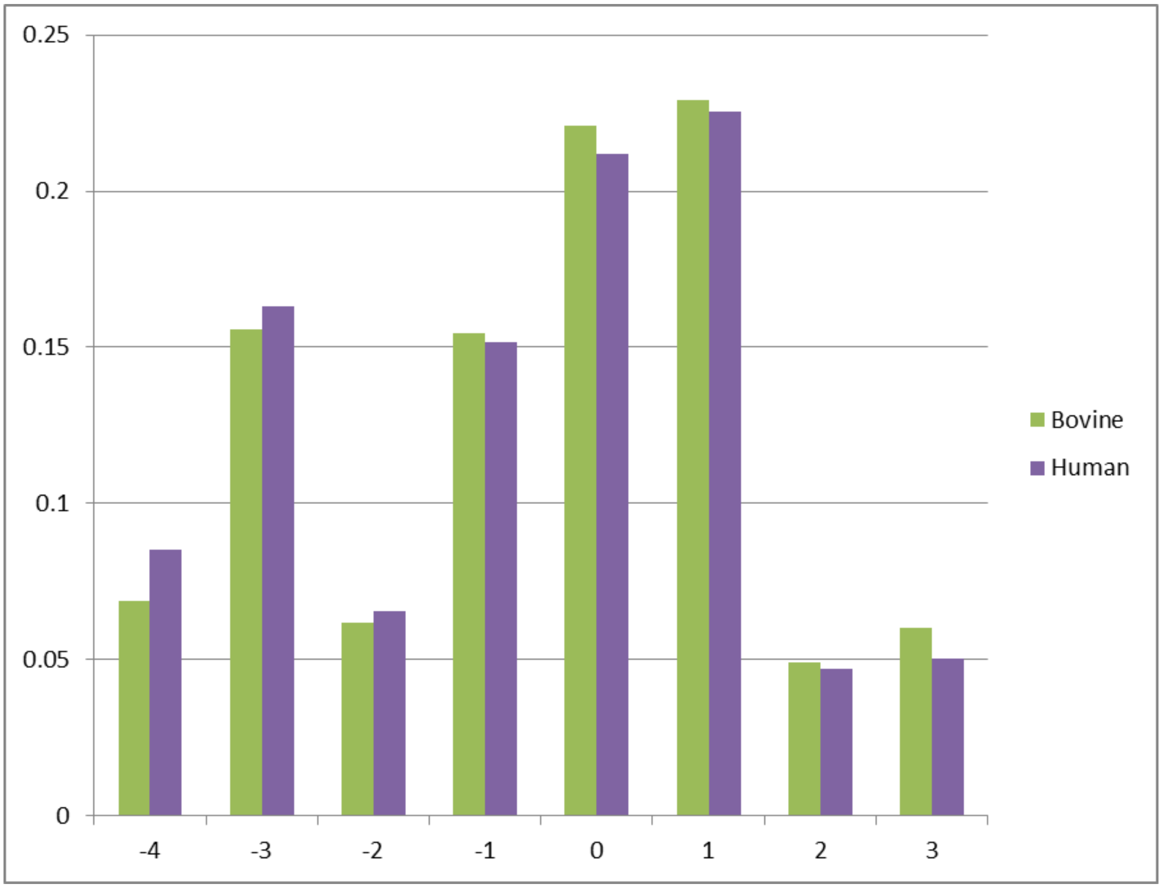


Figure 20: Comparison of substitution score counts between Human and Bovine nsSNPs.

The distribution of Blosum scores for non-synonymous substitutions in the human 1000 genomes dataset is very similar to the distribution found in the bovine 150 genomes dataset.

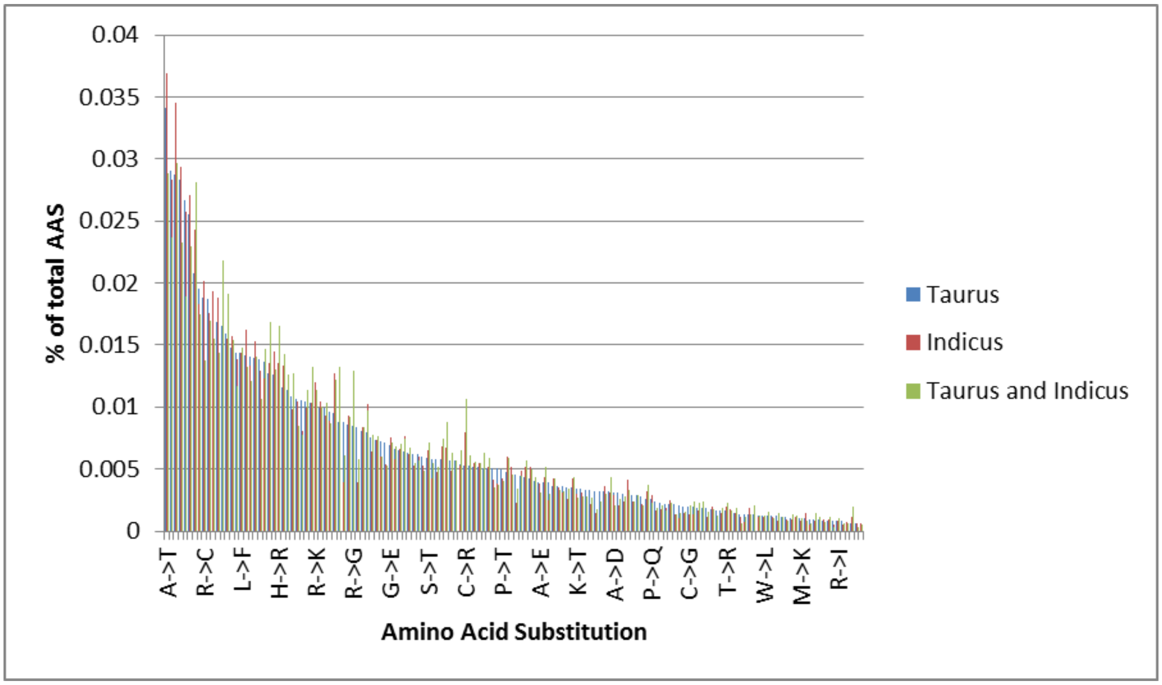


Figure 21: Comparison of substitution counts by substitution type between Human and Bovine nsSNPs. There are 150 different types of AAS in the human and bovine SNP data; these AAS are plotted on the x-axis. The total number of AAS that occur only in taurine breeds, only in indicine breeds, or occur in both breeds are plotted as bars on y-axis.

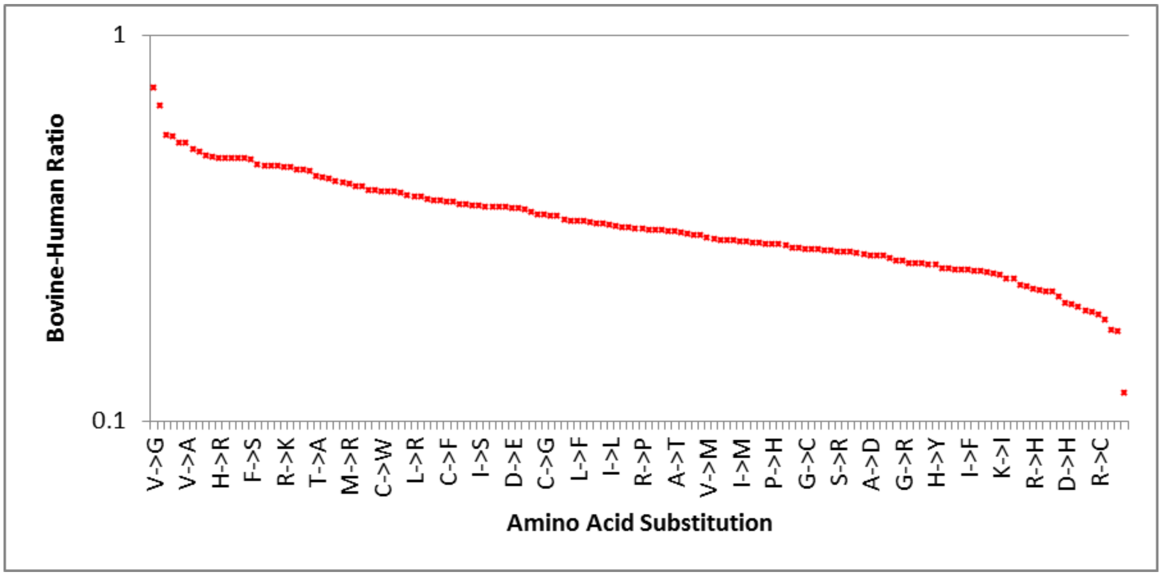


Figure 22. Bovine-Human AAS ratio. The 150 AAS in the human and bovine data are plotted on the x-axis and the ratio of bovine to human counts is plotted on the y-axis.

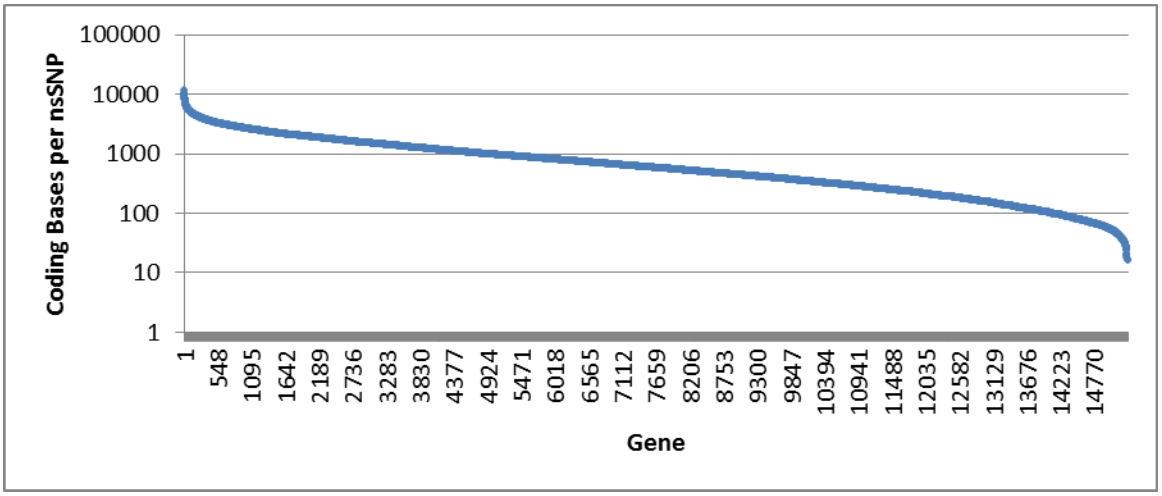


Figure 23: Distribution of nsSNP density. The 18,612 genes containing nsSNPs are plotted on the x-axis sorted by nsSNP density which is plotted on the logarithmic y-axis.

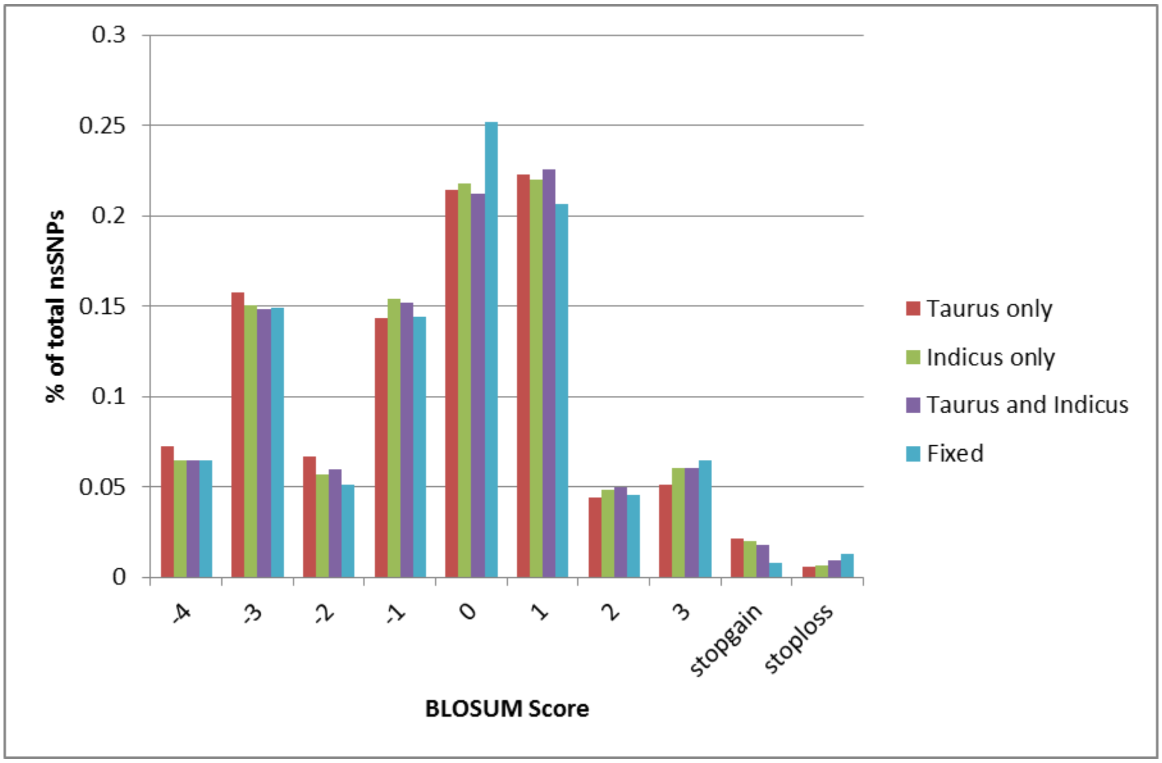


Figure 24: nsSNP BLOSUM score distributions. Distribution of BLOSUM scores for nsSNPS unique to Taurus breeds or Indicus breeds; nsSNPs with both alleles present in both types of breeds; and nsSNPs with alleles fixed by breed type. The distributions are very similar except for BLOSUM score 0 where fixed SNPs outnumber the other types.

5 Aim 4: Biomedical annotation of the porcine genome

5.1 Background

Pigs serve as important biological models for human diseases and are widely used in biomedical research [44-46]. Transgenic pig models have been developed for neurodegenerative and cardiovascular diseases, cystic fibrosis, and diabetes mellitus. The anatomical and physiological similarity of pigs to humans and their advantageous reproductive characteristics make them well suited for use in biomedicine. Advanced reproductive and genome-engineering technologies have been developed enabling efficient and precise development of new porcine models [47]. Pigs can model human diseases provided there are loss of function, gain of function, hypermorph, hypomorph and haplo-insufficient alleles. They can be identified in current swine populations, or more reasonably, can be created through targeted genome engineering. However, the current biomedical annotation of the pig genome is far from complete, presenting an obstacle in designing genome modification strategies to develop new pig models of human diseases. Integrating and agglomerating pig genome annotations as well as transferring human and mammalian annotations can greatly increase the knowledge guiding the engineering of the pig genome.

The emergence of sequence-specific genome-engineering platforms are revolutionizing biological research and affecting the medical industry. These platforms include: Meganucleases [48], Zinc-finger nucleases (ZFNs) [49], Transcription activator-like effector nucleases (TALENs)[50] and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs)[51]. Such a development has enabled the targeted modification of the swine genome. Each of these nucleases has different DNA binding characteristics that require specialized computation for identifying putative target sites. Selecting the appropriate tool for targeting a specific gene requires computing putative sites for each tool and compiling together the results, a time-consuming process that could be eliminated if target sites for all of the tools for the entire genome were pre-computed and compiled together. The current swine genome assembly (Sscrofa10.2) [52, 53] lacks a logical interface for users to identify, inspect, and engineer genes in the pig genome that correspond to causal alleles responsible for congenital diseases in humans.

In this paper we introduce the MedSwine system, a new genome web portal that agglomerates human disease related phenotypic annotation from human and model mammalian genomes to the pig genome, and integrates it with data on putative target sites for an assortment of genome engineering tools. The graphically guided genome engineering functionality also allows researchers to navigate from disease coordinates to nuclease binding sites with respect to the genes associated with specific phenotypes or diseases of interest. This resource provides a valuable tool for guiding the engineering of the swine genome for biomedical and agricultural applications.

5.2 Methods

5.2.1 From human disease causal variations to the swine alleles

Human variation data was mined from two resources, Online Mendelian Inheritance in Man (OMIM) [54] and dbSNP [39]. The web APIs of both resources were used to retrieve variants with pathogenic clinical significance in dbSNP and all variations associated with human genetic diseases in OMIM. To transfer these disease-causing alleles to the pig genome, we aligned the protein products of each human gene with its swine ortholog (Figure 25). The orthologous gene pairs were retrieved through Ensembl Core and Compara API [55]). We use both the protein alignment as well as the open reading frame (ORF) information to infer the orthologous positions of these alleles within the pig genome. Inferred disease-associated alleles in the pig genome can later serve as important guides to selecting nucleases to target the allele.

5.2.2 Data Incorporation

MedSwine also incorporates information from a variety of resources and laboratories (Figure 26). Two types of data are included: annotations and putative nuclease-binding sites. In summary, MedSwine identifies the gene ID of each mouse, human, and rat gene orthologous to each pig gene from Ensembl Biomart [55, 56] using the Ensembl Perl API. These orthologous gene IDs are used for transferring mouse, human, and rat gene annotations to pig genes. OMIM vocabulary and the disease causal variant associations are imported through the OMIM web API. The MedSwine pipeline also integrates human phenotype ontology (HPO) annotation of the human gene build by monitoring the HPO ftp site, mouse phenotype ontology (MPO) annotation of the mouse from the Jackson Laboratory ftp site, pig gene ontology (GO) annotation from Ensembl, pathway (PW) and rat disease ontology (RDO) annotations for the rat from the RGD RatMine Python API, and OMIA annotations by loading the OMIA SQL dump.

We implemented a data-integrating pipeline to collect data from databases through different access points. Due to the heterogeneous structure of these data sources, this pipeline involves FTP handling, Perl API access for annotation and python scripts to retrieve rat orthologous disease related data from RGD. All these data are then imported into the MedSwine database schema.

5.2.3 Identification of Genome Modification Target Sites

CRISPR/Cas9 Target Sites/Seeds

We identified Cas9 target sites [51] in the latest gene build from Ensembl. Coding exon sequences with 20bp flanking regions were retrieved using the Ensembl Perl API. We identified all 23bp sequences with the form 5’-GBBBB BBBBB BBBBB BBBBB NGG-3’ where B’s are the exon bases. For each of the 23bp candidate targets, we used Bowtie 0.12.8 [31] to filter out those with form 5'-NNNNN NNBBB BBBBB BBBBB NGG-3' within the pig genome to prevent off-target cuts [51]. Those sequences that passed the filter are unique Cas9 targets that can be applied in pig genome engineering.

Transcription Activator-Like Effector Nucleases (TALENs)

TALEN target sites are also calculated for all pig genes and integrated into MedSwine. We applied the target finder python scripts used in TALE-NT 2.0 [57] to the pig genome and filtered for candidate TALENs in coding exons. Due to the high frequency of TALEN in the genome, we limited the number of target sites to no more than one cut site per 10bp window. TALENs with 16 repeat variable di-residues (RVDs) and 16bp spacers were favoured in MedSwine; characteristics that have reliably resulted in high TALENs using the GoldyTALEN scaffold [58].

ZFN

Target sites of ZFNs were calculated using two methods, Oligomerized Pool Engineering (OPEN) and Context-Dependent Assembly (CoDA)[59, 60]. OPEN sites were scored using a zinc-finger OPEN Targeter (ZiFOpT) confidence score [61].

5.2.4 MedSwine Website

The MedSwine web interface is implemented as a blend of HTML, CSS, JavaScript, and PHP components. Annotations, homology data, and nuclease binding site data are stored in a MySQL relational schema. Jbrowse version 1.11.6 is used to display the current Ensembl and NCBI gene builds[62]. Seven additional tracks are included to display nuclease-binding sites for TALENs, CRISPR/Cas9 sites, and OPEN and CoDA ZnFN sites, as well as OMIM human diseases causative variants mapped to the pig genome, and genomic variants from Ensembl Variation and deep catalog of autosomal single nucleotide variation in pigs published by Bianco *et al* [63].

5.3 Results

5.3.1 From Human Disease to Pig Alleles

The major focus of MedSwine is to facilitate the modelling of human diseases. OMIM provides a comprehensive compendium of human congenital disorders along with their causal variants within disease related genes. MedSwine tries to map these disease causal variants to the pig genome. MedSwine incorporates a total of 8,834 associations between 1,767 OMIM genes and 4,162 unique disease causal variants. Of the 4,162 variants, 3,987 are associated with 2,349 OMIM phenotypes. Due to multiple alignments the 4,162 variants map to 6,566 locations in the swine genome in 1,488 genes. This is a significant improvement over OMIA alone where only 23 genes in pigs were associated with human disease.

5.3.2 Integrating Biomedical Data and Genome Engineering Tools

Annotations: There are 25,322 genes in the pig gene build in Ensembl release 75. Of these, 18,356 genes have orthologous human genes, 18,511 have orthologous mouse genes, and 17,847 have orthologous rat genes. Table 18 summarizes the transfer of annotation from the human, mouse, and rat genomes to the pig genome. Most ontology entries (about 90%) can be mapped to the pig genome based on human-pig, mouse-pig or rat-pig orthology. Existing GO annotation for the pig genome, comprised of 99,262 annotations to 17,495 genes, was also incorporated into the MedSwine database.

Nucleases: A total of 225,593 CRISPR/Cas9 targets were identified in the Ensembl 75 gene build; Due to nature of abundance for TALENs integration sites, we set the threshold to retain only one TALENs pair every 10 bases.

Browser Tracks: The genome browser contains the Sscrofa10.2 genome assembly along with seven annotation tracks. The “GENE_Ensembl” track displays the Ensembl gene build and the

“GENE_NCBI” track displays the NCBI gene build. HUGO Gene Nomenclature Committee (HGNC) gene names are shown if available, otherwise the Ensembl or NCBI gene IDs are listed (Gray & Daugherty, 2013). Four putative nuclease binding site tracks are available: CRISPR/Cas9, TALENs, ZFN CoDA Sites, and ZFN OPEN Sites. Due to the large number of CRISPR and TALEN sites in the genome, only those occurring in exons are displayed in the browser tracks. Human disease causal variants mapped to the swine genome are also included in the browser, providing precise locations in the swine genome to guide nuclease selection. Two tracks showing known swine variants are also included.

5.3.3 Workflow and User Interface

The MedSwine web interface provides two approaches for retrieving information from the database. Users can either search for a gene of interest using the gene tab, or ontology terms in the annotations tab. For the ontology-based search, MedSwine retrieves a list of genes matching the search term from the database along with their genomic coordinates, and presents the user with a list of genes as well as an interactive snapshot of the genome showing the locations of the genes. Links are provided for each gene to a detailed view of the gene in a genome browser and gene report pages that list all associated annotations and nuclease binding sites within or near the gene. The ontology page lists the number of genes annotated with each ontology term as well as cross-reference to other ontology data.

To demonstrate the utility of MedSwine, we focus below on two genes associated with familial hypercholesterolemia (FH), a major aetiology of cardiovascular disease [64]. Mutations in the low density lipoprotein receptor (*LDLR*) and proprotein convertase subtilisin/kexin type 9 (*PCSK9*) genes are the primary causes of FH. Using the search term “hypercholesterolemia” in the MedSwine query box in the annotation tab (Figure 27) brings up the gene distribution viewer (Figure 28), gene list panel (Figure 29) and ontology list panel (Figure 30). The gene list panel provides a much more comprehensive set of genes (Table 19) than searching using the same term against the HPO and OMIM websites alone. Clicking on an ontology link in the result list will bring up the ontology page with genes associated to the specific entry (Figure 31). Any gene from the list can be chosen to display their respective detail page as exemplified in Figure 32 by selecting PCSK9. In *PCSK9* there are 5 pathogenic variants mapped to the gene from its human ortholog (Table 20).

Nuclease design: We found two Ensembl *LDLR* genes with very low consensus regions: One located on the reverse strand of chromosome 2:70,193,418-70,206,818 with 4 exons, the other located on scaffold GL896440: 2,830-14,728 with 16 exons. *LDLR* is better annotated on the scaffold than on chromosome 2. All disease causal mutations in the human *LDLR* were mapped to the *LDLR* gene on GL896440. On the *LDLR* gene located on chromosome two we find a TALEN target with spacer size 15 located at exon 2 (chr2:70195973-70195989, negative strand) (Figure 33). This sequence was successfully verified in a previous study [65].

5.3.4 Keeping Data Current

Due to the constant improvement to the pig genome assembly there is a need for updating the MedSwine database on a regular basis since further studies may reveal inaccuracies. The Sscrofa10.2 genome assembly is a significant improvement over the previous build, but some problems still remain. There are genes incorrectly mapped to chromosomes and some annotated genes remain in unplaced scaffold contigs. One example would be the LDLR annotation discrepancy mentioned previously. MedSwine has an automatic updating pipeline that pulls data from different external resources to enable frequent updates and minimize manual effort. As major databases such as Ensembl are updated quarterly, this pipeline will continue to keep MedSwine up-to-date, as it has for the last three years.

5.3.5 Future work

We plan to merge pig genome annotations as well as corresponding genome engineering tools into MedSwine to facilitate the manipulation of target genes. As molecular biology techniques are developed, MedSwine has a long-term focus on adding new pre-computed nuclease targets (other enzymes, C31, recombinases) and regular updates of annotations by including additional model organisms. We will also be extending the system to support broader and more specific user interface interactions for keeping track of queries as well as designed nucleases.

5.4 Conclusions

MedSwine provides biomedical oriented ontological annotation of the pig genome that incorporates annotations and ontologies from both human and mammalian model animal genomes. As a resource for disease modelling in swine, MedSwine also incorporates greater than one billion target sites encompassing six different genome modification platforms. The synergy between the constant proceedings of swine genome assembly and MedSwine's integrative architecture makes MedSwine an invaluable research tool for engineering the swine genome for both biomedical and agricultural applications.

Table 18 - A Summary of annotations transferred from human, mouse, and rat genomes to the pig genome

	OMIM	HPO	MPO	PW	RDO
Terms defined in original database	4,247[#]	10,686	9,440	1,329	12,772
Annotations in the source genome	4,247	5,672	7,563	370	5,998
Source genes with annotations	4,253	3,774	6,915	1,732	10,319
Source genes with pig orthologs	3,356	3,270	7,294	1,817	8,932
Transferred phenotypic annotations to the pig genome	3,673	5,481	7,452	363	1,230
Annotation transfer rate*	89%	96%	98%	98%	95%

The number of human genes with OMIM and HPO terms, mouse genes with MPO terms, and rat genes with PW and RDO terms are shown along with the total number of annotations per genome. 89% or more of these annotations are successfully transferred to the pig genome. [# OMIM entries include "Phenotype description, molecular basis known"] [*Annotation transfer rate is calculated as the ratio of the transferred phenotypic annotations to the pig genome to annotations from the source genome]

Table 19 - MedSwine gene list for hypercholesterolemia

ID	Symbol
ENSSSCG00000003088	APOE
ENSSSCG00000003549	LDLRAP1
ENSSSCG00000008595	APOB
ENSSSCG00000008596	
ENSSSCG00000024093	
ENSSSCG00000025020	PCSK9
ENSSSCG00000028330	
ENSSSCG00000016866	GHR
ENSSSCG00000028512	
ENSSSCG00000030900	LDLR
ENSSSCG00000006225	TTPA
ENSSSCG00000007067	JAG1
ENSSSCG00000008453	ABCG8
ENSSSCG00000008454	ABCG5
ENSSSCG00000010450	LIPA
ENSSSCG00000011001	APTX
ENSSSCG00000012157	PHKA2
ENSSSCG00000012650	OCRL
ENSSSCG00000015336	SLC25A13
ENSSSCG00000016634	CAV1
ENSSSCG00000027197	DYRK1B
ENSSSCG00000009759	SCARB1

A total of 21 pig genes are found to be associated with the term “hypercholesterolemia” when searching against OMIM and HPO, of these 16 are identified in HPO and 11 in OMIM.

Table 20 - PCSK9 mutations in OMIM mapped to pig.

Human SNP ID	Transcript	Location within transcript	Assigned pig label
rs28362286	ENSSSCT00000029382	54%	SW_6089
rs28942112	ENSSSCT00000029382	22%	SW_6249
rs67608943	ENSSSCT00000029382	16%	SW_6352
rs28942111	ENSSSCT00000029382	15%	SW_6248
rs11591147	ENSSSCT00000029382	10%	SW_4950

There are five disease-associated SNPs in human gene PCSK9; all of them have corresponding pig homologs.

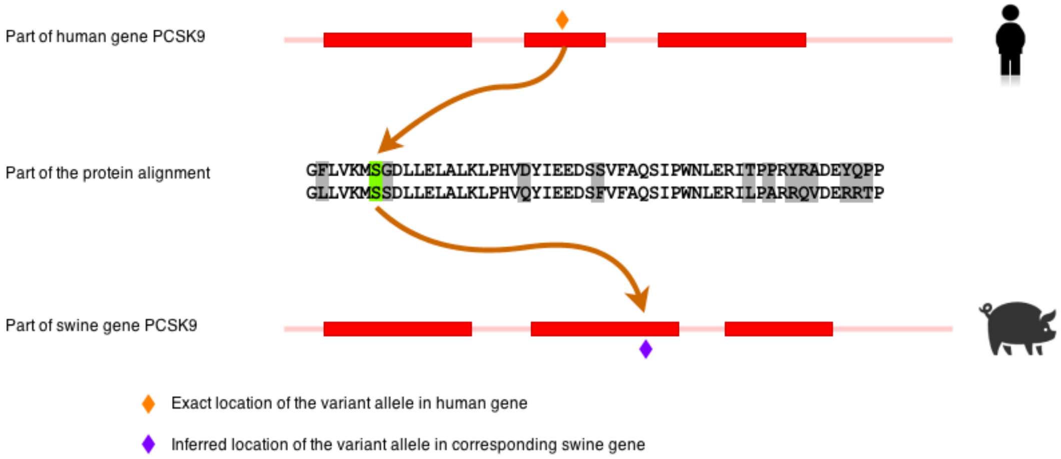

Part of human gene PCSK9
Part of the protein alignment
GFLVKMSGDLLELALKLPHVDYIEEDSSVFAQSIPWNLERITPPRYRADEYQPP
GLLVKMSSDLLELALKLPHVQYIEEDSFVFAQSIPWNLERILPARRQVDERRTP
Part of swine gene PCSK9
Exact location of the variant allele in human gene
Inferred location of the variant allele in corresponding swine gene

Figure 25 – Medswine causal variant mapping. Disease causal variants are mapped from the human genome to the swine genome via their corresponding orthologous protein sequences. In this figure, the human variant rs28942111 has an inferred location in the pig genome due to the identical amino acid (indicated by arrow). Variation rs28942111 is located at the third base within the codon causing the change from AGT to AGA. It is pathogenic due to its cause of amino acid change from Ser (S) to Arg (R) at position 127 in the human protein which is the translation product of transcript *PCSK9*-001 (Ensembl id: ENST00000302118).

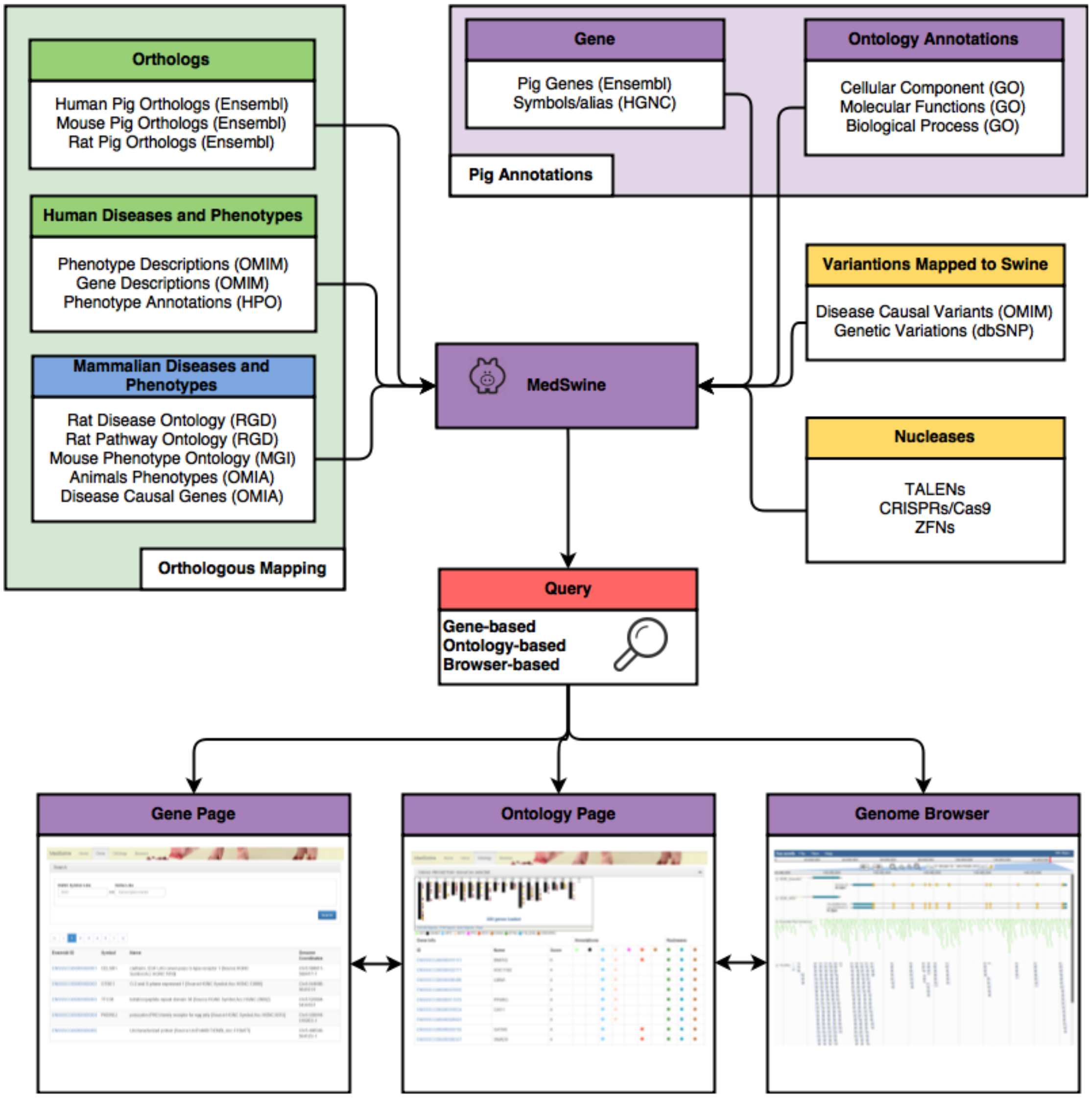
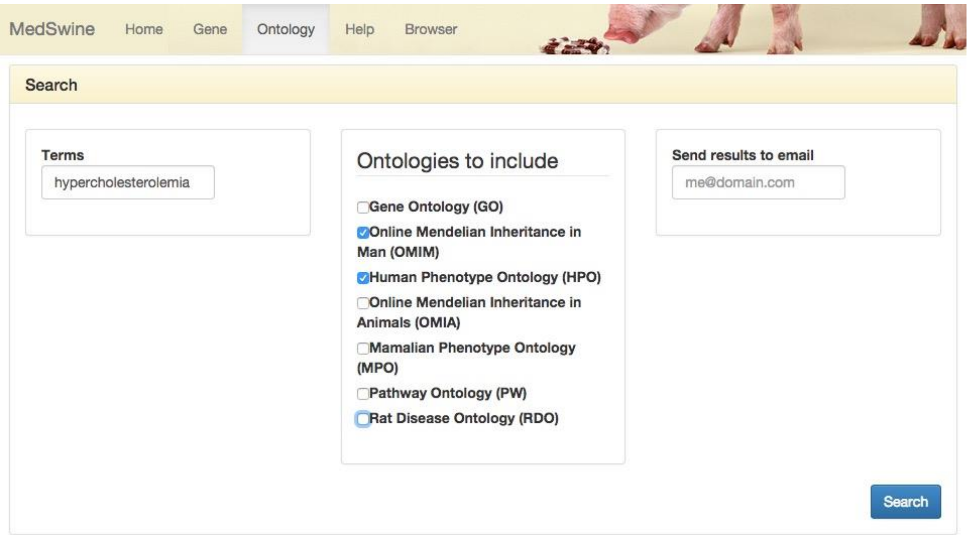

Orthologs
Human Pig Orthologs (Ensembl)
Mouse Pig Orthologs (Ensembl)
Rat Pig Orthologs (Ensembl)
Human Diseases and Phenotypes
Phenotype Descriptions (OMIM)
Gene Descriptions (OMIM)
Phenotype Annotations (HPO)
Mammalian Diseases and Phenotypes
Rat Disease Ontology (RGD)
Rat Pathway Ontology (RGD)
Mouse Phenotype Ontology (MGI)
Animals Phenotypes (OMIA)
Disease Causal Genes (OMIA)
Orthologous Mapping
Gene
Pig Genes (Ensembl)
Symbols/alias (HGNC)
Ontology Annotations
Cellular Component (GO)
Molecular Functions (GO)
Biological Process (GO)
Pig Annotations
Variantions Mapped to Swine
Disease Causal Variants (OMIM)
Genetic Variations (dbSNP)
MedSwine
Nucleases
TALENs
CRISPRs/Cas9
ZFNs
Query
Gene-based
Ontology-based
Browser-based
Gene Page
Ontology Page
Genome Browser

Figure 26 – Medswine data architecture. MedSwine includes a query front end and a data processing back end. MedSwine accepts three kinds of queries: gene-based, ontology-based and browser-based. The MedSwine backend automatically transfers annotations from external resources and converts this data to fit MedSwine's database. MedSwine uses latest Ensembl gene build and HGNC gene names and aliases.



Search

Terms

hypercholesterolemia

Ontologies to include

□ Gene Ontology (GO)

☑ Online Mendelian Inheritance in Man (OMIM)

☑ Human Phenotype Ontology (HPO)

□ Online Mendelian Inheritance in Animals (OMIA)

□ Mamalian Phenotype Ontology (MPO)

□ Pathway Ontology (PW)

□ Rat Disease Ontology (RDO)

Send results to email

me@domain.com

Figure 27 – Medswine search interface. Annotation Tab of the query box

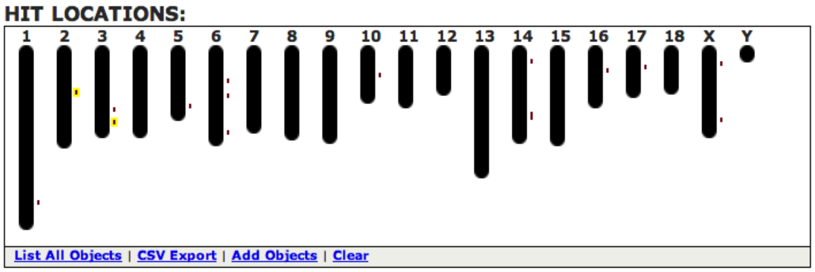
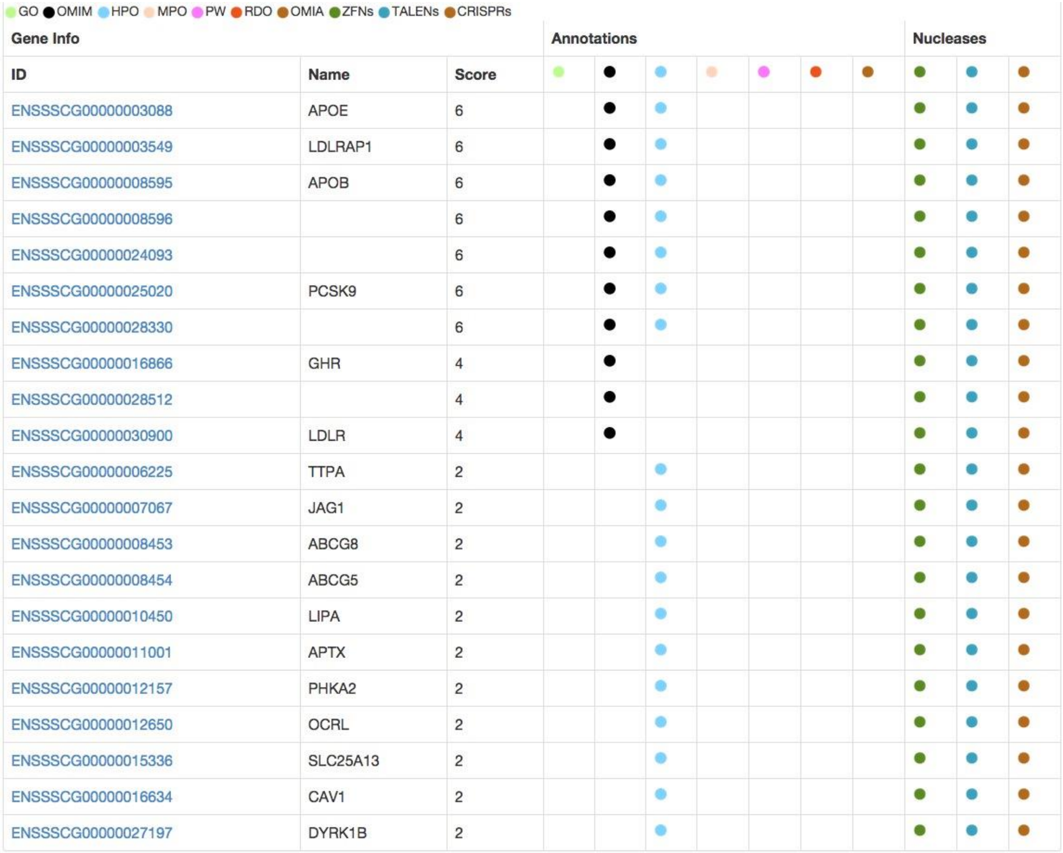


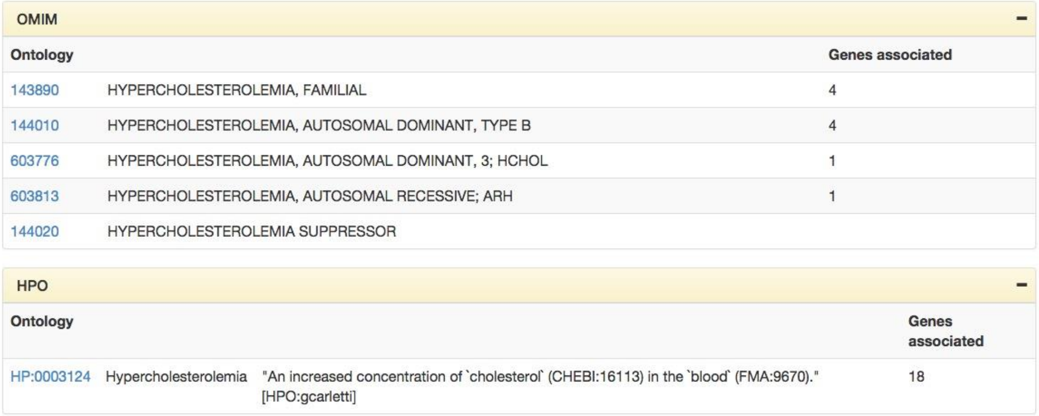
Figure 28 – Medswine search result graphic. Snapshot of gene distribution viewer where multiple hits are highlighted with shaded boxes.



● GO ● OMIM ● HPO ● MPO ● PW ● RDO ● OMIA ● ZFNs ● TALENs ● CRISPRs

Gene Info			Annotations							Nucleases		
ID	Name	Score	●	●	●	●	●	●	●	●	●	●
ENSSSCG00000003088	APOE	6		●	●					●	●	●
ENSSSCG00000003549	LDLRAP1	6		●	●					●	●	●
ENSSSCG00000008595	APOB	6		●	●					●	●	●
ENSSSCG00000008596		6		●	●					●	●	●
ENSSSCG00000024093		6		●	●					●	●	●
ENSSSCG00000025020	PCSK9	6		●	●					●	●	●
ENSSSCG00000028330		6		●	●					●	●	●
ENSSSCG00000016866	GHR	4		●						●	●	●
ENSSSCG00000028512		4		●						●	●	●
ENSSSCG00000030900	LDLR	4		●						●	●	●
ENSSSCG00000006225	TTPA	2			●					●	●	●
ENSSSCG00000007067	JAG1	2			●					●	●	●
ENSSSCG00000008453	ABCG8	2			●					●	●	●
ENSSSCG00000008454	ABCG5	2			●					●	●	●
ENSSSCG00000010450	LIPA	2			●					●	●	●
ENSSSCG00000011001	APTX	2			●					●	●	●
ENSSSCG00000012157	PHKA2	2			●					●	●	●
ENSSSCG00000012650	OCRL	2			●					●	●	●
ENSSSCG00000015336	SLC25A13	2			●					●	●	●
ENSSSCG00000016634	CAV1	2			●					●	●	●
ENSSSCG00000027197	DYRK1B	2			●					●	●	●

Figure 29 – Hypercholesterolemia gene search results. Gene list panel of the search result from "hypercholesterolemia"



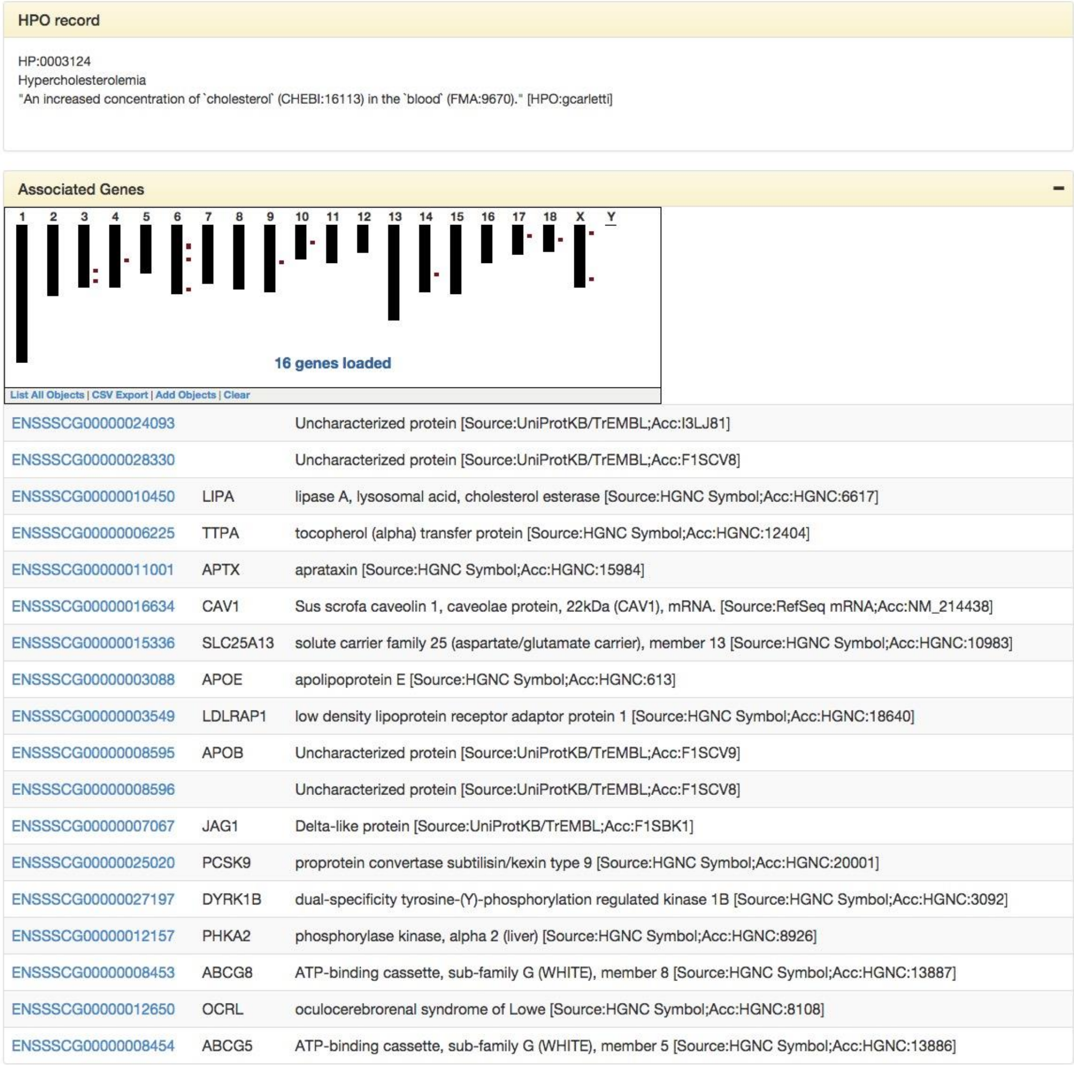
OMIM −

Ontology		Genes associated
143890	HYPERCHOLESTEROLEMIA, FAMILIAL	4
144010	HYPERCHOLESTEROLEMIA, AUTOSOMAL DOMINANT, TYPE B	4
603776	HYPERCHOLESTEROLEMIA, AUTOSOMAL DOMINANT, 3; HCHOL	1
603813	HYPERCHOLESTEROLEMIA, AUTOSOMAL RECESSIVE; ARH	1
144020	HYPERCHOLESTEROLEMIA SUPPRESSOR	

HPO −

Ontology			Genes associated
HP:0003124	Hypercholesterolemia	"An increased concentration of `cholesterol` (CHEBI:16113) in the `blood` (FMA:9670)." [HPO:gcarletti]	18

Figure 30 – Hypercholesterolemia ontology search results. Snapshot of result entries from query of ontology term 'hypercholesterolemia'. The top term in each ontology are the desired entries. Clicking on those entries brings up the ontology report page shown as in Figure 31.

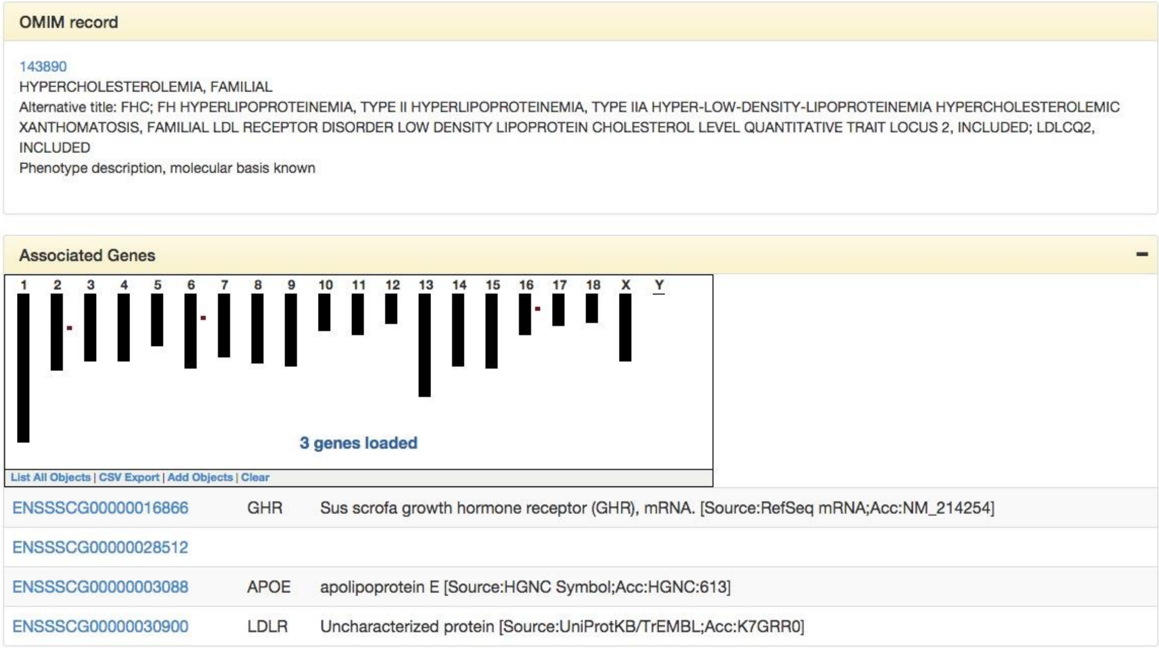


HPO record

HP:0003124
Hypercholesterolemia
"An increased concentration of `cholesterol` (CHEBI:16113) in the `blood` (FMA:9670)." [HPO:gcarletti]

Associated Genes

ENSSSCG00000024093		Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:I3LJ81]
ENSSSCG00000028330		Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F1SCV8]
ENSSSCG00000010450	LIPA	lipase A, lysosomal acid, cholesterol esterase [Source:HGNC Symbol;Acc:HGNC:6617]
ENSSSCG00000006225	TTPA	tocopherol (alpha) transfer protein [Source:HGNC Symbol;Acc:HGNC:12404]
ENSSSCG00000011001	APTX	aprataxin [Source:HGNC Symbol;Acc:HGNC:15984]
ENSSSCG00000016634	CAV1	Sus scrofa caveolin 1, caveolae protein, 22kDa (CAV1), mRNA. [Source:RefSeq mRNA;Acc:NM_214438]
ENSSSCG00000015336	SLC25A13	solute carrier family 25 (aspartate/glutamate carrier), member 13 [Source:HGNC Symbol;Acc:HGNC:10983]
ENSSSCG00000003088	APOE	apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613]
ENSSSCG00000003549	LDLRAP1	low density lipoprotein receptor adaptor protein 1 [Source:HGNC Symbol;Acc:HGNC:18640]
ENSSSCG00000008595	APOB	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F1SCV9]
ENSSSCG00000008596		Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F1SCV8]
ENSSSCG00000007067	JAG1	Delta-like protein [Source:UniProtKB/TrEMBL;Acc:F1SBK1]
ENSSSCG00000025020	PCSK9	proprotein convertase subtilisin/kexin type 9 [Source:HGNC Symbol;Acc:HGNC:20001]
ENSSSCG00000027197	DYRK1B	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1B [Source:HGNC Symbol;Acc:HGNC:3092]
ENSSSCG00000012157	PHKA2	phosphorylase kinase, alpha 2 (liver) [Source:HGNC Symbol;Acc:HGNC:8926]
ENSSSCG00000008453	ABCG8	ATP-binding cassette, sub-family G (WHITE), member 8 [Source:HGNC Symbol;Acc:HGNC:13887]
ENSSSCG00000012650	OCRL	oculocerebrorenal syndrome of Lowe [Source:HGNC Symbol;Acc:HGNC:8108]
ENSSSCG00000008454	ABCG5	ATP-binding cassette, sub-family G (WHITE), member 5 [Source:HGNC Symbol;Acc:HGNC:13886]



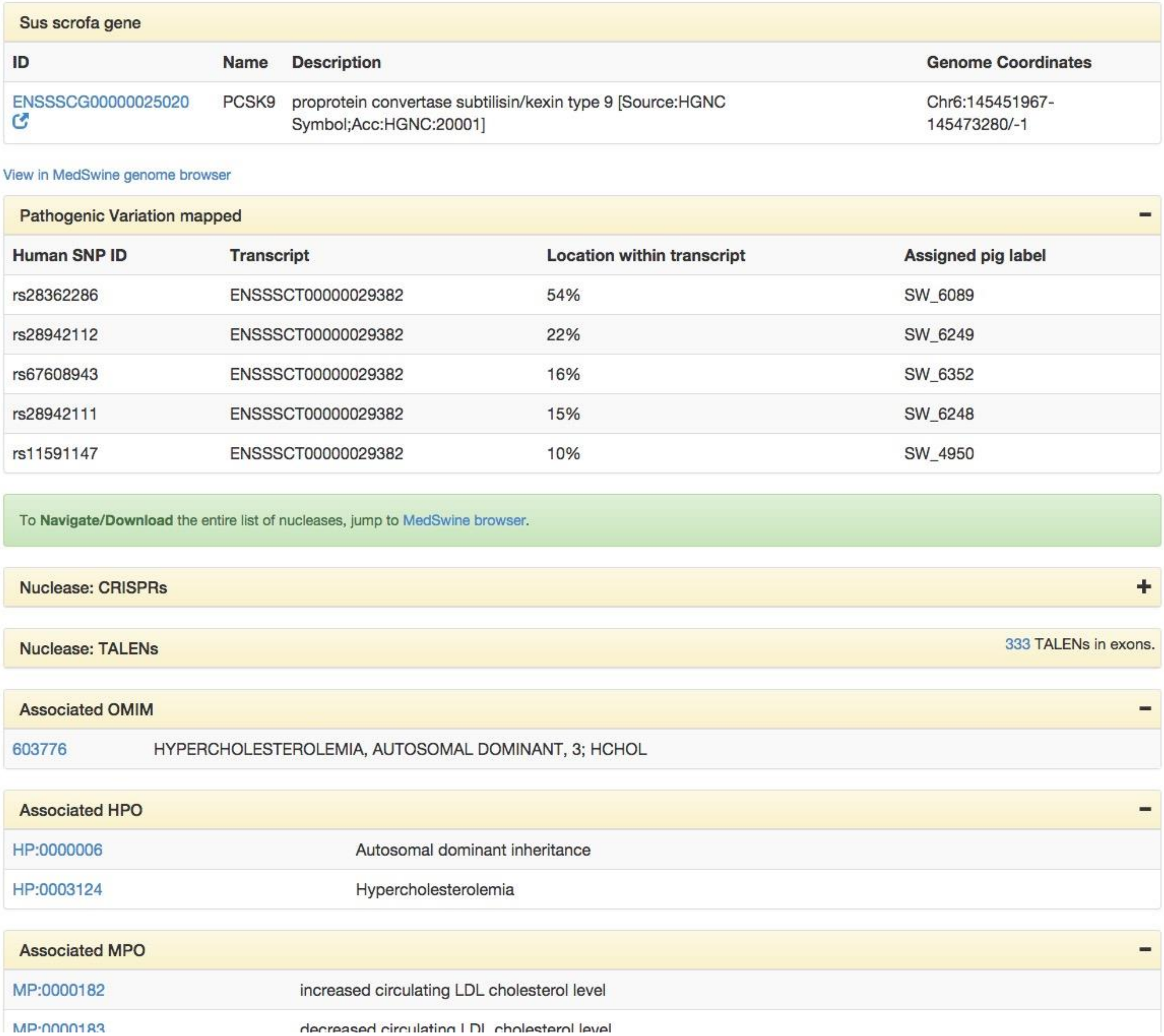
OMIM record

143890
HYPERCHOLESTEROLEMIA, FAMILIAL
Alternative title: FHC; FH HYPERLIPOPROTEINEMIA, TYPE II HYPERLIPOPROTEINEMIA, TYPE IIA HYPER-LOW-DENSITY-LIPOPROTEINEMIA HYPERCHOLESTEROLEMIC XANTHOMATOSIS, FAMILIAL LDL RECEPTOR DISORDER LOW DENSITY LIPOPROTEIN CHOLESTEROL LEVEL QUANTITATIVE TRAIT LOCUS 2, INCLUDED; LDLCQ2, INCLUDED
Phenotype description, molecular basis known

Associated Genes

ENSSSCG00000016866	GHR	Sus scrofa growth hormone receptor (GHR), mRNA. [Source:RefSeq mRNA;Acc:NM_214254]
ENSSSCG00000028512		
ENSSSCG00000003088	APOE	apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613]
ENSSSCG00000030900	LDLR	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:K7GRR0]

Figure 31 – Familial hypercholesterolemia ontology page. Snapshots of FH ontology page from both HPO (on the top) and OMIM (on the bottom) provide detailed information and a comprehensive list of associated genes with all available external links provided for cross-reference.



Sus scrofa gene

ID	Name	Description	Genome Coordinates
ENSSSCG00000025020	PCSK9	proprotein convertase subtilisin/kexin type 9 [Source:HGNC Symbol;Acc:HGNC:20001]	Chr6:145451967-145473280/-1

View in MedSwine genome browser

Pathogenic Variation mapped

Human SNP ID	Transcript	Location within transcript	Assigned pig label
rs28362286	ENSSSCT00000029382	54%	SW_6089
rs28942112	ENSSSCT00000029382	22%	SW_6249
rs67608943	ENSSSCT00000029382	16%	SW_6352
rs28942111	ENSSSCT00000029382	15%	SW_6248
rs11591147	ENSSSCT00000029382	10%	SW_4950

To **Navigate/Download** the entire list of nucleases, jump to MedSwine browser.

Nuclease: CRISPRs +

Nuclease: TALENs 333 TALENs in exons.

Associated OMIM

603776	HYPERCHOLESTEROLEMIA, AUTOSOMAL DOMINANT, 3; HCHOL

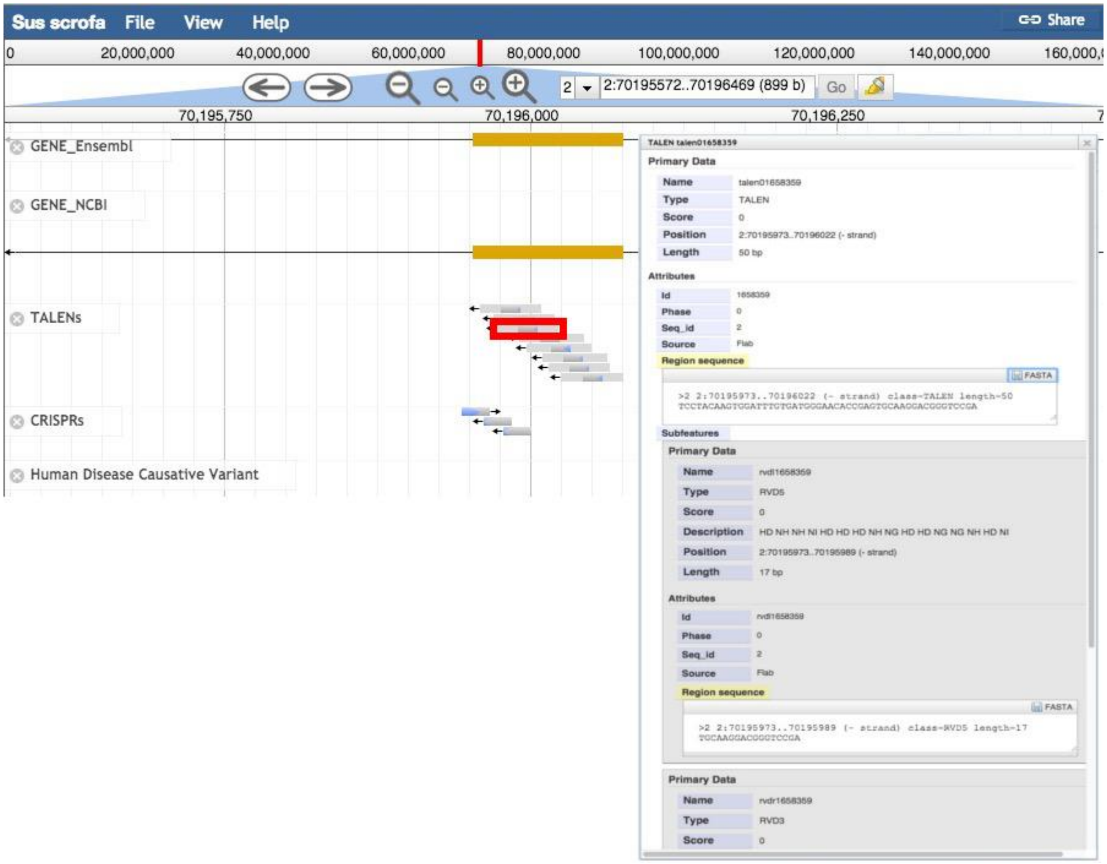
Associated HPO

HP:0000006	Autosomal dominant inheritance
HP:0003124	Hypercholesterolemia

Associated MPO

MP:0000182	increased circulating LDL cholesterol level
MP:0000183	decreased circulating LDL cholesterol level

Figure 32 – PCSK9 gene page. On the result pig gene page, MedSwine displays ontological annotations from GO, disease and phenotype associations from HPO, OMIM, MPO, PW and RDO. Statistics and links of the nucleases associated with this gene are also listed.



TALEN talen01658359

Primary Data

Name	talen01658359
Type	TALEN
Score	0
Position	2:70195973..70196022 (- strand)
Length	50 bp

Attributes

Id	1658359
Phase	0
Seq_id	2
Source	Flab

Region sequence

FASTA

```
>2 2:70195973..70196022 (- strand) class=TALEN length=50
TCCTACAAGTGGATTTGTGATGGGAACACCGAGTGCAAGGACGGGTCCGA
```

Subfeatures

Primary Data

Name	rvdl1658359
Type	RVDS
Score	0
Description	HD NH NH NI HD HD HD NH NG HD HD NG NG NH HD NI
Position	2:70195973..70195989 (- strand)
Length	17 bp

Attributes

Id	rvdl1658359
Phase	0
Seq_id	2
Source	Flab

Region sequence

FASTA

```
>2 2:70195973..70195989 (- strand) class=RVDS length=17
TGCAAGGACGGGTCCGA
```

Primary Data

Name	rvdr1658359
Type	RVD3
Score	0

Figure 33 – Medswine genome browser. Genome browser display shows a 500bp section of pig chromosome 2 surrounding the second exon of LDLR. In the "TALENs" track, nuclease designed with sequence TCTCCTACAAGTGGATTTgtgatgggaacaccgAGTGCAAGGACGGGTCCGA (in highlighted box) was selected and successfully applied to generate LDLR KO Ossabaw swine [65].

6 References

1. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS *et al*: **Development and characterization of a high density SNP genotyping assay for cattle**. *PLoS One* 2009, **4**(4):e5350.
2. **Understanding Genomic Predictions** [http://www.holsteinusa.com/pdf/print_material/genomics.pdf]
3. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigo R *et al*: **The genome sequence of taurine cattle: a window to ruminant biology and evolution**. *Science* 2009, **324**(5926):522-528.
4. Liu Y, Qin X, Song XZ, Jiang H, Shen Y, Durbin KJ, Lien S, Kent MP, Sodeland M, Ren Y *et al*: **Bos taurus genome assembly**. *BMC Genomics* 2009, **10**:180.
5. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS *et al*: **A whole-genome assembly of the domestic cow, Bos taurus**. *Genome Biol* 2009, **10**(4):R42.
6. Hu ZL, Reecy JM: **Animal QTLdb: beyond a repository. A public platform for QTL comparisons and integration with diverse types of structural genomic information**. *Mamm Genome* 2007, **18**(1):1-4.
7. Everts RE, Band MR, Liu ZL, Kumar CG, Liu L, Loor JJ, Oliveira R, Lewin HA: **A 7872 cDNA microarray and its use in bovine functional genomics**. *Vet Immunol Immunopathol* 2005, **105**(3-4):235-245.
8. Suchyta SP, Sipkovsky S, Kruska R, Jeffers A, McNulty A, Coussens MJ, Tempelman RJ, Halgren RG, Saama PM, Bauman DE *et al*: **Development and testing of a high-density cDNA microarray resource for cattle**. *Physiol Genomics* 2003, **15**(2):158-164.
9. Zhao SH, Recknor J, Lunney JK, Nettleton D, Kuhar D, Orley S, Tuggle CK: **Validation of a first-generation long-oligonucleotide microarray for transcriptional profiling in the pig**. *Genomics* 2005, **86**(5):618-625.
10. Li X, He Z, Zhou J: **Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation**. *Nucleic Acids Res* 2005, **33**(19):6114-6123.
11. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.
12. Galbraith DW, Elumalai R, Gong FC: **Integrative flow cytometric and microarray approaches for use in transcriptional profiling**. *Methods Mol Biol* 2004, **263**:259-280.
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

14. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data**. *Genome Res* 2004, **14**(1):160-169.
15. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P *et al*: **Integrating biological data--the Distributed Annotation System**. *BMC Bioinformatics* 2008, **9 Suppl 8**:S3.
16. Lee M, Xiang CC, Trent JM, Bittner ML: **Performance characteristics of 65-mer oligonucleotide microarrays**. *Anal Biochem* 2007, **368**(1):70-78.
17. Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJ *et al*: **An atlas of human gene expression from massively parallel signature sequencing (MPSS)**. *Genome Res* 2005, **15**(7):1007-1014.
18. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P *et al*: **A compendium of gene expression in normal human tissues**. *Physiol Genomics* 2001, **7**(2):97-104.
19. Arthur PF, Archer JA, Johnston DJ, Herd RM, Richardson EC, Parnell PF: **Genetic and phenotypic variance and covariance components for feed intake, feed efficiency, and other postweaning traits in Angus cattle**. *J Anim Sci* 2001, **79**(11):2805-2811.
20. Koch RM, Swiger LA, Chambers D, Gregory KE: **Efficiency of Feed Use in Beef Cattle**. *Journal of Animal Science* 1963, **22**:486-494.
21. Herd RM, Archer JA, Arthur PF: **Reducing the cost of beef production through genetic improvement in residual feed intake: Opportunity and challenges to application**. *Journal of Animal Science* 2003, **81**:E9-E17.
22. Archer JA, Richardson EC, Herd RM, Arthur PF: **Potential for selection to improve efficiency of feed use in beef cattle: a review**. *Aust J Agr Res* 1999, **50**(2):147-161.
23. Crews DH, Jr.: **Genetics of efficient feed utilization and national cattle evaluation: a review**. *Genet Mol Res* 2005, **4**(2):152-165.
24. Nkrumah JD, Sherman EL, Li C, Marques E, Crews DH, Bartusiak R, Murdoch B, Wang Z, Basarab JA, Moore SS: **Primary genome scan to identify putative quantitative trait loci for feedlot growth rate, feed intake, and feed efficiency of beef cattle**. *Journal of Animal Science* 2007, **85**(12):3170-3181.
25. Barendse W, Reverter A, Bunch RJ, Harrison BE, Barris W, Thomas MB: **A validated whole-genome association study of efficient food conversion in cattle**. *Genetics* 2007, **176**(3):1893-1905.
26. Bolormaa S, Hayes BJ, Savin K, Hawken R, Barendse W, Arthur PF, Herd RM, Goddard ME: **Genome-wide association studies for feedlot and growth traits in cattle**. *J Anim Sci* 2011, **89**(6):1684-1697.
27. Garbe JR, Elsik CG, Antoniou E, Reecy JM, Clark KJ, Venkatraman A, Kim JW, Schnabel RD, Michael Dickens C, Wolfinger RD *et al*: **Development and application of bovine and porcine oligonucleotide arrays with protein-based annotation**. *J Biomed Biotechnol* 2010, **2010**:453638.

28. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models**. *J Comput Biol* 2001, **8**(6):625-637.

29. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.

30. Storey JD: **A direct approach to false discovery rates**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002, **64**(3):479-498.

31. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

32. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

33. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-515.

34. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, **18**(9):1509-1517.

35. Loor JJ, Dann HM, Everts RE, Oliveira R, Green CA, Guretzky NA, Rodriguez-Zas SL, Lewin HA, Drackley JK: **Temporal gene expression profiling of liver from periparturient dairy cows reveals complex adaptive mechanisms in hepatic function**. *Physiol Genomics* 2005, **23**(2):217-226.

36. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Jr., Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK *et al*: **Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows**. *BMC Genomics* 2011, **12**:408.

37. Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C: **Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping**. *BMC Genomics* 2011, **12**(1):557.

38. Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S *et al*: **Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds**. *Science* 2009, **324**(5926):528-532.

39. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

40. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**(16):e164.

41. **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.

42. Yang Z, Nielsen R: **Synonymous and nonsynonymous rate variation in nuclear genes of mammals**. *J Mol Evol* 1998, **46**(4):409-418.

43. Kumar S: **Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates**. *Genetics* 1996, **143**(1):537-548.

44. Aigner B, Renner S, Kessler B, Klymiuk N, Kurome M, Wunsch A, Wolf E: **Transgenic pigs as models for translational biomedical research**. *J Mol Med (Berl)* 2010, **88**(7):653-664.

45. Lunney JK: **Advances in swine biomedical model genomics**. *Int J Biol Sci* 2007, **3**(3):179-184.

46. Matsunari H, Nagashima H: **Application of genetically modified and cloned pigs in translational research**. *J Reprod Dev* 2009, **55**(3):225-230.

47. Clark KJ, Carlson DF, Fahrenkrug SC: **Pigs taking wing with transposons and recombinases**. *Genome Biol* 2007, **8 Suppl 1**:S13.

48. Grizot S, Epinat JC, Thomas S, Duclert A, Rolland S, Paques F, Duchateau P: **Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds**. *Nucleic Acids Research* 2010, **38**(6):2006-2018.

49. Kim YG, Cha J, Chandrasegaran S: **Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain**. *Proc Natl Acad Sci U S A* 1996, **93**(3):1156-1160.

50. Boch J: **TALEs of genome targeting**. *Nat Biotechnol* 2011, **29**(2):135-136.

51. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM: **RNA-guided human genome engineering via Cas9**. *Science* 2013, **339**(6121):823-826.

52. Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MA, Harlizius B, Lee KT, Milan D, Rogers J, Rothschild MF *et al*: **Pig genome sequence--analysis and publication strategy**. *BMC Genomics* 2010, **11**:438.

53. Walters EM, Wolf E, Whyte JJ, Mao J, Renner S, Nagashima H, Kobayashi E, Zhao J, Wells KD, Critser JK *et al*: **Completion of the swine genome will simplify the production of swine as a large animal biomedical model**. *BMC Med Genomics* 2012, **5**:55.

54. Amberger J, Bocchini CA, Scott AF, Hamosh A: **McKusick's Online Mendelian Inheritance in Man (OMIM)**. *Nucleic Acids Res* 2009, **37**(Database issue):D793-796.

55. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al*: **Ensembl 2013**. *Nucleic Acids Res* 2013, **41**(Database issue):D48-55.

56. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal--unified access to biological data**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W23-27.

57. Doyle EL, Booher NJ, Standage DS, Voytas DF, Brendel VP, Vandyk JK, Bogdanove AJ: **TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction**. *Nucleic Acids Res* 2012, **40**(Web Server issue):W117-122.

58. Carlson DF, Fahrenkrug SC, Hackett PB: **Targeting DNA With Fingers and TALENs**. *Mol Ther Nucleic Acids* 2012, **1**:e3.

59. Maeder ML, Thibodeau-Beganny S, Sander JD, Voytas DF, Joung JK: **Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays**. *Nat Protoc* 2009, **4**(10):1471-1501.

60. Sander JD, Dahlborg EJ, Goodwin MJ, Cade L, Zhang F, Cifuentes D, Curtin SJ, Blackburn JS, Thibodeau-Beganny S, Qi Y *et al*: **Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)**. *Nat Methods* 2011, **8**(1):67-69.

61. Sander JD, Reyon D, Maeder ML, Foley JE, Thibodeau-Beganny S, Li X, Regan MR, Dahlborg EJ, Goodwin MJ, Fu F *et al*: **Predicting success of oligomerized pool engineering (OPEN) for zinc finger target site sequences**. *BMC Bioinformatics* 2010, **11**:543.

62. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome Res* 2009, **19**(9):1630-1638.

63. Bianco E, Nevado B, Ramos-Onsins SE, Perez-Enciso M: **A deep catalog of autosomal single nucleotide variation in the pig**. *PLoS One* 2015, **10**(3):e0118867.

64. Austin MA, Hutter CM, Zimmern RL, Humphries SE: **Familial hypercholesterolemia and coronary heart disease: a HuGE association review**. *Am J Epidemiol* 2004, **160**(5):421-429.

65. Tan WS, Carlson DF, Walton MW, Fahrenkrug SC, Hackett PB: **Precision editing of large animal genomes**. *Adv Genet* 2012, **80**:37-97.