

University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Honors Theses, University of Nebraska-Lincoln

Honors Program


Spring 3-9-2018

Copy Number Variation in the Porcine Genome Detected from Whole-Genome Sequence

Rebecca Anderson

University of Nebraska-Lincoln

Follow this and additional works at: <https://digitalcommons.unl.edu/honorsthesis>

 Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), [Genetics Commons](#), [Genomics Commons](#), [Molecular Biology Commons](#), [Molecular Genetics Commons](#), and the [Structural Biology Commons](#)

Anderson, Rebecca, "Copy Number Variation in the Porcine Genome Detected from Whole-Genome Sequence" (2018). *Honors Theses, University of Nebraska-Lincoln*. 63.

<https://digitalcommons.unl.edu/honorsthesis/63>

This Article is brought to you for free and open access by the Honors Program at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Honors Theses, University of Nebraska-Lincoln by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Copy Number Variation in the Porcine Genome Detected from Whole-Genome Sequence

Presented to the College of Agriculture and Natural Resources at the University of
Nebraska-Lincoln in partial fulfillment of requirements for a Degree of Bachelor of
Science with Distinction and completion of the Honors Program

By

**Rebecca Anderson
Biochemistry**

The University of Nebraska- College of Agriculture and Natural Resources

May 2018

Completed under the supervision of S. Madhavan, Ph. D.
In the Department of Biochemistry

Acknowledgements

I would like to thank Dr. Brittney Keel for acting not only as a research and internship advisor, but also as a significant mentor during my undergraduate career. Her continuous support and encouragement to challenge myself cannot go unnoticed in my academic achievements. I would also like to thank the other scientists and research assistants at the U.S. Meat Animal Research Center in Clay Center, Nebraska. Additionally, I would like to thank Dr. Madhavan for acting as an additional research advisor. He has been a vital motivator during my undergraduate career in the UNL Biochemistry Department. He has made me realize the applicable knowledge biochemistry provides. Finally, thank you to my friends and family that have encouraged me relentlessly to pursue my greatest abilities.

Table of Contents

Acknowledgements.....	2
Abstract.....	4
1. Introduction.....	6
2. Background.....	8
a. Deoxyribose Nucleic Acid.....	8
b. Genome Sequencing.....	8
c. Sequencing Approaches.....	9
d. Applications of NGS.....	11
e. Paired-end vs. Single-end Sequence Reads.....	11
f. Genomic Coverage.....	12
d. Structural Variations.....	12
e. Copy Number Variations.....	13
f. CNV Detection Methods.....	14
3. Material and Methods.....	16
a. Sequencing of USMARC Swine.....	16
b. Sequence Data Processing.....	16
c. CNV Detection and Defining CNVRs.....	17
d. Gene Content and Ontology in CNVRs.....	18
4. Results and Discussion.....	19
a. Read Mapping.....	19
b. CNVR Discovery and Statistics.....	19
c. Function of CNV Genes.....	19
d. Overlap with Known Quantitative Trait Loci.....	21
5. Conclusion.....	22
6. References.....	23
7. Supporting Information.....	28

List of Supporting Information

Figure 1. Approaches to Detect CNVs from NGS Short Reads.....	25
Figure 2. Distribution of Copy Number Variations Across the Genome.....	26
Figure 3. Enrichment Analysis of Molecular Function Gene Ontology Terms	27
Figure 4. Enrichment Analysis of Biological Function Gene Ontology Terms	27
Figure 5. Enrichment Analysis of Cellular Function Gene Ontology Terms	28
Appendix A. Library Preparation.....	29
Appendix B. Perl Script that Formatted Final CNVR Tables.....	30
Appendix C. Perl Script that Identified Gene Overlaps.....	31

Abstract

Copy number variations (CNVs) are large insertions, deletions, and duplications in the genome that vary between individuals in a species. These variations are known to impact a broad range of phenotypes from molecular-level traits to higher-order clinical phenotypes. CNVs have been linked to complex traits in humans such as autism, attention deficit hyperactivity disorder, nervous system disorders, and early-onset extreme obesity. In this study, whole-genome sequence was obtained from 72 founders of an intensely phenotyped experimental swine herd at the U.S. Meat Animal Research Center (USMARC) in Clay Center, Nebraska. This included 24 boars (12 Duroc and 12 Landrace) and 48 sows (Yorkshire-Landrace composites) for a total of 72 swine animals. Copy number variations were identified and analyzed using next generation sequencing and bioinformatics software. A total of 4566 copy number variations regions (CNVRs) were discovered in this study, covering 3.02% of the swine genome. A total of 593 genes were overlapped by CNVRs. These genes were further analyzed to determine function and relevance. Enrichment analysis determine function of CNVRs included sensory perception of smell (OR4D10), G-protein coupled receptor signaling pathway, cellular response to stimulus, and cell communication Quantitative trait loci (QTL) that were discovered included carcass weight (hot), average daily gain, fat-to-meat ratio, estimated carcass lean content, and birth weight.

Introduction

Variations in the genome, i.e. genetic mutations, are permanent changes in the chemical structure of the genome that result in variation in observed phenotypes and diseases within an organism. Whole-genome sequencing is performed to determine the complete DNA sequence of an organism and to identify structural variation. Genomic sequencing has become increasingly popular in the livestock industry as it allows the discovery of genetic variants underlying economically important traits, such as reproduction, feed efficiency, product quality, and disease resistance and susceptibility.

Copy number variations (CNVs) are large insertions, deletions, and duplications in the genome that are classified as being greater than 1 kilobases in length [1]. CNV studies have been performed in a number of animals; including cattle [2][3], dogs [4][5], sheep [6][7], pig [8][9], chicken [10][11], and goat [12] [13]. Analyzing CNVs, their location, and gene overlap allows for determination of their effects on the phenotype [2].

Analysis of the human genome and the impact of mutations have contributed significantly to the understanding of numerous diseases and phenotypes. A significant portion of research has been focused on single-nucleotide polymorphisms (SNPs), yet copy-number variation findings have become crucial in efforts to characterize genetic underpinnings of psychiatric and neurodevelopmental [14]. CNV analysis in human populations has discovered that a CNV encompassing the *UGT2B17* gene is associated with osteoporosis, graft-versus-host disease (GVHD), and variability in testosterone glucuronidation rate [14]. Other research has reported common copy number polymorphisms associated with several complex diseases phenotypes, including HIV acquisition and progression, lupus glomerulonephritis and three

systemic autoimmune diseases (systemic lupus erythematosus, microscopic polyangiitis and Wegener's granulomatosis) [15].

A major goal of livestock genomics research is to identify the genetic differences that are responsible for variations in phenotypic traits that are economically significant. For example, cattle are an important source of meat, milk, and other goods that are provided to millions around the world. Cattle have been selectively bred to increase meat and milk production, all a result of improved genetics [16]. CNV research may provide insight to development of more accurate tools for genomic selection. CNVs in Black Angus cattle have been associated with growth and immune system process, while Holstein CNVs were associated with reproduction and enzyme regulator activity [16]. Therefore, insight into CNV within livestock can be impactful to breeding selections and the future of meat, milk, and other goods production of numerous livestock species.

Background

Deoxyribose Nucleic Acid

Deoxyribose nucleic acid (DNA) provides important information for an organism to develop, survive, and reproduce. It is composed of a sequence of single and double-ringed nitrogenous bases bonded to a repeating sugar-phosphate backbone. The possible nitrogenous bases include adenine, cytosine, guanine, and thymine. Double-stranded DNA complements itself by pairing adenine and thymine bases together, as well as cytosine and guanine. Lengths of DNA are measured in base pairs- one base pair being a base that has been paired with its complementary partner. The overall combination of base, sugar, and phosphate composes a nucleotide [18]. A small portion of DNA contains genes- sequences that code for specific amino acids to produce a protein- and an even smaller portion within the genes encode for amino acids.

Genome Sequencing

Genome sequencing is the identification of the nucleotides in order in a genome. This chemical alphabet can then be interpreted and analyzed. The purpose of genomic analysis is to understand the underlying biology; not only the structural components that make up the genome, but also the functions of individual genes, the process by which genes work together, and the role that genetic mutations play in the phenotype. Several factors come into play regarding the effectiveness of genome sequencing, such as read-length, accuracy, and perhaps most importantly cost [19]. These factors have guided the performance and direction of genome sequencing technologies.

Sequencing approaches

The Sanger method of sequencing was one of the earliest methods of sequencing. Also known as chain termination, the Sanger method uses dideoxynucleotides (ddNTPs) along with the normal nucleotides found in the DNA. The DNA fragment is prepared by shotgun *de novo* or PCR amplification, where both approaches result in an amplified template. Dideoxynucleotides contain only a hydrogen on the 3' carbon rather than a hydroxyl group. This integration of ddNTPs prevents the addition of nucleotides via hydrogen bonding- thus terminating the DNA chain. This method allows for the nucleotides to be identified since the termination occurs at all positions where the same nucleotide is required. The DNA is then processed through electrophoresis and exposed to UV light. The label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Software then translates the identity into DNA sequence. With use over three decades, this process gradually improved and was capable of achieving read lengths of 1,000 base pairs. In the current context, Sanger sequencing is estimated to cost \$0.50 per kilobase [21]. This method has been classified as 'first generation' technology as it was one of the first widely adopted sequencing techniques.

Second-generation DNA sequencing resulted from the advances of Sanger sequencing and the high demand for low-cost sequencing. These strategies are grouped into the following categories: microelectrophoretic methods, sequencing by hybridization, real-time observation of single molecules, and cyclic-array sequencing. These approaches differ in how the array is generated, yet they have similar workflow- random fragmentation of the DNA is accomplished to prepare the library followed by *in vitro* ligation of common adaptor sequences. Array-based sequencing allows a greater degree of parallelism, produce hundreds of millions of sequencing

reads, and can be enzymatically manipulated easily. Limitations to this approach include bioinformatic challenges, read-lengths, and subpar raw accuracy [22].

Newer methods of sequencing have been classified as ‘next-generation’ sequencing (NGS). NGS generates DNA sequence data that is more complete and accurate than with previous methods. This approach can deliver data output ranging from 300 kilobases up to 1 terabase in a single run. There are several ‘next-generation’ innovations, such as sequencing-by-synthesis, sequencing-by-ligation, ion semiconductor sequencing, and others. Though, the most predominant method is sequencing-by-synthesis due the success of genetic medicine using Illumina devices. Sequencing by synthesis uses fluorescently labeled nucleotides to sequence clusters on the cell surface in parallel. During each cycle, a labeled ddNTP is added to the nucleic acid chain. The dye is then identified through laser excitation and imaging. This method reduces sequence-context-specific errors [23].

The critical difference between early approaches and NGS is that instead of sequencing a single DNA fragment, NGS extends this process across millions of fragments in a parallel manner. The four basic steps of NGS sequencing are the following: 1) library preparation- random fragmentation of the DNA, 5’ and 3’ adapter ligation, PCR amplification and gel purification; 2) cluster generation- library is loaded, fragments are captured by surface-bound complementary oligonucleotides, fragments amplified into clusters through bridge amplification (a PCR technique that embeds DNA on a surface while cloning); 3) sequencing- detection of single bases as they are incorporated into DNA template strands, identification of nucleotides by fluorophore excitation; 4) data analysis- sequence reads are aligned to reference genome to be utilized in downstream analyses [23].

Applications of NGS

NGS can be applied to numerous fields of study, such as genomics, transcriptomics, and epigenomics. Within genomics, NGS allows for whole-genome sequencing, as well as targeted sequencing such as exome sequencing. In transcriptomics, sequencing of total RNA and mRNA, targeted RNA, and even small RNA and noncoding RNA is available. Methylation sequencing, ChiP sequencing, and ribosome profiling can be completed for the field of epigenomics. The ability for NGS to parallel sequence has enabled advancements within numerous fields.

Paired-end vs. Single-end Sequence Reads

Single-end sequence reads are produced by reading a fragment from one end to the other while generating the sequence of base pairs. In comparison, paired-end sequence reads involves sequencing both ends of the DNA fragments in a sequencing library and aligning the forward and reverse reads as read pairs. Paired-end reading improved the ability to identify the relative positions of various reads in the genome [24].

Paired-end sequencing with NGS allows for sequencing in high throughput. This allows for twice the amount of reads for the same amount of time, as well as more accurate read alignment and detection of indels. NGS sequencing allows for high-throughput at a moderate cost [25]. The Illumina HiSeq- a highly utilized NGS machine- has been estimated to cost as little as \$41 per gigabase with a throughput of 600 gigabases [26].

Genomic Coverage

Sequence reads are not evenly distributed over the entire genome due to the random and independent manner of sequencing. As a result, multiple observations per base is required to

come to a reliable base call. This is required since read lengths are short and an error is difficult to distinguish from a sequence variant in the sample. Genomic coverages are the average number of reads that align to the bases in the reference genome. Most often, the desired genomic coverage for an experiment depends on the application. Higher coverage of a sequence inevitably results in more reliable data but at a greater cost.

Structural Variation

Structural variation refers to large scale structural differences in the genomic DNA of an organism. It is a result of chromosomal rearrangement via deletions, small kilobase duplications, inversion, translocations, insertions, recombination, single nucleotide polymorphism (SNP), and tandemly repeated DNA. Some genetic diseases are suspected to be caused by structural variations.

Copy Number Variations

Copy number variations (CNVs) are large insertions, deletions, and duplications in the genome that vary between individuals of a species. They are classified as being at least 1000 bases or greater. These variations are known to impact a broad range of phenotypes from molecular-level traits to higher-order clinical phenotypes. Copy number variations formation occurs by non-homologous end joining, non-allelic homologous recombination, transposition of transposable elements or pseudogenes, variable numbers of tandem repeats, and replication errors [27].

Copy number variation has emerged as an important type of genetic risk factor for developmental disorders in humans, including the neurodevelopmental disorders schizophrenia,

autism and mental retardation [28]. Recently, genome-wide CNV surveys have associated a number of CNVs with early-onset extreme obesity [29]. Research has shown significant association of a CNV encompassing the salivary amylase gene (*AMY1*) with BMI and obesity, providing the first link between carbohydrate metabolism and BMI [29]. Additionally, recent studies have shown that there are CNV distribution hotspots in the human genome. The following categories of genes are enriched for structural variants in humans: 1) genes involved in immunity and signaling, 2) genes encoding proteins involved with the environment (immune response, perception of smell), 3) retrovirus and transposition related protein coding genes [30].

There have been findings in regards to CNVRs in the livestock industry. Studies have shown breed specific CNVRs in cattle. A particular CNVR adjacent to the *KIT* gene has been linked to the white face in Hereford cattle [31]. Other studies involving numerous cattle breeds have shown CNVR significantly enriched for immunity, lactation, reproduction and rumination [32].

CNV Detection Methods

Several methods for detecting CNV in NGS data exist and have been previously reviewed [33]. Early methods of detection include fluorescence *in situ* hybridization and array comparative genomic hybridization. *In situ* hybridization uses labeled complementary DNA to localize a specific DNA sequence which a reporter molecule is attached. Fluorescence is then used to identify the location of the reporter molecule [34]. Comparative genomic hybridization used a test and control genome. They are differentially labeled and hybridized to metaphase chromosomes. The fluorescent signal intensity of the labeled test DNA relative to the reference allows for identification of structural variations [35].

There are currently three main strategies for CNV detection through NGS technology. Those strategies include (1) paired-end mapping, (2) split read, and (3) read depth (Figure 1) [36]. Paired-end mapping utilizes the fact that DNA fragments from the same library preparation protocol have a specific distribution of insert size, i.e. the distance between read pairs. In paired-end reading, knowing the read length between the two ends allows for detection of genomic rearrangements and repetitive sequence elements. Paired-end mapping then identifies both single nucleotide polymorphism and copy number variations from discordantly mapped paired-reads whose distances are significantly different than the average insert size [36].

The split read-based approach provides the precise location, size and types of variants found in a genome. It is a powerful method for finding small and medium-sized insertions, deletions, and inversions [37]. This method reads pairs in which one pair is aligned to the reference genome while the other fails to map or only partially maps to the genome. This is due the sequencing DNA from only one end. This method splits the incompletely mapped reads into multiple fragments. This approach delivers large volumes of high-quality data, rapidly and economically. This method is not commonly used due to the occurrence of false positives or false negative results and is limited with large variants or those in repetitive regions [37].

The read depth-based approach is based on the hypothesis that the depth of coverage in a genomic region is correlated with the copy number of the region. This method can detect the exact copy number, while the previous methods only gave an estimate based on position information [36]. In general, read-depth methods assume a random distribution in mapping depth. Read-depth methods are more effective for larger (> 1 kb) CNVs. Challenges with this method include the inability to identify copy number neutral variants like inversions [37].

Material and Methods

The DNA samples sequenced for this study were extracted from semen collected by commercial AI services and from blood and tail tissue archived under standard operating procedures for the U.S. Meat Animal Research Center (USMARC) tissue repository. The search did not involve experimentation on animals requiring IACUC approval

Sequencing of USMARC Swine

Blood or semen samples were obtained from the 72 of the founders of a USMARC composite swine population. These animals included 12 Landrace boars, 12 Duroc boars, and 48 Yorkshire-Landrace composite sows. Genomic DNA was extracted from semen and blood using standard DNA extraction protocols (standard phenol-chloroform extraction for semen and salt extraction for blood samples). Genomic DNA was sheared to 300-500bp using a Covaris S220 ultrasonicator (Woburn, MA, USA) and libraries prepared using TrueSeq DNA sample prep kit, version 2 (Illumina, San Diego, CA) were sequenced using a HiSeq2500 (Illumina Inc., San Diego, CA, USA) at the Iowa State DNA Core Facility (Ames, IA, USA) and at DNA Landmarks (Saint-Jean-sur-Richelieu, Quebec, CN). The bases of the resulting 100 cycle paired-end reads were identified with the Illumina BaseCaller and fastq files were produced for downstream analysis of the sequence data. (Appendix A explains library preparation).

Sequencing Data Processing

The fastq file contains base call and quality information for all reads passing filtering as it is used as sequence input for alignment and other secondary analysis softwares used in bioinformatics. The format includes four lines, which are sequence identifier, sequence, quality

score identifier line, and quality score [38]. The fastq files were processed as follows: The trimmomaticPE software (version 0.35) was used to trim Illumina adapter sequences and low-quality bases [39]. After quality filtering, the remaining reads were mapped to the Sscrofa 10.2 genome using Burrow-Wheeler Alignment (BWA) (version 0.7.12 with the default parameters) [40] [41]. The Burrows-Wheeler Alignment software is an algorithm that maps low-divergent sequences against a large reference genome. All output SAM files were converted to sorted BAM files using SortSam from Picard (version 1.1; <http://broadinstitute.github.io/picard/>), and duplicates in the BAM files were marked by applying MarkDuplicates from Picard. MarkDuplicates locates and tags duplicate reads which originated from the same fragment of DNA. Genomic coverage for each of the BAM files was computed using SAMtools version 1.3 [42].

CNV Detection and Defining CNVRs

In order for putative copy number variations to be detected and defined from the 72 pigs, BAM files were run through the cn.MOPS program to construct a set of copy number variable regions [43]. cn.MOPS is a “relative” CNV detection algorithm that applies a Bayesian approach to decompose read variation across multiple samples into integer copy numbers and noise by its mixture components and Poisson distributions, respectively. cn.MOPS avoids read count biases along the chromosomes by modeling the depth of coverage across all samples at each genomic position. This approach is known to have a lower false-positive rate than other CNV detection methods [43]. The program was run using a window length of 2500, mean normalization mode, and default values for all other parameters.

A copy number variation region (CNVR) was constructed by merging CNVs within and across samples that exhibited at least 50% pairwise reciprocal overlap in their genomic coordinates. For example, suppose we have two CNVs, CNV1 beginning at position a and ending at position b and CNV2 running from c to d with $a < c < b < d$. If the reciprocal overlap between the two CNVs is at least 50% then they are merged into a CNVR which runs from a to d . The final table of CNVRs was formatted using a custom Perl script (see Appendix B).

Gene content and Ontology in CNVRs

Genes from the Ensembl genome annotation of *Sus scrofa* 10.2 overlapping with CNVRs were identified using a custom Perl script (see Appendix C). Functions of genes containing detected variants were determined using the PANTHER classification system (Version 10.0) [44]. Enrichment analysis of gene function was performed using PANTHER's implementation of the binomial test of overrepresentation. PANTHER (Protein ANalysis THrough Evolutionary Relationships) classifies proteins and their corresponding genes through high-throughput analysis. Gene ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. Significance of GO terms was assessed using the default Ensembl *Sus scrofa* GO annotation as background for the enrichment analysis. Data from PANTHER were considered statistically significant at Bonferroni corrected P -value < 0.05 .

Results and Discussion

Read Mapping

Genomic coverage for the 72 samples ranged from 1.15x – 21.11x with an average of 5.75x. When the NGS sequence data was generated approximately 10x genomic coverage for the boars and 3x coverage for the sows were targeted. A mean genomic coverage of 11.7x and 3.2x for boars and sows respectively, was obtained. The observed variations from the target coverage may be due to several different technical aspects, including the stochasticity of NGS technology, DNA quality, or library preparation.

CNVR Discovery and Statistics

Copy number variations were identified from the 72 pigs using the cn.MOPS algorithm. CNVs were merged across the genome and across samples into CNVRs, and CNVRs were filtered out if they were not present in at least 2 samples. This criterion resulted in 4,566 CNVRs on the 18 autosomes of the sampled animals. Sizes of the CNVRs ranged from 7.5 Kb (kilobases) to 500 Kb, with an average of 22 Kb and median of 15 Kb. CNVRs occupied 3.02% of the Sscrofa 10.2 genome assembly. Of the identified CNVRs, 1,268 showed copy number gain, 1,982 copy number loss, and 1,316 showed both copy number loss and gain- a result of possible insertion and deletion in the same position of the chromosome. The distribution of CNVRs along the genome is pictured in Figure 2.

Function of CNV Genes

A total of 593 Ensembl genes from the Sscrofa 10.2 assembly overlapping with our CNVRs were identified. Gene ontology enrichment analysis through PANTHER indicated that

genes overlapped with CNVRs were mostly involved in receptor activity, catalytic activity, binding, transporter activity, apoptotic process, cellular process, multicellular organismal process, response to stimulus, membrane, cell part, organelle, and macromolecular complex.

Enrichment analysis of GO terms was used via the GO slim database. GO slim terms are a subset of the terms in the entire gene ontology that provides an overview of the ontology content. Analysis showed that the terms sensory perception of smell, G-protein coupled receptor signaling pathway, cellular response to stimulus, and cell communication were significantly enriched in the protein-coding of genes overlapped by CNVRs (Bonferroni corrected P -value < 0.05). Results from the GO slim analysis are shown in Figures 2-4.

The Humane Society has published the evolution of the pig and displayed how sensory organs have become critical for this specie's survival [45]. They adapt their diet to the seasonal availability of edible plant food in their home ranges. Although pigs subsist primarily on plant matter, they are omnivores and supplement their diets with earthworms, insects, amphibians, reptiles, and rodents in the wild [45]. Additionally, a pig's snout provides heightened senses to navigate and interact with the environment. This part of the animal is designed for rooting in the ground in searching for food. The numerous sensory receptors that innervate the snout provides pigs with an well-developed sense of smell. This species has developed a strong method of communication using olfactory signals through their saliva and urine called pheromones. This is essential for reproduction and a sow's maternal behavior with her offspring [45]. *S. scrofa* has been identified to have one of the largest olfactory receptor repertoires. The significant number of unique and expanding olfactory receptor genes in the pig genome may suggest or provide insight to the presence of swine specific olfactory stimulation [46].

Previous studies have found that olfactory receptors (OR) are found in CNV regions. ORs are seven-transmembrane G protein-coupled receptors that compose one of the largest gene families in mammalian genomes [47]. Variations in OR genes can result in partial or total insensitivity to certain odorants. It is interesting to note that in humans a subset of ORs could function outside the olfactory system. An example would be *OR1D2*, which has been found to mediate sperm chemotaxis toward its ligand; therefore, impactful to male fertility [Gilad]. It has also been discovered that a low olfactory copy number have an early age of onset of Alzheimer disease [48].

Overlap with Known Quantitative Trait Loci

Quantitative trait loci analysis is a statistical method that links phenotypic data, such as traits, and genotypic data, usually molecular markers, in an attempt to explain the genetic basis of variation in complex traits [49]. To reveal the potential relationships between CNVR and QTL, we analyzed the overlap between our CNVRs and known swine QTL. Swine QTL from the Sscrofa 10.2 genome build were downloaded from the Animal QTL database (<http://www.animalgenome.org/cgi-bin/QTLdb/SS/index>). The most frequent included carcass weight (hot), average daily gain, fat-to-meat ratio, estimated carcass lean content, and birth weight. These regions will require further analysis to gain a better understanding of the impact of the CNV may play in QTL.

Conclusion

Genomic research has concentrated on single nucleotide polymorphisms as the most relevant source of structural variation in the genome. However, studies have linked changes in copy number to complex diseases and unique phenotypic traits. Therefore, copy number variations have a role in reshaping gene structure, modulating gene expression, and contributing to phenotypic variations that may impact future research discoveries. Additional research is required since evolutionary and functional aspects of the copy number variations in organisms is not completely understood.

In this study, we examined whole-genome sequences from 72 of the founders of a heavily phenotyped experimental swine herd at the U.S. Meat Animal Research Center. Findings identified 4566 copy number variations from 24 boars and 48 sows in the sampled population. Genes overlapped by CNVs were enriched for sensory perception, G-protein coupled receptors, and cellular response to stimuli. Additionally, CNVs overlapped with many QTL for economically relevant traits, which included carcass weight (hot), average daily gain, fat-to-meat ratio, estimated carcass lean content, and birth weight.

References

1. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., ... Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. <http://doi.org/10.1038/nature05329>
2. Keel, B. N., Lindholm-Perry, A. K., & Snelling, W. M. (2016). Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome. *Frontiers in Genetics*, *7*, 207. <http://doi.org/10.3389/fgene.2016.00207>
3. Gao, Y., Jiang, J., Yang, S., Hou, Y., Liu, G. E., Zhang, S., ... Sun, D. (2017). CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics*, *18*, 265. <http://doi.org/10.1186/s12864-017-3636-3>
4. Berglund, J., Nevalainen, E. M., Molin, A., Perloski, M., André, C., Zody, M. C., ... Webster, M. T. (2012). Novel origins of copy number variation in the dog genome. *Genome Biology*, *13*(8). doi:10.1186/gb-2012-13-8-r73
5. Molin, A., Berglund, J., Webster, M. T., & Lindblad-Toh, K. (2014). Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. *BMC Genomics*, *15*(1), 210. doi:10.1186/1471-2164-15-210
6. Jenkins, G. M., Goddard, M. E., Black, M. A., Brauning, R., Auvray, B., Dodds, K. G., ... McEwan, J. C. (2016). Copy number variants in the sheep genome detected using multiple approaches. *BMC Genomics*, *17*(1). doi:10.1186/s12864-016-2754-7
7. Liu, J., Zhang, L., Xu, L., Ren, H., Lu, J., Zhang, X., ... Du, L. (2013). Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC Genomics*, *14*, 229. <http://doi.org/10.1186/1471-2164-14-229>
8. Chen, C., Qiao, R., Wei, R., Guo, Y., Ai, H., Ma, J., ... Huang, L. (2012). A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics*, *13*(1), 733. doi:10.1186/1471-2164-13-733
9. Fadista, J., Nygaard, M., Holm, L., Thomsen, B., & Bendixen, C. (2008). A Snapshot of CNVs in the Pig Genome. *PLoS ONE*, *3*(12). doi:10.1371/journal.pone.0003916
10. Jia, X., Chen, S., Zhou, H., Li, D., Liu, W., & Yang, N. (2012). Copy number variations identified in the chicken using a 60K SNP BeadChip. *Animal Genetics*, *44*(3), 276-284. doi:10.1111/age.12009
11. Wang, X., & Byers, S. (2014). Copy Number Variation in Chickens: A Review and Future Prospects. *Microarrays*, *3*(1), 24–38. <http://doi.org/10.3390/microarrays3010024>

12. Fontanesi, L., Martelli, P., Beretti, F., Riggio, V., Dallolio, S., Colombo, M., . . . Portolano, B. (2010). An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics*, *11*(1), 639. doi:10.1186/1471-2164-11-639
13. Fontanesi, L., Beretti, F., Riggio, V., González, E. G., Dall'Olio, S., Davoli, R., . . . Portolano, B. (2009). Copy Number Variation and Missense Mutations of the Agouti Signaling Protein (*ASIP*) Gene in Goat Breeds with Different Coat Colors. *Cytogenetic and Genome Research*, *126*(4), 333-347. doi:10.1159/000268089
14. Gamazon, E. R., & Stranger, B. E. (2015). The impact of human copy number variation on gene expression. *Briefings in Functional Genomics*, *14*(5), 352–357. <http://doi.org/10.1093/bfgp/elv017>
15. Mccarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, *39*(7s). doi:10.1038/ng2080
16. Stothard, Paul, et al. “Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery.” *BMC Genomics*, vol. 12, no. 1, 2011, doi:10.1186/1471-2164-12-559.
17. Wang J, Jiang J, Wang H, Kang H, Zhang Q, Liu J-F (2014) Enhancing Genome-Wide Copy Number Variation Identification by High Density Array CGH Using Diverse Resources of Pig Breeds. *PLoS ONE* 9(1): e87571. doi:10.1371/journal.pone.0087571.
18. Eccles, D. A. (2011). *Genomic analysis of human population structure a thesis submitted to the Victoria University of Wellington in fulfilment of the requirements for the degree of Doctor Philosophy in Biomedical Science* (Unpublished master's thesis). Victoria University of Wellington.
19. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. <http://doi.org/10.1016/j.ygeno.2015.11.003>
20. Janitz, M. (2008). *Next-generation genome sequencing: Towards personalized medicine*. Weinheim: Wiley-VCH.
21. Shendure, J., & Ji, H. (2008, October 9). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135-1145. Retrieved from http://www.nature.com/nbt/journal/v26/n10/full/nbt1486.html?type=access_denied
22. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. <http://doi.org/10.1007/s13353-011-0057-x>
23. An Introduction to Next-Generation Sequencing Technology. (n.d.). Retrieved June 10, 2017, from http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

24. “Genome Center Home Genome Center.” *Genome Sequencing: Defining Your Experiment | Columbia University Department of Systems Biology*, systemsbiology.columbia.edu/genome-sequencing-defining-your-experiment.
25. Metzker, M. L. (2010, January). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*, 31-46. Retrieved from <http://www.nature.com/nrg/journal/v11/n1/full/nrg2626.html>
26. Comparing Price and Tech. Specs. of Illumina MiSeq, Ion Torrent PGM, 454 GS Junior, and PacBio RS. (2012, August 05). Retrieved July 21, 2016, from <http://nextgenseek.com/2012/08/comparing-price-and-tech-specs-of-illumina-miseq-ion-torrent-pgm-454-gs-junior-and-pacbio-rs/>
27. Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481. <http://doi.org/10.1146/annurev.genom.9.081307.164217>
28. O'donovan, Michael C., George Kirov, and Michael J. Owen. "Phenotypic Variations on the Theme of CNVs." *Nature Genetics Nat Genet* 40.12 (2008): 1392-393.
29. Aerts, E., Beckers, S., Zegers, D., Van Hoorenbeeck, K., Massa, G., Verrijken, A., ... Van Hul, W. (2016). CNV analysis and mutation screening indicate an important role for the NPY4R gene in human obesity. *Obesity*, *24*(4), 970–976. doi:10.1002/oby.21435
30. Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics*, *45*(1), 1-16. doi:10.1152/physiolgenomics.00082.2012
31. Keel, B. N., et al. “Genome-Wide copy number variation in the bovine genome detected using low coverage sequence of popular beef breeds.” *Animal Genetics*, vol. 48, no. 2, 2016, pp. 141–150., doi:10.1111/age.12519.
32. Hou, Yali, et al. “Genomic characteristics of cattle copy number variations.” *BMC Genomics*, vol. 12, no. 1, 2011, doi:10.1186/1471-2164-12-127.
33. Abel, H. J., & Duncavage, E. (2013). Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics*, *206*(12), 432–440. <http://doi.org/10.1016/j.cancergen.2013.11.002>
34. In Situ Hybridization (ISH). (2004). *Encyclopedic Dictionary of Genetics, Genomics and Proteomics*. doi:10.1002/0471684228.egp06373
35. Theisen, A. (2008) Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education* 1(1):45

36. Jia, P., Wang, Q., Wang, Q., Zhao, M., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*.
37. K, Ye, and Hall G. "Structural Variation Detection from Next Generation Sequencing." *Journal of Next Generation Sequencing & Applications*, vol. 01, no. S1, 2015, doi:10.4172/2469-9853.s1-007.
38. Fastq files- Illumina Support. (n.d.). Retrieved February 28, 2018, from http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm
39. Bolger, Anthony M., et al. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics*, vol. 30, no. 15, Jan. 2014, pp. 2114–2120., doi:10.1093/bioinformatics/btu170.
40. Groenen, Martein, et al. "Analyses of pig genomes provide insight into porcine demography and evolution." *Nature*, vol. 491, Nov. 2012, p. 393398.
41. Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-60.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009;25:2078
43. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D., Mitterecker, A., Bodenhofer, U., & Hochreiter, S. (2012). Cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, 40(9). doi:10.1093/nar/gks003
44. Mi, Huaiyu, et al. "PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees." *Nucleic Acids Research*, vol. 41, no. D1, 2012, doi:10.1093/nar/gks1118.
45. More about Pigs. (n.d.). Retrieved February 17, 2018, from http://www.humanesociety.org/animals/pigs/pigs_more.html?referrer=https%3A%2F%2Fwww.google.com%2F#snout
46. Nguyen, D., Lee, K., Choi, H., Choi, M., Le, M., Song, N., Park, C. (2012). The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. *BMC Genomics*, 13(1), 584. doi:10.1186/1471-2164-13-584
47. Gilad, Yoav. "Faculty of 1000 evaluation for Extensive copy-Number variation of the human olfactory receptor gene family." *F1000 - Post-Publication peer review of the biomedical literature*, Aug. 2008, doi:10.3410/f.1120705.576941.

48. “Olfactory copy number association with age at onset of Alzheimer disease.” *Neurology*, vol. 76, no. 22, 2011, pp. 1945–1945., doi:10.1212/wnl.0b013e318221c187.
49. Miles, C. & Wayne, M. (2008) Quantitative trait locus (QTL) analysis. *Nature Education* 1(1):208
50. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639-1645. doi:10.1101/gr.092759.109

Supporting Information

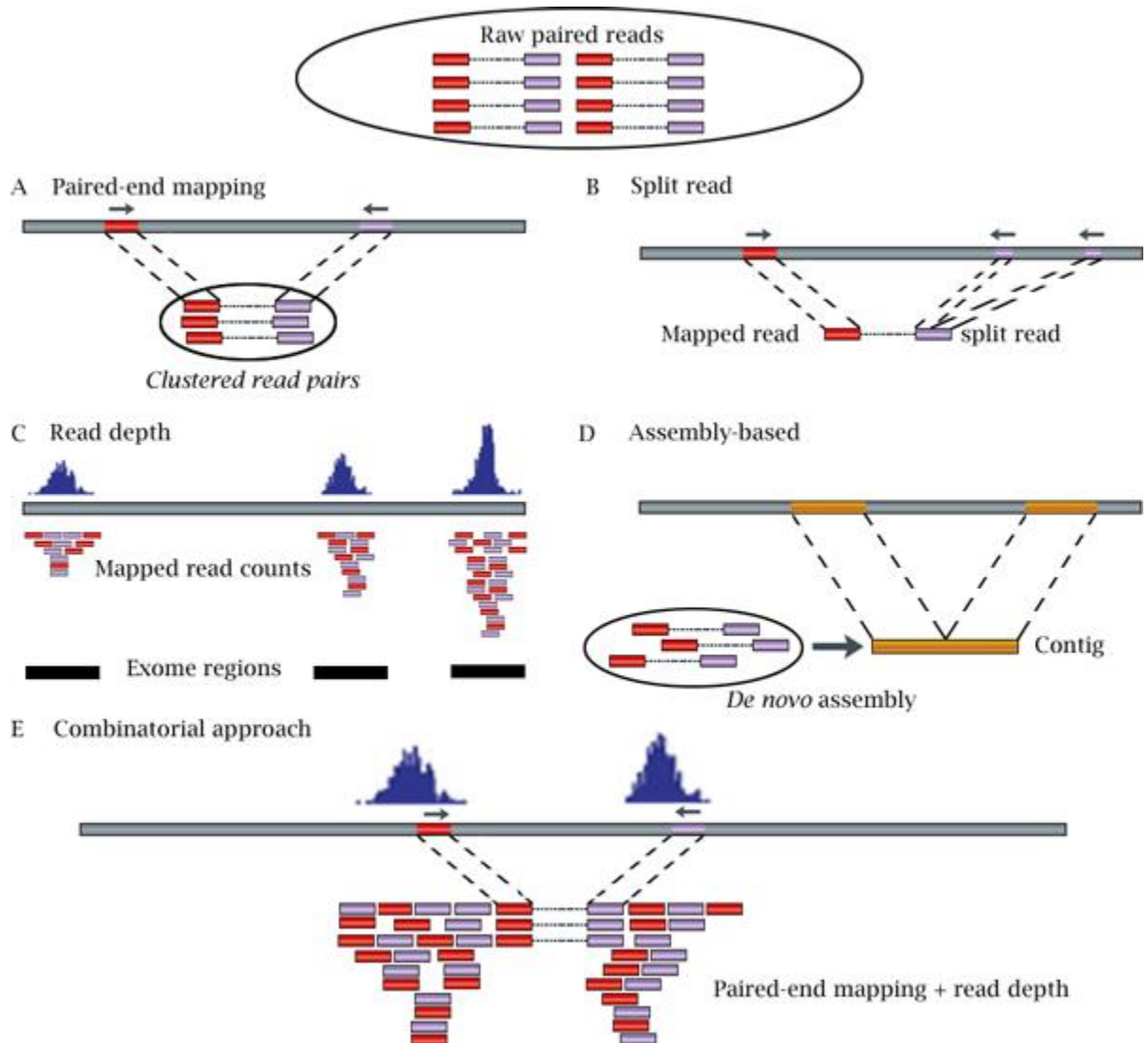


Figure 1. Approaches to detect CNVs from NGS short reads. A. Paired-end mapping strategy detects CNVs through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read pair is significantly different from the average insert size. B. Split read-based methods use incompletely mapped read from each read pair to identify small CNVs. C. Read depth-based approach detects CNV by counting the number of reads mapped to each genomic region. In the figure, reads are mapped to three exome regions. D. Assembly-based approach detects CNVs by mapping contigs to the reference genome. E. Combinatorial approach combines RD and PEM information to detect CNVs [36].



Figure 2. Distribution of Copy Number Variations Across the Genome

Positions of CNVR identified from the 181 sequenced swine genomes in Circos format [50]. The outer ideogram runs clockwise from chromosome 1 to chromosome Y with levels in Mb of physical distance. The copy number data is represented in the inner tracks. The two innermost tracks show scatter plots of the CNVR, where the red track shows copy number loss and the green track shows copy number gain. Concentric circles within these tracks indicate y-axis values in the scatter plot. The ten concentric circles in the red track mark values $0 \leq y < 2$, with 0 being the inner-most track, while the eleven concentric circles in the green track mark values $2 \leq y \leq 20$. The size of the dot in the scatter plot is proportional to the number of samples containing the CNVR. The other track shows a heat map which indicates the parts of the genome that contain copy number gain and loss. This plot simply collapses the scatter plot values onto a single radial position.

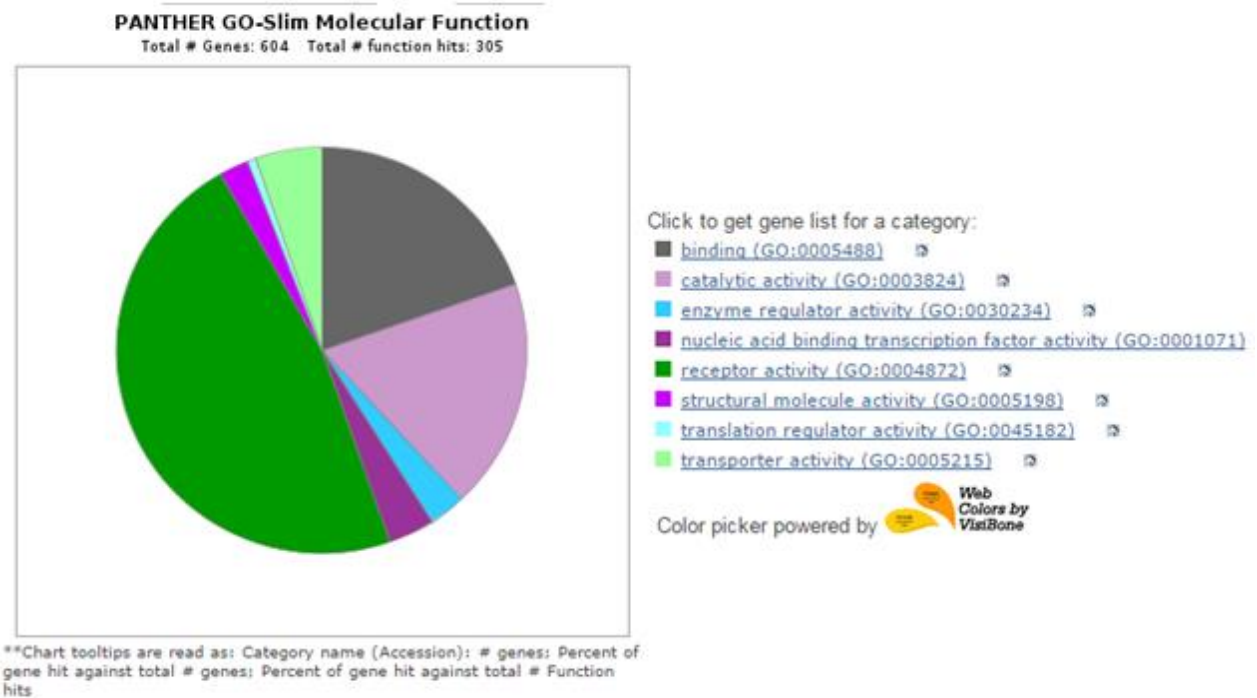


Figure 3. Enrichment Analysis of Molecular Function Gene Ontology Terms
The molecular function was analyzed using the PANTHER classification system.

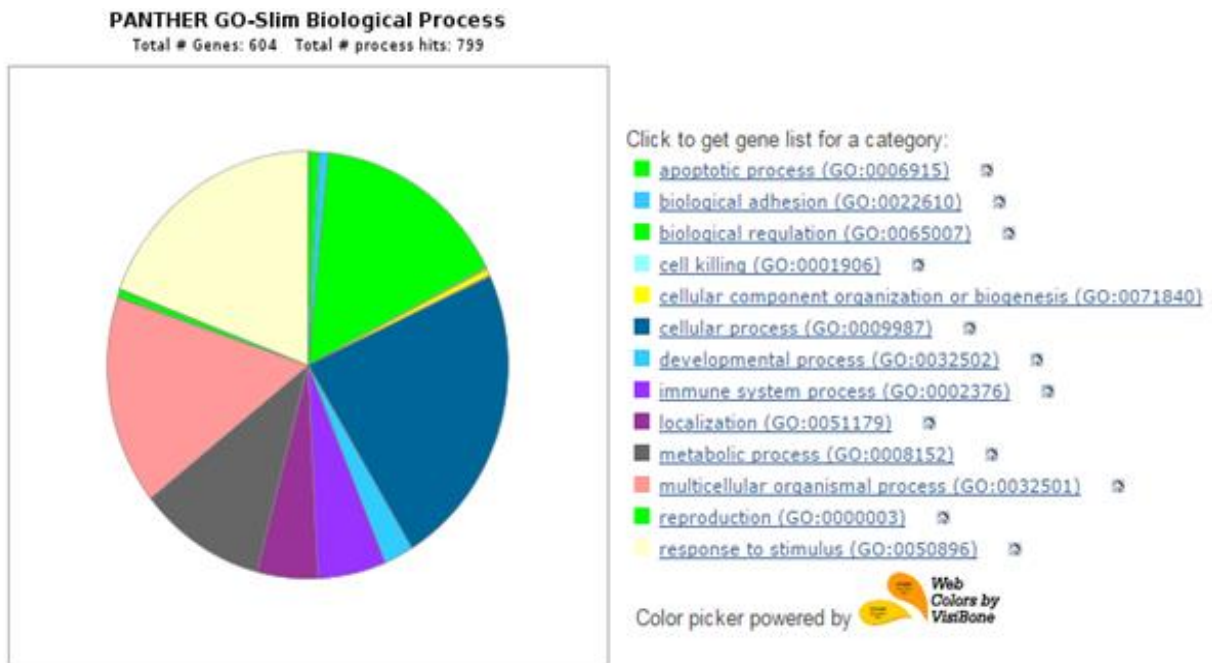


Figure 4. Enrichment Analysis of Biological Function Gene Ontology Terms
Biological function was analyzed using the PANTHER classification system.

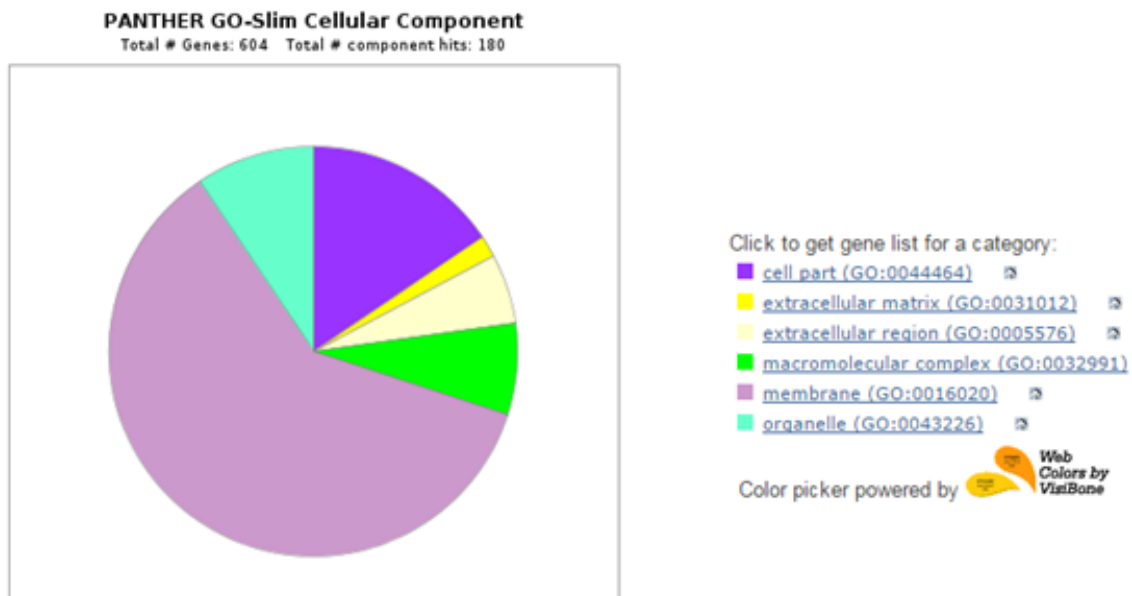


Figure 5. Enrichment Analysis of Cellular Function Gene Ontology Terms
 Cellular function was analyzed using the PANTHER classification system.

Appendixes

Appendix A.- Library Preparation

TrueSeq DNA sample prep kit (version 2) was the library preparation used for the given samples. I have been able to perform TruSeq DNA PCR-Free Library Prep kit (Revision D) during my time at US-Meat Animal Research Center. 80 blood samples were obtained from feed-efficiency steers being analyzed. Buffy coat was taken from the whole blood and the DNA was extracted. The 80 samples of extracted DNA were analyzed on a Nanodrop. 24 samples were selected to prepare libraries for whole-genome sequencing. The gDNA was normalized with resuspension buffer to a final volume of 55 microliters according to concentration, and then sheared on the Covaris instrument according to the instrument guidelines for a 350 base pair library. The fragmented DNA was cleaned using sample purification beads, magnetic stand, and 80% ethanol. Library size was selected and ends were repaired by using the provided end repair mix and a thermal cycle of 30 degC. Purification beads were diluted to 160 microliters per 100 microliters for large DNA fragments, immediately followed by the same procedure with concentrated purification beads for small DNA fragments. The 3' ends of the samples were adenylated using the A-Tailing Mix provided and a thermal cycle at 37, 70, and 4 degC. Ligase, control, and the corresponding DNA adapters were added to ligate adapters followed by a thermal cycle at 30 degC and stop ligase at time of completion. The ligated fragments were then cleaned with 2 rounds of differing amounts of sample purification beads, magnetic stand and 80% ethanol. Quality of libraries were analyzed using the Bioanalyzer and quantified. The 24 samples were pooled in groups of 8 to achieve genomic coverage of 5-10x.

Appendix B. Perl Script that Formatted Final CNVR Tables

The following Perl Script formatted the final CNVR table.

```
#!/usr/bin/perl
use strict;
my $infile = '</home/brittney.keel/rebecca/cnvs.txt';
my $INPUT;
open($INPUT, $infile) or die "can't open $infile";
my $outfile= '>/home/brittney.keel/rebecca/cnvs_all.txt';
my $OUTPUT;
open($OUTPUT, $outfile) or die "can't open $outfile";
my $firstline = readline($INPUT); # remove header line
my $count = 1; #to get CNV number
while (my $line = readline($INPUT)) {
    chomp($line); #cuts out the new line
    my @parts = split /\s+/, $line; #splitting line on white space
    my $start = $parts[1]; #just grabbing postion 1 (starts at 0)
    my $chrom = $parts[0];
    my $end = $parts[2];
    my $sample = $parts[5];
    $sample=~ s/.bam//;

    my $CN = $parts[8];
    $CN=~ s/CN//; #get rid of CN

    my $type;
    if ($CN >2){
        $type= 'Amp';
    } else {
        $type= 'Del';
    }
    print $OUTPUT "$count\t$chrom\t$start\t$end\t$sample\t$type\t$CN\n"; # tab over is t

    $count= 1 + $count; #go to next CNV number
}
close ($INPUT);
close ($OUTPUT);
```

Appendix C. Perl Script that Identified Gene Overlaps

The following Perl Script identified genes that overlapped with copy number variation regions. The formatted CNVR file was ran against the Ensembl gene bank for *Sus scrofa*.

```
#!/usr/bin/perl
use strict;
use List::MoreUtils qw(uniq);
# Read in the gene file and store it.
my $infile = 'C:\Users\rebecca.anderson2\Documents\Data\genes.txt';
my $INPUT;
open ($INPUT, $infile) or die "can't open $infile";
my $firstline = readline($INPUT);
my %genes;
my @genenames;
while (my $line = readline($INPUT)) {
    chomp($line);
    my @parts = split /\s+/, $line;
    my $gene = $parts[1];
    my $chromID = $parts[2];
    my $start = $parts[3];
    my $end = $parts[4];
    my $genebiotype = $parts[5];
    $genes{$gene}{'chrom'} = $chromID;
    $genes{$gene}{'start'} = $start;
    $genes{$gene}{'end'} = $end;
    $genes{$gene}{'type'} = $genebiotype;
    push(@genenames, $gene);
}
close($INPUT);
# Read in the CNV file and check for gene overlaps.
my $infile2 = 'C:\Users\rebecca.anderson2\Documents\Data\filtered_cnvr.txt';
my $INPUT2;
open ($INPUT2, $infile2) or die "can't open $infile2";
my $firstline = readline($INPUT2);
my @overlapped;
while (my $line = readline($INPUT2)) {
    chomp($line);
    my @parts = split /\s+/, $line;
    my $cnv_chrom = $parts[1];
    my $cnv_start = $parts[2];
    my $cnv_end = $parts[3];
    foreach my $g_id (@genenames) {
        my $gene_chrom = $genes{$g_id}{'chrom'};
        my $gene_start = $genes{$g_id}{'start'};
        my $gene_end = $genes{$g_id}{'end'};
```

```
        if ($cnv_chrom eq $gene_chrom) {
            if (($cnv_start < $gene_end) && ($gene_start < $cnv_end)) {
                push(@overlapped, $g_id);
            }
        }
    }
}
close($INPUT2);
my @uniquegenes = uniq @overlapped;
my $outfile= '>C:\Users\rebecca.anderson2\Documents\Data\overlapgenes.txt';
my $OUTPUT;
open ($OUTPUT, $outfile) or die "can't open $outfile";
foreach my $g (@uniquegenes){
    print $OUTPUT "$g\n";
}
close($OUTPUT);
```