EXPANDING THE FRONTIERS OF TRANSCRIPTOME SEQUENCING DATA (RNA-SEQ).

SELECTION SIGNATURES IN CHICKENS

by

Modupeore O. Adetunji

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

Spring 2019

EXPANDING THE FRONTIERS OF TRANSCRIPTOME SEQUENCING DATA (RNA-SEQ).

SELECTION SIGNATURES IN CHICKENS

by

Modupeore O. Adetunji

Approved: _____
Limin Kung, Jr., Ph.D.
Chair of the Department of Animal and Food Sciences

Approved: _____
Mark W. Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Carl J. Schmidt, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Shawn W. Polson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Behnam Abasht, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Erin L. Crowgey, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation advisor, Dr. Carl J. Schmidt, for the opportunity to pursue research in his lab. I am grateful for all his support, guidance and encouragement throughout my graduate degree. The experiences gained working with him has definitely helped me become a better scientist. My deepest thanks to my dissertation committee members, Dr. Shawn W. Polson, Dr. Behnam Abasht and Dr. Erin L. Crowgey for their feedback and guidance, which greatly helped shape my research direction.

I thank the Schmidt lab group, for all the interestingly eccentric discussions and providing an enjoyable environment to learn. I greatly appreciate Heidi Van Every for all her assistance and help in the presentation of this dissertation. My sincerest thanks to Dr. Susan Lamont who also provided valuable feedback towards the projects tackled in this dissertation.

This dissertation is dedicated to my parents, Hon. Arch. & Mrs Adetunji for their love, encouragement and guidance in all things me, and to my vivacious sisters, Damilola and Doyin Adetunji for being my support system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Transcriptome sequencing (RNA-seq) analysis is a highly exploited technique for defining transcript abundance and differential expression analysis but is underutilized for nucleotide variant detection. Given the ability of RNA-seq to reveal active regions of the genome, detection of RNA-seq SNPs can prove valuable in understanding the phenotypic diversity between populations. This dissertation showcases the applicability of RNA-seq data in currently unexplored but important areas of biological research; such as variant analysis and detection of selection signatures in commercial broilers. I have developed a novel computational workflow that takes advantage of multiple RNA-seq splice aware aligners to call SNPs using RNA-seq data only. Our workflow achieved >97% precision and >99% sensitivity when applied to RNA-seq data and the matching whole genome sequencing (WGS) data from the Fayoumi line. Furthermore, our results discovered SNPs resulting from post-translational events that would have been missed in WGS data. The results demonstrate SNP identification from RNA-seq data be reliable and a potential resource in determining selection signatures from variant data. The identification of regions that have undergone selection is important in understanding the variation patterns responsible for the underlying phenotypic changes between populations. Modern broilers are characterized from decades of extensive genetic selection for traits of economic importance. However, improvement in economic traits also resulted in negative complications, such as skeletal abnormalities, inability to adapt to heat stress and susceptibility to diseases. These dramatic phenotypic changes imply strong

positive selection for the causal loci or polymorphisms controlling these traits. To offer insight into the variation patterns responsible for the underlying phenotypic changes, we investigated regions of selection using the SNPs derived from our RNA-seq workflow in commercial broilers.

Given the vast amounts of data generated by next-generation sequencing (NGS) data for the today's -omics era, the ability to efficiently manage the massive throughput from NGS analysis becomes a major challenge, especially when dealing with data that range on a terabyte to petabyte scale. Thus, innovative storage solutions that address this computational bottleneck are paramount. To this aim, we designed a hybrid (Relational & NoSQL) database framework, called TransAtlasDB, that addresses the crucial need for a smart and innovative storage solution for archival, management and retrieval of large-scale transcriptome analysis data output relevant to basic, medical and agriculture research.

Chapter 1

INTRODUCTION

1.1 Transcriptome Sequencing (RNA-seq)

Over the last decade, the application of massively rapid parallel sequencing technologies is proved to be the best approach for transcriptome studies. Asides for the decreasing cost of sequencing and ease in generating high-throughput sequencing data [1], transcriptome sequencing (RNA-seq) is free from many of the limitations of previous methods, like quantitative PCR and microarrays [2]. RNA-seq is a powerful tool for quantitative research in understanding biological systems [3]. The application of RNA-seq is proven to be useful in validating known systems and uncovering novel networks in response to differing treatment or environmental conditions [4].

Depending on the goal and biological questions, different strategies can be implemented for RNA-seq analysis; a typical data analysis includes quality control, read preprocessing, alignment to a reference or de novo assembly and downstream analysis such as transcripts annotation, differential gene expression (DGE), gene fusion analysis and alternative splicing [5,6]. So far, RNA-seq is a well-known technique for gene enrichment and differential expression analysis [7], but there has been little exploration in other areas such as variant detection, and allele specific expression [8]. This may be because of the complexity of the transcriptome or high false positive rates due to RNA splicing, RNA editing or sequencing errors [9]. However, recent studies have proven genomic variants can be accurately detected from RNA-seq providing adequate coverage and use of current state-of-the-art-

assemblers [10–12]. Thus, detecting variants from RNA-seq will be beneficial in providing an efficient option to validate whole genome sequencing (WGS) or whole exome sequencing (WES) variants; in identifying potentially disease-associated variations that would not have otherwise been found in genomic data; and in offering an additional deliverable of the existing RNA-seq data [11,12].

1.2 Big Data Storage

A major challenge in transcriptome studies is that the scale of data, which includes DNA sequences, transcript quantification, and polymorphism analysis generates data on a terabyte to petabyte scale. This is clearly a problem of dealing with "big data". Storage options available involve online repositories either entailing a one-line summary of published projects or archives of actual files; with the former being almost useless, the latter requires sophisticated, expensive storage requirements. Furthermore, assessing the data in the near future will be tedious, leading to either reprocessing the data files or replicating the sample study, wasting time and effort. Hence, an innovative storage solution that addresses these limitations, such as hard disk storage requirements, efficiency and reproducibility are paramount.

Databases are an excellent platform for warehousing such large amounts of interrelated data. The most commonly used database management systems are the relational database management systems (RDBMS), such as MySQL, Oracle, MariaDB and PostgreSQL. RDBMS provide a systematic storage of data by maintaining the relationship between the data members. They prevent data redundancy, enforce data consistency and concurrency, maintain data integrity, provide data security, and data sharing among numerous applications and users. Their

ease of application development and simplicity makes them the most popular database system. However, RDBMS performance significantly declines with increase in scalability and user requirements, thus a new class of database known as NoSQL [13] has arisen. NoSQL describes a broad class of technologies that provide an alternative data storage compared with RDBMS. They follow the non-relational database model and are designed to handle terabytes of unstructured data seamlessly [14]. NoSQL systems do not use the RDBMS principles; they are schema-less and do not store data in tables, instead they assign identification keys to data [15]. The types of NoSQL systems include key-value stores (e.g. HBase), document stores (e.g. MongoDB), MapReduce framework (Hadoop), and graph databases (e.g. Neo4J) [16].

The scalability and flexibility of NoSQL technologies to store and manage massive volumes of data is beneficial for handling the large complex data from next generation sequencing analysis over the traditional relational databases. However, NoSQL systems have their drawbacks; they lack data security, they do not provide ACID (Atomicity, Consistency, Isolation and Durability) transactional properties data consistency and there is no standardized method of performing transactions across the different NoSQL databases. All of these drawbacks make them difficult to adopt [16–18].

While NoSQL technologies are better suited to handle such Big Data, migrating data to the schema-less world of NoSQL will be difficult for developers to accomplish. Hence, software infrastructures have been created to coerce a structured query (SQL) environment to interact with a NoSQL database, like DQE – Distributed Query Engine and SOS – Save Our Systems [17,19]. These programming environments provide a means to combine the structure and uniformity of RDBMS

with the scalability and schema flexibility of a NoSQL database. This strategy provides a hybrid architecture that contain both types of technologies (MySQL and NoSQL) allowing efficient performance in managing huge amounts of data without compromising on either database types limitations.

1.3 Poultry Production – Chicken as a Model Organism

The poultry industry is a major source of income and important contributor of meat and egg production for human consumption [20]. Since poultry is an essential resource to the economy, the poultry industry has been undergoing huge growth in order to meet the challenges imposed by the growing world population. Consequently chickens have become the most widespread livestock irrespective of culture and religion [21,22]. Chickens are also extensively used as a model organism for research in agriculture, phylogenetics, developmental biology, virology, human diseases and many more, thus, the benefits of research in chickens can help in improving chicken production, gain knowledge in biology and shed light on diseases and human medicine [23,24].

The main ancestor of today's chickens is the red junglefowl (*Gallus gallus*), and minor introgression with other junglefowl species like the grey junglefowl (*Gallus sonneratii*) or the green junglefowl (*Gallus varius*), which resulted from many generations of controlled genetic and phenotypic changes [25–27]. In the process of domestication, natural and artificial selection has led to a wide spectrum of phenotypically diverse chicken breeds, and this is largely due to selection for commercial traits; such as growth, production and reproduction [22].

Decades of extensive genetic selection have improved the traits of economic importance in commercial birds, such as growth rate, feed efficiency and body composition. For instance, comparison of the modern broilers with the heritage/legacy broilers (i.e. broilers that had not been subjected to selection since 1957) found that the average body weight of modern broilers at 42 days of age increased by over 400%, with concurrent 50% improvement in feed conversion ratio compared to the heritage broilers [28,29]. However, improvement in economic traits also resulted in negative complications, such as skeletal abnormalities, abdominal fatness , reduced reproductive performance, inability to adapt to environmental changes (i.e. heat stress) and increased susceptibility to diseases (e.g. Wooden Breast) in commercial broilers [30–32]. These dramatic phenotypic changes imply strong positive selection for the causal loci or polymorphisms controlling these traits, generating selection signatures. Investigating these signatures of selection can aid in identifying variation patterns responsible for the underlying phenotypic changes and to better understand the biological mechanism controlling these traits [33,34].

Publication of the draft chicken genome sequence was a significant achievement for biologists for genomics research [35], and ongoing improvements to the chicken reference genome, *Gallus gallus*, has greatly enhanced the insights in avian genomics [36]. Furthermore, the availability of the reference genome and transcriptome sequences provides the avenue for genome-wide studies in chickens, including: comparative genomics, comparative transcriptomics, functional genomics and genome-wide association studies. Genetic factors controlling growth, development, reproduction and production have been extensively studied [37,38]. The studies show that the different bird breeds differ in many morphological features and

phenotypes of commercial relevance, they also differ in adaptations to environmental pressures. For example, the Fayoumi breed, indigenous to Egypt, is prized for its robustness in harsh environment and disease resistance compared to other breeds [39–41]. Therefore, identifying the genetic variants between breeds that contribute to these differences can provide insights into understanding the biological mechanisms that defines phenotypic diversity [42]. With the availability of next generation sequencing data and computational tools, it is possible to screen for candidate genes affected by selection signatures in the whole genome.

Using whole-genome resequencing data, several candidate sweeps have been confirmed using Z-transformed pooled heterozygosity (ZHp) scores [43]. The statistic estimates the reduced heterozygosity in chromosomal regions affected by selection and is used for detecting loci that are at or near fixation. Most of the candidate genes were identified in regions associated growth, appetite and metabolic regulation, like the thyroid stimulating growth hormone (TSHR) [42], Beta-carotene oxygenase 2 (BCO2/BCDO2) [44,45], and many others. Investigating the selection footprints in the genome allows for better understanding of the evolutionary pressures during domestication.

1.4 Thesis Aims

This thesis has three main objectives. First, I aim to expand the current applications of transcriptomics, proffering evidences to prove the beneficial uses of RNA-seq in other unexplored areas of bioinformatic analysis such as in variant detection and annotation analysis. Secondly, I aim to contribute to the understanding of genetic diversity influenced by domestication and selection that characterize the

commercial broilers, Ross708 and Heritage lines using selection sweeps analysis. Thirdly, given the lack of suitable storage solutions for complex analysis output typically generated from RNA-seq analysis, I aim to provide a database storage system for smart storage for the complex analysis data generated from RNA-seq analysis.

1.4.1 Thesis Overview

This thesis is explained in the following chapters:

Chapter 2 evaluates the identification of variants in RNA-seq data. Our variant analysis pipeline (VAP) utilizes stringent mapping and variant calling methodologies to ensure accurate detection of RNA-seq SNPs. VAP includes (1) a mapping procedure, which consists of the application of three splice aware tools for accurate mapping of RNA-seq reads to the reference genome, (2) the variant calling procedure using the Genome Analysis Toolkit (GATK) and, (3) the quality control and filtering procedures at each stage of the pipeline. Our computational pipeline assessment of sensitivity and specificity highlights the ability and importance of SNP discovery from RNA-seq data.

Chapter 3 examines the signatures of selection in commercial broilers, both the modern and legacy broiler lines. Identifying the potential selection signatures can help identify causal polymorphisms controlling traits. Thus, the candidate genes identified may be of great interest for future research into the genetic architecture of traits relevant to modern broiler breeding.

Chapter 4 presents the database storage solution, called TransAtlasDB. TransAtlasDB is a hybrid database system for smart storage of the complex analysis output generated from the current state-of-the-art open source tools for RNA-seq samples metadata, gene expression and quantification analysis and variant analysis.

Chapter 2

VARIANT ANALYSIS PIPELINE FOR ACCURATE DETECTION OF GENOMIC VARIANTS FROM TRANSCRIPTOME SEQUENCING DATA

2.1 Abstract

The wealth of information deliverable from transcriptome sequencing (RNA-seq) is significant, however current applications for variant detection still remain a challenge due to the complexity of the transcriptome. Given the ability of RNA-seq to reveal active regions of the genome, detection of RNA-seq SNPs can prove valuable in understanding the phenotypic diversity between populations. Thus, we present a novel computational workflow named VAP (Variant Analysis Pipeline) that takes advantage of multiple RNA-seq splice aware aligners to call SNPs in non-human models using RNA-seq data only. We applied VAP to RNA-seq from a highly inbred chicken line and achieved >97% precision and >99% sensitivity when compared with the matching whole genome sequencing (WGS) data. Over 65% of WGS coding variants were identified from RNA-seq. Further, our results discovered SNPs resulting from post translational modifications, such as RNA editing, which may reveal potentially functional variation that would have otherwise been missed in genomic data. Even with the limitation in detecting variants in expressed regions only, our method proves to be a reliable alternative for SNP identification using RNA-seq data.

2.2 Introduction

Detection of single nucleotide polymorphisms (SNPs) is an important step in understanding the relationship between genotype and phenotype. The insights achieved with next generation sequencing (NGS) technologies provide an unbiased view of the entire genome, exome or transcriptome at a reasonable cost [46]. Most methods for variant identification utilize whole-genome or whole-exome sequencing data, while variant identification using RNA-seq remains a challenge because of the complexity in the transcriptome and the high false positive rates [9]. However, having access to RNA sequences at a single nucleotide resolution provides the opportunity to investigate gene or transcript differences across species at a nucleotide level.

RNA-seq is applicable to numerous research studies, such as the quantification of gene expression levels, detection of alternative splicing, allele-specific expression, gene fusions or RNA editing [4]. Workflows have been developed to address identifying SNPs from RNA-seq reads in human, including SNPiR and eSNV-detect. SNPiR [11] employs BWA aligner and variant calling using GATK UnifiedGenotyper, eSNV-detect [47] relies on combination of two aligners (BWA and TopHat2) followed by variant calling with SAMtools and Opposum + Platypus [48]. Opposum reconstructs RNA alignment files to make them suitable for haplotype-based variant calling with Platypus [49]. These workflows require adequate sampling of RNA-seq reads and accurate mapping of the RNA-seq reads to the reference genome to avoid false positive SNP calls. In addition to the limitation of these workflows being specifically designed for human samples, they either rely on outdated variant calling procedures, or preprocessing RNA-seq data to make it suitable for variant calling, thus making it difficult to sufficiently compare their performance.

Due to the aforementioned limitations, we designed a workflow, called VAP (Variant Analysis Pipeline), to reliably identify SNPs in RNA-seq in non-human models. VAP takes into consideration current state-of-the-art RNA-seq mapping, variant calling algorithms and the GATK best practices recommended by the Broad Institute [50], Our workflow consists of (i) multiple splice-aware reference-mapping algorithms that make use of the transcripts annotation data, (ii) variant calling following the Genome Analysis Toolkit (GATK) best practices, and (iii) stringent filtering procedures. We propose that calculating specificity will estimate the likelihood of detecting a true variant in RNA-seq and sensitivity will determine how likely RNA-seq is able to detect an expressed SNP if it is present in a transcribed gene [51]. Overall the results indicate that RNA-seq can be an accurate method of SNP detection using our VAP workflow.

2.3 Materials and Method

2.3.1 VAP Workflow

Figure 2.1 shows the flowchart of the VAP workflow. Read quality was assessed using FastQC and preprocessed using Trimmomatic [52] and/or AfterQC [53] when required. Pre-processed RNA-seq reads were mapped to the reference genome and known transcripts employing three splice-aware assembly tools; TopHat2 [54], HiSAT2 [55] and STAR [56]. All three programs are open-source and are highly recommended for reliable reference mapping of RNA-seq data [57]. SAMtools was used to convert the alignment results to BAM format [58]. The mapped reads undergo sorting, adding read groups, and marking of duplicates using Picard tools package

(https://broadinstitute.github.io/picard/). The SNP calling step uses the GATK toolkit for splitting "N" cigar reads (i.e. splice junction reads), base quality score recalibration and variant detection using the GATK HaplotypeCaller [59]. Lastly, the filtering steps entail assigning priority to SNPs found in all three mapping plus SNP calling steps, to minimize false positive variant calls. The priority SNPs were filtered using the GATK Variant Filtration tool and custom Perl scripts. SNPs were filtered using the set of read characteristics summarized in Table 2.1; low quality calls (QD < 5), or variants with strong strand bias (FS > 60), or low read depth (DP < 10) and SNP clusters (3 SNPs in 35bp window) were excluded from further analysis. Custom filtering was described as follows: nucleotide positions with less than 5 alternative allele supporting reads and nucleotide positions with heterozygosity scores < 0.10 are eliminated to prevent ambiguous SNP calls. Heterozygosity score (Het) is calculated by $Het_i = \frac{aa_i}{t_i}$; where i is the nucleotide base pair, aa_i is the alternate read depth at the location i and t_i is the total number of reads at location i. After filtering, the variants were annotated using the ANNOVAR [60] and VEP [61] software.

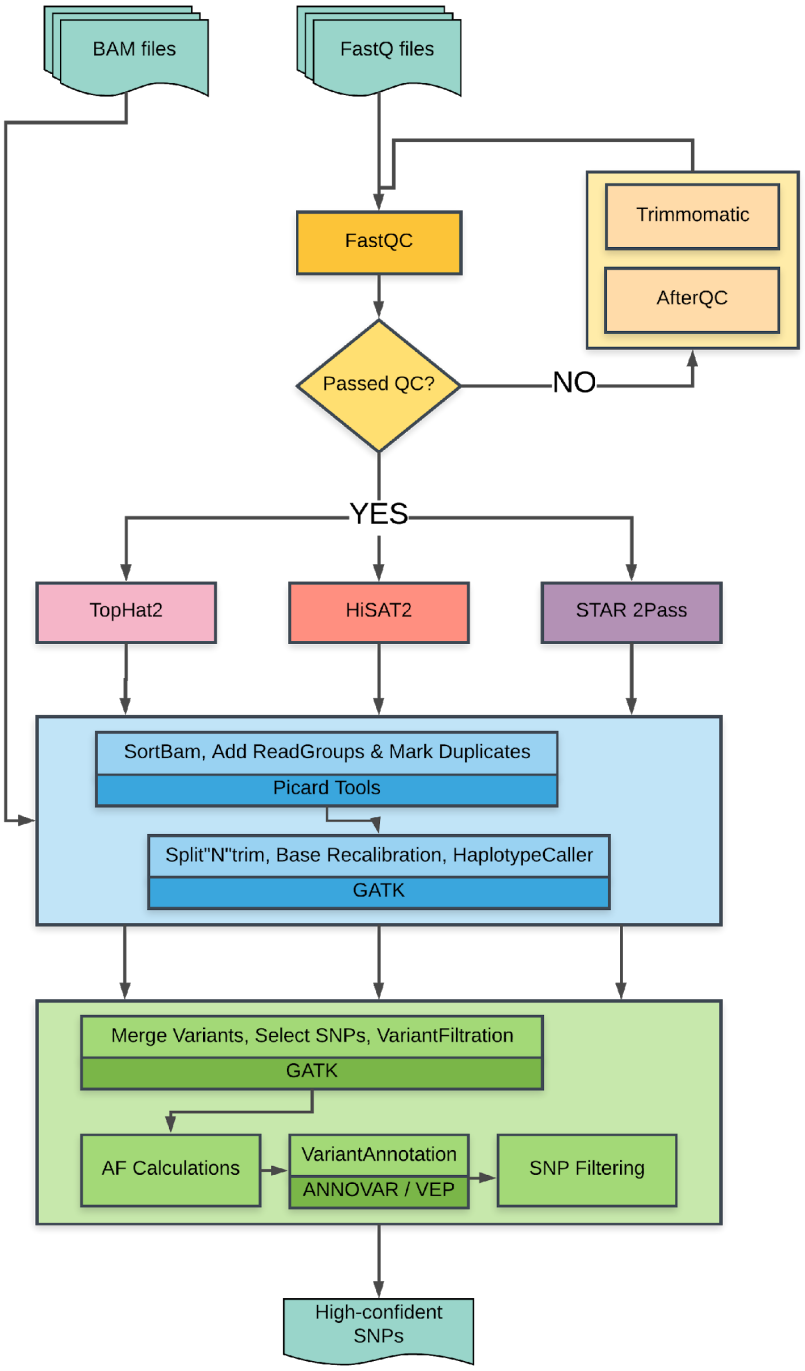


Figure 2.1 Flow chart of the VAP workflow. FastQ files are QC using FastQC, mapped using three aligners. BAM files are pre-processed by Picard and GATK, then merged, annotated and filtered to achieve high-confident SNPs.

Table 2.1 Criteria used in the VAP filtering workflow.

Criteria	**Threshold**
GATK - VariantFiltration tool	
ReadRankPosSum (RRPS)	RRPS < -8
Quality by depth (QD)	QD < 5
Read depth (DP)	DP < 10
Fisher's exact test p-value (FS)	FS > 60
Mapping Quality (MQ)	MQ < 40
SnpCluster	3 SNPs in 35bp
Mann-Whitney Rank-Sum (MQRankSum)	MQRankSum < -12.5
Alternative allele supporting read depth	ALTreads < 5
Alternative allele frequency	$Het = \frac{aa}{t} \leq 0.10$

2.3.2 DNA and RNA Sequencing data

Raw RNA-seq and whole genome sequencing (WGS) data were obtained from previously published works. Both sequencing data sets from highly inbred Fayoumi chickens were sequenced on the Illumina HiSeq platform. For RNA-seq, pooled samples were collected from the brain and liver generating 117 million 75bp pair-end reads are available in the NCBI Sequence Read Archive with accession number SRP102082 [62]. For WGS, pooled DNA samples were constructed from individual DNA isolates from blood from 16 birds, contributing to 241 million 100bp pair-end reads [39]. The transcriptome and whole genome of these samples have been deeply sequenced to provide sufficient coverage for accurate identification of variants from RNA and DNA of the same line. Having matched RNA and DNA samples allows for suitable verification of RNA SNP calls, making our datasets good candidates for evaluating the accuracy of our VAP methodology.

2.3.3 600K Genotyping data

Pooled samples from two different projects; Feed Efficiency [63,64] and Wooden Breast Disease [65], were genotyped with the ThermoFisher Axiom Chicken Genotyping Array [66]. The raw genotyping data (cel files) was analyzed with the *Gallus gallus* 5.0 genome (from Axiom server) using the Axiom Analysis Suite Software (version 3.0.1) following the software's Best Practices Workflow using recommended settings for agricultural animals. The final results were exported, including a raw VCF of all the genotype calls and a *txt* file of all variants with >= 97% call rate. The *txt* file was utilized to filter low quality variants from the raw VCF.

2.3.4 RNA-seq Mapping, Variant Calling and Filtering

RNA-seq samples were mapped with the three RNA-seq mapping tools; TopHat2 (v 2.1.1), HiSAT2 (v 2.1.0) and STAR (v 2.5.2b) 2-pass method using default parameters to the NCBI *Gallus gallus* Build 5.0 reference genome and the mapping files were converted to BAM using SAMtools (v 1.4.1). The BAM files were processed, and variants were called using Picard tools (v 2.13.2) and GATK (v 3.8-0-ge9d806836) through the VAP pipeline. We used ANNOVAR (v 2017Jul16) and VEP (v 91) to annotate variants on the basis of gene model from RefSeq, Ensembl and the UCSC Genome Browser. We retained SNPs found with all three mapping tools and those that fulfilled the filtering criteria in Table 2.1. SNPs found in WGS data or present in dbSNP (Build 150) are identified as "verified" variants, while those not found are tagged as "novel". The precision of the VAP workflow was determined as the number of all known RNA-seq variants divided by the total number of known and novel RNA-seq variants (Equation 2.1).

$$Precision = \frac{verified_{SNPs}}{verified_{SNPs} + novel_{SNPs}}$$

Equation 2.1 Precision of VAP workflow equation.

2.3.5 WGS Mapping, Variant Calling and Filtering

We mapped the WGS data with BWA-mem (v 0.7.16a-r1181) [67] using default parameters to the NCBI *Gallus gallus* Build 5.0 reference genome. Variant calling was performed using Picard and GATK HaplotypeCaller, following the recommendations proposed by Van der Auwera et al [68] and Yiyuan Yan et al [69]. Similar filtering parameters for RNA-seq as previously described were applied using the GATK Variant Filtration tool and custom scripts (Table 2.1). To allow a fair comparison between RNA-seq and WGS variants, we estimated specificity with the fraction of coding exonic variants identified from WGS.

2.3.6 To Calculate Sensitivity and Specificity of Verified RNA-seq SNPs

To determine the accuracy of detecting a true variant from RNA-seq using our VAP workflow, we calculated the specificity and sensitivity of the verified RNA-seq SNPs. Because we are using transcriptome data, we should only be theoretically able to detect SNPs at sites expressed in our data. Sensitivity analysis will evaluate the accuracy of our pipeline to correctly detect known SNPs using RNA-seq, and specificity analysis will assess how likely a SNP is detected by RNA-seq compared to WGS. To do this, we further characterized our verified RNA-seq SNPs as "true-verified" and "non-verified" SNPs. A true-verified SNP (TS) is a SNP with the same corresponding dbSNP and/or WGS data, and a non-verified SNP (NS) is where the genotype does not match the dbSNP/WGS data. Also, SNPs not detected in RNA-seq but found in WGS and validated using dbSNP are called "DNA-verified" SNPs (DS).

Sensitivity is calculated as the number of TS divided by the number of TS plus the number of PS (Equation 2.2). While specificity is estimated as the number of TS divided by the number of TS plus the number of DS (Equation 2.3) [11,51].

$$Sensitivity = \frac{TS}{TS + NS}$$

Equation 2.2 Sensitivity equation.

$$Specificity = \frac{TS}{TS + DS}$$

Equation 2.3 Specificity equation.

2.3.7 Gene Expression Analysis

Variants in expressed regions were identified by gene quantification analysis using StringTie v1.3.3 [70] on the TopHat2, HISAT2 and STAR BAM files. The average FPKM (fragments per kilobase of transcript per million fragments mapped) was calculated for specificity analysis.

2.4 Results

2.4.1 The Multi-Aligner Concept

VAP uses a multi-aligner concept to call SNPs confidently. The application of multiple aligners reduces false discovery rates significantly, as shown in the eSNV-detect pipeline [47,71]. However, we do not assign a confidence hierarchy on candidate SNP calls, rather SNP detected from all three aligners are weighted equally, thus all consensus SNPs are obtained and filtered based on the filtering criteria listed

above. High percentages of similar SNPs were observed between all three tools, which shows that using a splice-aware read mapper is appropriate for reference mapping using RNA-seq, unlike with BWA. Table 2.2 provides the summary of mapping and variant calling statistics from the multiple aligners.

Table 2.2 Summary from the multiple aligners; read mapping statistics and variant calls.

Tools	% Reads mapped	% Reference covered	Variants	SNPs	% similar SNPs
TopHat	87.70	23.07	578655	535505	96.12
HiSAT	90.53	23.44	636948	583547	88.21
STAR	87.81	23.70	798696	708391	72.66

2.4.2 SNPs detected in RNA-seq data.

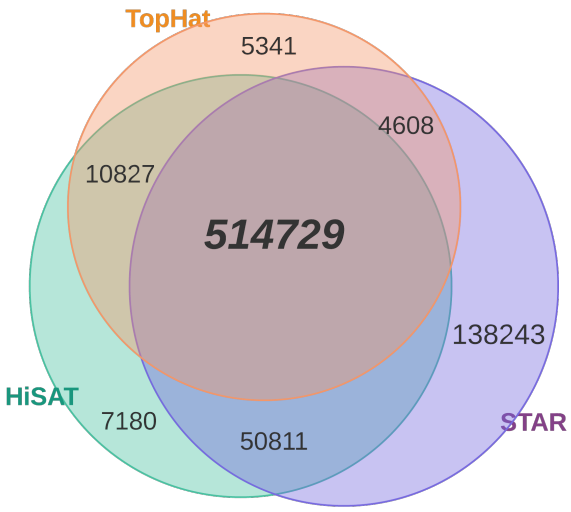


Figure 2.2 Comparison of RNA-seq SNPs identified in the different mapping tools.

Our method identified 514,729 SNPs from all 3 aligners before filtering, which assures reduction of false positives calls (Figure 2.2). After filtering, 282,798 (54.9%) high confidence SNPs remain, of which 97.2% (274,777 SNPs) were supported by evidence from WGS or dbSNP v.150 (Figure 2.3). The verified sites exhibited a transition-to-transversion (ts/tv) ratio of 2.84 and estimated ts/tv ratio of ~5 for exonic regions and thus a good indicator of genomic conservation in transcribed regions. For the remaining (novel) 8,021 SNPs, we observed slightly lower ts/tv ratio (2.81) than for the verified sites. The variant sites showed a clear enrichment of transitions, inclusive of A>G and T>C mutations (73.9%), indicative of mRNA editing and the dominant A-to-I RNA editing [72] (Figure 2.4).

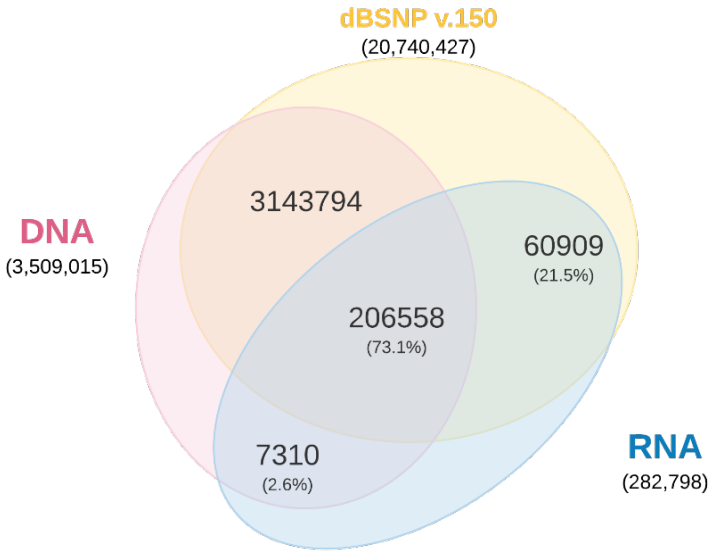


Figure 2.3 Comparison of the number of RNA-seq SNPs and the percentage found in either dbSNP or WGS.

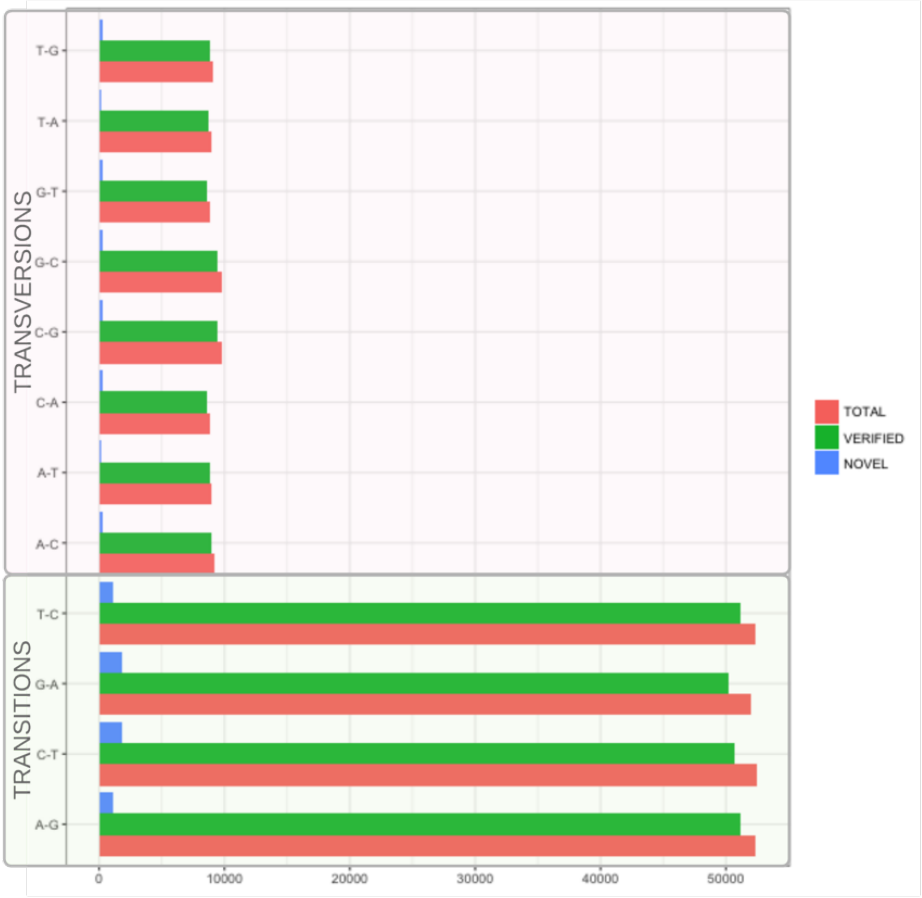


Figure 2.4 The mutational profile of RNA-seq variants.

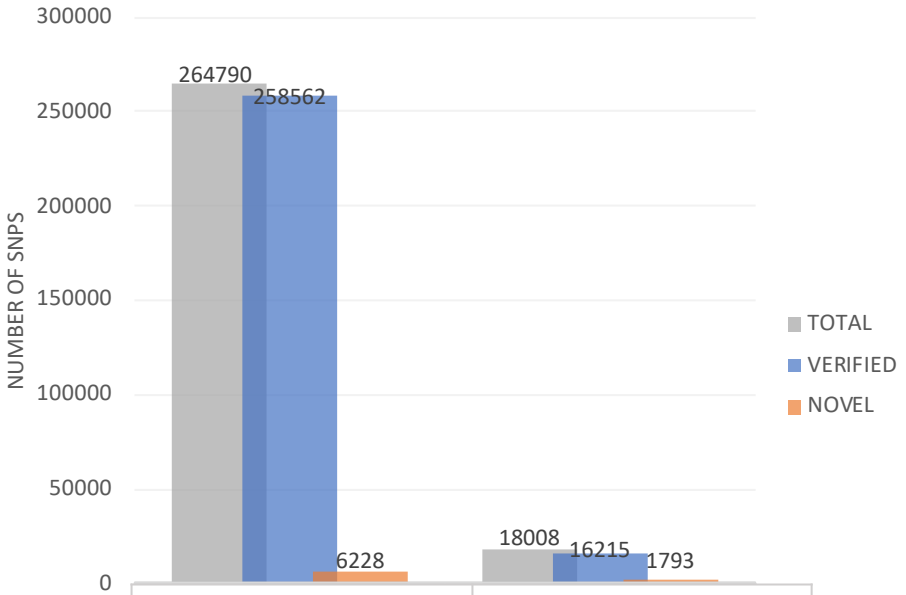
2.4.3 SNPs Allele Frequencies

The 282,798 SNPs called, were grouped based on their variant allele frequencies (VAF). VAFs were calculated by dividing the number of reads supporting the variant allele by the total number of reads obtained. SNPs were grouped as homozygous alternate with VAF ≥ 0.99, and heterozygous with VAF < 0.99. We found 264,790 (93.6%) and 18,008 (6.4%) SNPs were classified as homozygous alternate and heterozygous, respectively. Not surprisingly, most of the predicted SNPs were homozygous to the non-reference allele, suggesting genetic diversity of the

Fayoumi breed compared to the reference genome *Gallus gallus* (Red Jungle Fowl) is influenced by polymorphisms [32,73]. This will aid in identifying the variations enriched by selection.

2.4.4 Precision and Sensitivity of RNA-seq SNPs

A high proportion of SNPs detected in RNA-seq data are true variants. The sensitivity of SNP calls are similar for both heterozygous and homozygous sites (Figure 2.5). With the high number of calls verified via dbSNP, the precision is much higher for homozygous variants compared to heterozygous variants, indicating that a high proportion of expected variants can be detected using RNA-seq with adequate coverage. The decreased precision in heterozygous SNPs may suggest expression of the non-reference allele, and this provides the opportunity to study the effects of genetic variation on the different transcriptional events, such as RNA editing, alternate splicing and allelic specific expression, which cannot be explained using DNA sequencing data [8].



	HOMOZYGOUS	HETEROZYGOUS
SENSITIVITY	99.8%	99.7%
PRECISION	97.6%	90.1%
TS/TV	2.83	2.94

Figure 2.5 Comparison of SNPs identified as homozygous and heterozygous in RNA-seq.

2.4.5 Functional classification of variants identified from RNA-seq and WGS

Thirteen percent of the RNA-seq SNPs were predicted to be within protein-coding regions while >1% of the WGS SNPs were in coding regions when annotated against both the NCBI and ENSEMBL gene database for chicken; the remaining SNPs were found in non-coding or regulatory regions (Table 2.3). Due to difficulty in annotating and determining the impact of polymorphisms on non-coding or regulatory regions, only polymorphisms found on coding regions were further evaluated.

Table 2.3 SNPs belonging to different annotation categories.

	Annotation categories	Number (%)	Mean VAF (± SD)	No. homozygous (VAF ≥ 0.99) [a]
RNA	Intergenic	162240 (57)	0.99 (0.06)	152732 (94%)
	Up/downstream	11793 (4)	0.99 (0.07)	10817 (92%)
	Intronic	58028 (20)	0.99 (0.05)	55744 (96%)
	Exonic	36702 (13)	0.99 (0.08)	33051 (90%)
	Non-synonymous	8599 (3)	0.98 (0.11)	7664 (89%)
	Synonymous	28094 (10)	0.99 (0.07)	25353 (90%)
	Stop-gain/loss	39 (>1)	0.96 (0.16)	34 (87%)
	Splicing	8 (>1)	1 (0)	8 (100%)
	UTR3/UTR5	13421 (5)	0.98 (0.09)	11895 (88%)
	ncRNA	106 (>1)	0.97 (0.13)	100 (94%)
WGS	Intergenic	2865498 (82)	0.99 (0.07)	2659382 (92%)
	Up/downstream	30741 (>1)	0.99 (0.08)	28558 (93%)
	Intronic	565323 (16)	0.99 (0.07)	522577 (92%)
	Exonic	34294 (1)	0.98 (0.09)	31875 (92%)
	Non-synonymous	8946 (>1)	0.97 (0.11)	8283 (86%)
	Synonymous	25274 (>1)	0.99 (0.08)	23526 (93%)
	Stop-gain/loss	74 (>1)	0.98 (0.11)	66 (69%)
	Splicing	17 (>1)	0.97 (0.13)	17 (100%)
	UTR3/UTR5	12476 (>1)	0.99 (0.07)	11515 (92%)
	ncRNA	302 (>1)	0.99 (0.07)	277 (91%)
Overlap RNA & WGS [b]	Intergenic	125218 (58)	1 (0.04)	112462 (89%)
	Up/downstream	9787 (4)	0.99 (0.04)	6908 (87%)
	Intronic	47894 (22)	1 (0.04)	43636 (91%)
	Exonic	22551 (10)	0.99 (0.05)	19533 (87%)
	Non-synonymous	5165 (2)	0.99 (0.06)	4486 (87%)
	Synonymous	17363 (8)	0.99 (0.05)	15030 (86%)
	Stop-gain/loss	23 (>1)	1 (0.01)	17 (39%)
	Splicing	5 (>1)	1 (0)	5 (100%)
	UTR3/UTR5	9943 (5)	0.99 (0.04)	8475 (85%)
	ncRNA	73 (>1)	0.99 (0.03)	63 (86%)

[a] The percentages are in relation to the number of SNPs within the annotation category.
[b] The percentages are in relation to the number of SNPs within the annotation category in RNA.

2.4.6 Specificity of RNA-seq SNPs

To calculate specificity of our VAP methodology, we focused on variants in coding regions to allow for fair comparison between RNA-seq and WGS data. Approximately 66% of the coding variants identified by WGS were discovered using RNA-seq alone (Figure 2.6). Given that RNA-seq required less sequencing effort and computational requirements (e.g. 234 million for RNA-seq compared to the 482 million for WGS sequencing reads used in our case study). Using RNA-seq data is advantageous because it enriches for expressed genic regions compared to WGS and therefore will increase the power to detect functionally important SNPs impacting protein sequence.

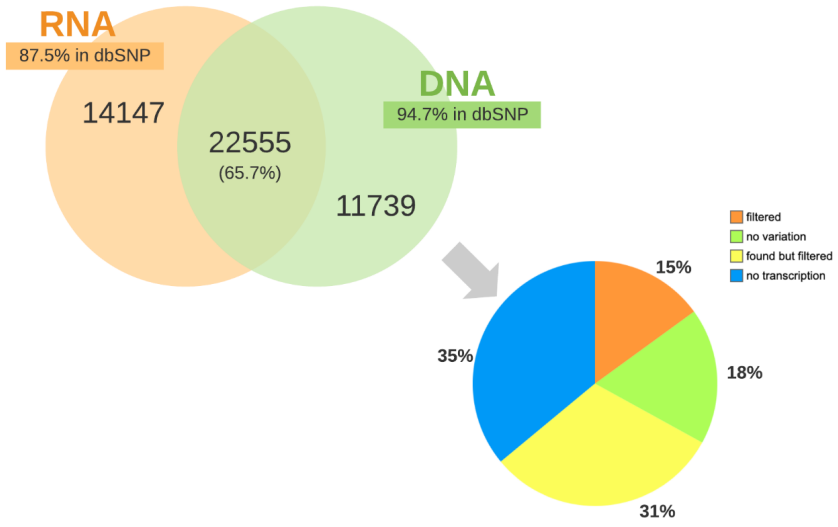


Figure 2.6 Overlap of SNPs found in coding regions from RNA-seq and WGS. 66% of the coding variants identified in WGS data were found in RNA-seq. However, the remaining WGS coding variants were not detected as a result of either: lack of expression/transcription ("no transcription"), the position was homozygous in RNA ("no variation"), "found but filtered" signifying that the position was detected but removed by one of our filtering steps, or "filtered" which indicates the position was heterozygous but filtered because it didn't meet the default parameters for variant detection.

We then compared the RNA-seq SNPs in expressed genes (having FPKM > 0.1), and the specificity increased from 66% to over 82% (Figure 2.7). This shows that a large fraction of genes are expressed at very low levels (Figure 2.8). Overall the results prove our methodology can achieve high specificity for variant calling in expressed regions of the genome.

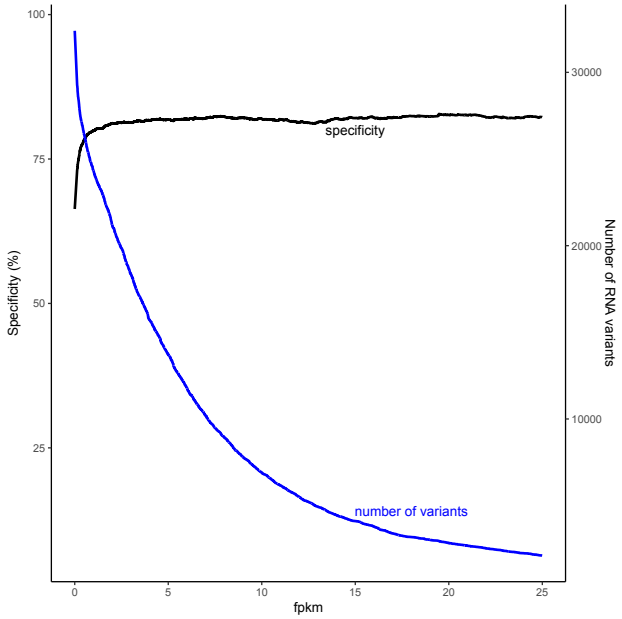


Figure 2.7 Specificity and the number of RNA-seq SNPs detected in relation to the genes expressed (FPKM values).

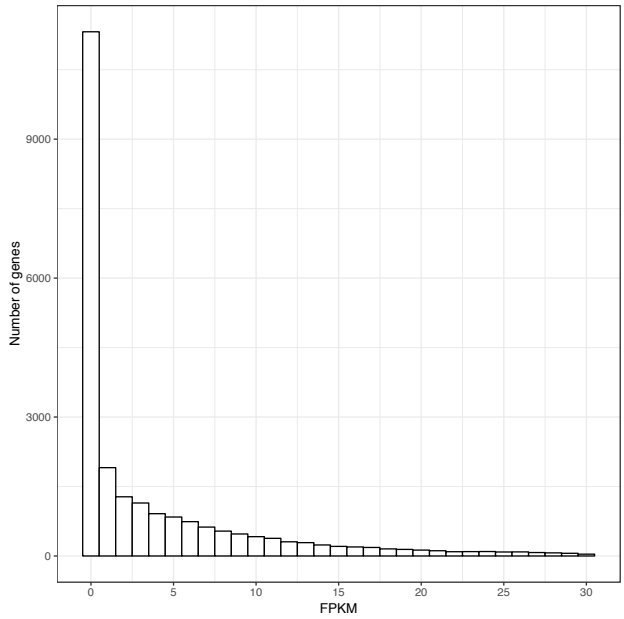


Figure 2.8 Distribution of expression levels for genes with RNA-seq SNPs.

2.4.7 Comparison of RNA-seq SNPs and 600k Genotyping Panel SNPs

Given the high accuracy of genotyping arrays for SNP discovery, we compared our initially verified RNA-seq SNPs with the 600k chicken genotyping panel. A low percentage (10%) of our RNA-seq SNPs overlap with the 600k SNPs (Figure 2.9), which is largely due to the limitation in the number of variants the genotyping panel is able to capture across different samples. However, 99.9% of the genotyping SNPs were found in dbSNP, proving dbSNP is an adequate method for *in silico* verification of our RNA-seq SNPs.

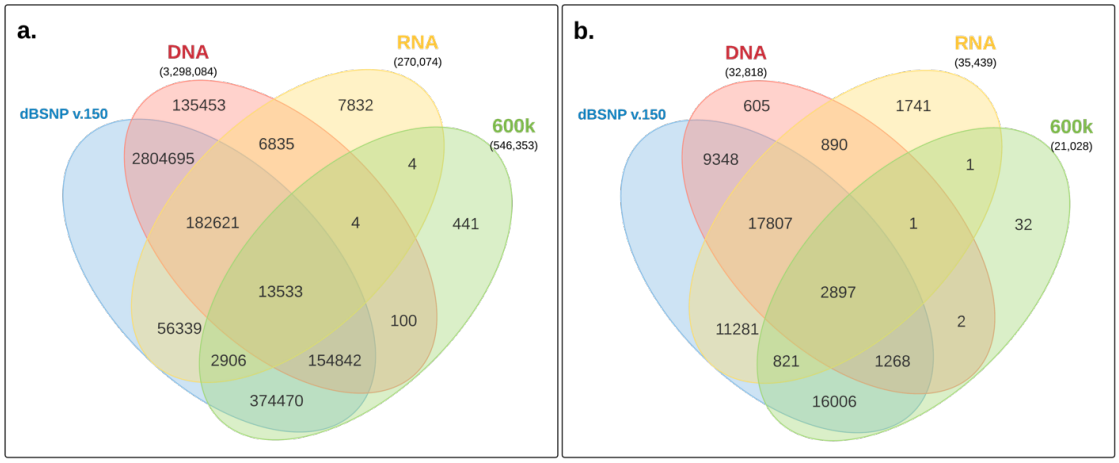


Figure 2.9 Comparison of SNP calls between 600k genotyping panel, RNA-seq SNPs, WGS SNPs and dbSNP v150, using (a) all autosomal SNPs and (b) autosomal SNPs found in exons.

2.4.8 RNA–DNA differences (RDD) sites

As mentioned before, our RNA-seq SNPs were notably contributed from transitions which may be attributed to mRNA editing. Further classifications of the RNA-seq SNPs detected in exons reveal 34% of the exonic SNPs verified by dbSNP were not identified in our WGS data. The majority of the RNA SNPs were not found in WGS because of the mapping and filtering parameters as shown in Table 2.4. Interestingly, 28% of these SNPs were not found because the alternate nucleotide was not present in the DNA sequence indicating RNA–DNA differences (RDD). Consequently, these RDD sites may result from post-transcriptional modification of the RNA sequence, such as RNA editing or alternative splicing.

Table 2.4 Explanation for the 14,147 RNA SNPs not found in WGS data.

Reason for absence	Number of SNPs
Position was heterozygous in WGS but filtered because it didn't meet the default parameters for variant detection.	1225
No reads were mapped to region/position.	1693
Position was homozygous in WGS	3471
Position was heterozygous in WGS but removed by our custom filtering criteria	7758

RNA editing is the most prevalent form of post-transcriptional maturation processes that contributes to transcriptome diversity. It involves the modification of specific nucleotides in the RNA sequence without altering its template DNA [72,74]. From our dataset, we identified the three non-synonymous RDD mutations on *CYFIP2*, *GRIA2* and *COG3* previously validated by Frésand et al. in chicken embryos[72] (Table 2.5). This demonstrates the VAP methodology ability to detect conserved RNA editing phenomena and that it can be used in further discovery of novel post-transcriptional editing events.

Table 2.5 Potentially functional RDD candidates found in Fayoumi.

Chromosome	Position	Nucleotide		Aminoacid change	Gene Short Name	VAF
		DNA	RDD			
chr 1	167798513	A	G	I/V	COG3	0.524
chr 4	21653669	A	G	R/G	GRIA2	0.703
chr 13	11398088	T	C	K/E	CYFIP2	0.375

2.5 Discussion

RNA-seq is instrumental in understanding the complexity of the transcriptome. Several methodologies have provided approaches to understanding the varied aspects occurring in the transcriptome, but little has been done in its application to identifying variants in functional regions of the genome. To this aim, we designed the VAP workflow, a multi-aligner strategy using a combination of splice-aware RNA-seq reference mapping tools, variant identification using GATK, and subsequent filtering that allows accurate identification of genomic variants from transcriptome sequencing. Our results show very high precision, sensitivity and specificity, though limited to SNPs occurring in transcribed regions.

Considering the mapping phase of RNA-seq reads is a crucial step in variant calling, we devised a reference mapping strategy using three RNA-seq splice-aware aligners to reduce the prevalence of false positives. The use of the splice-aware aligner allows for accurate assembly of reads because it makes use of both the genome and transcriptome information simultaneously for read mapping.

The ability to call variants from RNA-seq has numerous applications. It enables validation of variants detected by genome sequencing. It also uncovers potential post-transcriptional modifications for gene regulation (Table 2.5) and allows for detection of previously unidentified variants that may be functionally important but difficult to capture using DNA sequencing or exome sequencing at lower cost. For instance, 87.5% of RNA-seq variants were not found in WGS though well characterized in dbSNP (Figure 2.6). Therefore, RNA variants can be used in identifying genetic markers for genetic mapping of traits of interest, thus offering a better understanding of the relationship between genotype and phenotype.

Our VAP methodology shows high precision in calling SNPs from RNA-seq data. It is however limited by the RNA-seq experiments; RNA SNPs are detected only on the transcripts expressed. Regardless of comprehensive coverage, variant detection in some portions of the genome are not guaranteed by RNA-seq because of the potential lack of expression. Also, allele-specific gene expression or tissue-specific gene expression might hamper the discovery of genomic variants given that the allele carrying the variant might not be expressed or the tissues collected might not express the genes of interest. In addition, as a result of monoallelic expression (only one parental allele is expressed), RNA SNPs might be miscalled as homozygous rather than a heterozygous variant, attributing to the large number of homozygous SNPs identified in our case study (Figure 2.5).

SNP genotyping offers a highly accurate and alternative method of SNP discovery, and thus offers an additional *in silico* method of validation of our RNA-seq SNPs. However, a low overlap with the 600K chicken genotyping panel was observed (Figure 2.9). This low overlap is most likely due to the limitations in genotyping panels currently available for any given organism. The genotyping panels are limited by the number of variants they are able to capture across different genetic backgrounds. [66]. Not surprisingly, the majority of the 600K genotyping variants were also identified in dbSNP, proving that dbSNP an excellent choice for *in silico* validation.

Nevertheless, VAP allows the detection of variants even for lowly expressed genes. To obtain higher confidence in variant calls, pooling multiple data sets (i.e. RNA-seq from different tissues) can increase the coverage thereby facilitate variant discovery in regions of interest that would have otherwise been missed. Our study

demonstrates that variants calling from RNA-seq experiments can tremendously benefit from an increased number of reads increasing the coverage of genomic regions especially for whole genome analysis; nevertheless even our small sample size allowed for reliable calling of variants and enriching for variants in exonic regions.

Despite the limitations of calling genomic variants from RNA-seq data, our work shows high sensitivity and specificity in SNP calls from RNA-seq data. SNP calling from RNA-seq will not replace WGS or exome-sequencing (WES) approaches but rather offers a suitable alternative to either approaches and might complement or be used to validate SNPs detected from either WGS or WES. Overall, we present a valuable methodology that provides an avenue to analyze genomic SNPs from RNA-seq data alone.

Chapter 3

SIGNATURES OF SELECTION IN MODERN BROILERS USING TRANSCRIPTOME SEQUECING (RNA-SEQ) DATA

3.1 Introduction

Natural and artificial selection during chicken domestication, has led to phenotypically distinct livestock breeds. A major contributor for the extreme chicken phenotypes has been artificial selection for specific traits of commercial relevance including growth rate, egg production, body size and feed efficiency. In the United States, commercial chickens have been extensively selected into two groups: layers for egg production and broilers for meat production. Modern broilers exhibit enhanced growth, especially in the skeletal muscle compared with legacy broilers (i.e. broilers that had not been subjected to selection since the 1950s) [28,29,37]. However, the selection for commercial traits has led to some unanticipated consequences such as reduced resistance to infectious disease, increased skeletal deformities and increased mortality [75,76]. These undesirable effects may result from negative pleiotropic effects of the alleles under selection [77], alteration of causal polymorphisms due to selection for performance to these traits [78,79], or tight linkage of deleterious alleles to alleles under selection. Identifying these selection signatures can help researchers better understand the biological mechanisms controlling these traits.

Various statistical methods have been applied to detect selection signatures at the genomic level using high-throughput sequencing data or high density SNP chips in domestic animals. Methods include Z-transformed pooled heterozygosity scores (ZH_p)

[42,80], iHS (integrated haploytype homozygosity) [81], Wright fixation index (F_{ST}) , extended haplotype homozygosity (EHH), CLR (composite likelihood ratio) [82], and others all designed to assess evidence of selection from individual candidate variants [83]. Given the benefits of using transcriptome sequencing (RNA-seq) in both quantitative and exploratory studies as described in previous chapters, RNA-seq provides the means of identifying SNPs in transcribed regions of the genome. We applied the ZH_p statistic on the individual variants detected from RNA-seq in both the modern broilers (Ross 708 line) and legacy/heritage broilers (Illinois line). This statistic identifies chromosomal regions under selection and detecting alleles that have swept to fixation or near-fixation [42,45]. Since modern broilers breeding practices have a more recent selection history with significant success in phenotypic selection of polygenic traits, such as feed efficiency and meat yield. Comparative studies of legacy and modern broiler chickens provide an opportunity to identify regions of the genome that have undergone selection pressure by this human-directed evolution.

3.2 Materials and Method

3.2.1 Data Collection

To obtain the transcriptome, we collected 23 birds each comprising of 3 to 4 different tissues for the two commercial breeds (Ross and Illinois line), totaling 184 libraries and 164 libraries in the Ross and Illinois lines respectively (Table 3.1). The 56 pooled birds were selected based on their overall sequence read distribution and read quality using FastQC.

Table 3.1 Summary of sequencing data from each tissue of the respective chicken breeds; Ross and Illinois.

Bird #	Ross				Iliinois			
	# Tissues	Tissue(s)[a]	# Seqs[b]	% Seq. Cov.	# Tissues	Tissue(s)[a]	# Seqs[b]	% Seq. Cov.
1	4	K L P S	96.2	22.47	4	B L P S	90.8	21.22
2	4	K L P S	98.5	23.02	4	B L P S	95.5	22.31
3	4	B L P S	104.3	24.38	4	B L P S	64.4	15.06
4	4	B L P S	112.8	26.36	4	B K P S	89.6	20.94
5	4	K L P S	108.6	25.39	4	B K L S	97.1	22.69
6	4	K L P S	84.9	19.84	4	B K L S	95.2	22.26
7	4	K L P S	104.7	24.46	3	B K S	76.7	17.93
8	4	K L P S	83.9	19.61	3	B P S	64.8	15.15
9	4	K L P S	99.5	23.25	3	B K S	46.0	10.75
10	4	K L P S	102.8	24.03	3	B K S	54.5	12.73
11	4	B K L S	83.5	19.50	3	B K S	47.1	11.01
12	4	K L P S	101.9	23.82	3	B K S	54.3	12.68
13	4	A C K P	83.6	19.53	3	B K S	75.7	17.69
14	4	A B C K	67.2	15.71	3	B K S	59.1	13.81
15	4	B L P S	91.3	21.33	3	K S P	66.1	15.45
16	4	A B C K	60.3	14.10	3	C K S	88.7	20.73
17	4	A B C P	58.0	13.55	4	A C K S	95.4	22.30
18	4	K L P S	82.9	19.36	4	A C K S	83.1	19.41
19	4	K L P S	70.4	16.44	4	A C K S	96.9	22.65
20	4	K L P S	84.8	19.80	4	A C K S	80.2	18.74
21	4	B L P S	80.5	18.81	4	A C K S	73.2	17.10
22	4	K L P S	67.2	15.69	4	A C K S	96.1	22.46
23	4	K L P S	94.8	22.16	4	A C K S	97.1	22.69

Note: [a] Tissue names are indicated as follows: A – Abdominal adipose, B – Breast muscle, C – Cardiac adipose, K – Kidney, L- Liver, P – Pituitary, S – Spleen.
[b] The number of paired end sequences for each pooled-bird (Millions).

3.2.2 Pre-processing of RNA-seq Reads

The fastQ reads were pre-processed using AfterQC [53], a python program used for automatic filtering, trimming, error removing and quality control of single-end or pair-end sequencing data. The AfterQC program was used to remove adaptor

sequences, error-correct mismatch bases in overlapping pairs, filter low quality and abnormal sequences. Abnormal sequences are sequences with short lengths compared to the average sequence length, and with too many ambiguous nucleotides (N's) or with polyX (i.e. at least a string of 35 X nucleotides in the given sequence, X is one of A/T/C/G).

3.2.3 SNP Discovery using VAP workflow

SNPs were detected from the pooled RNA-seq data for each line using the VAP workflow and filtering criteria. The pair-end RNA-seq reads were aligned using three transcriptome reference assemblers, TopHat2 (v 2.1.1) [54], HiSAT2 (v 2.1.0) [55] and STAR (v 2.5.2b) [56]. The RNA-seq reads were mapped to the genome, NCBI *Gallus gallus* Build 5.0 reference genome, and converted to BAM format using SAMtools (v 1.4.1) [58]. The BAM files were processed using Picard Package (v 2.13.2) (https://broadinstitute.github.io/picard/) and variants were detected using GATK (v 3.8-0-ge9d806836) [59] and filtered as described in Chapter 2.

3.2.4 Allele Frequencies and Genotyping

Following determination of all unique variant sites, we *in silico* genotyped all variant sites using custom Perl scripts. This was done to compare all uniquely aligned reads to both the reference and variants alleles and calculate the variant allele frequencies (VAF) observed between 0 and 1 (i.e. $0 \leq VAF \leq 1$). SNPs were divided into two groups based on their VAF score; the first group are SNPs with $VAF > 0.99$ are tagged as homozygous to the alternate allele (homozygous alternate), because at least 99% of the reads mapped to the position had the mutant allele and, the second group are SNPs with $VAF \leq 0.99$ as heterozygous

3.2.5 Selection Sweep Analysis

Allele counts at SNP positions were used to identify signatures of selection in sliding windows. The sliding window approach involves overlapping a fixed window by the step size along the chromosomes and summing from sequence data the allele counts corresponding to the most and least abundant allele frequencies (n_{MAJ} and n_{MIN}) for each SNP in the given window. Unlike Rubin et al approach, each SNPs allele frequencies was used instead of the number of reads to normalize for gene expression bias. The heterozygosity score (H_w) for each window is calculated as shown in Equation 3.1, where $\sum n_{MAJ}$ is the sum of major allele frequencies, and $\sum n_{MIN}$ is the sum of minor allele frequencies within a window. Each H_w values are Z-transformed (ZH_w) as shown in Equation 3.2, where μH_w is the overall average heterozygosity and σH_w is the standard deviation of the overall heterozygosity.

$$H_w = \frac{2 \sum n_{MAJ} \sum n_{MIN}}{(\sum n_{MAJ} + \sum n_{MIN})^2}$$

Equation 3.1 Heterozygosity score equation.

$$ZH_w = \frac{H_w - \mu H_w}{\sigma H_w}$$

Equation 3.2 Z-transformed heterozygosity score equation.

We calculated the heterozygosity in sliding 20-kb windows with a 10-kb overlap step along the autosomes. For each window, we calculated the heterozygosity score and Z scores (Equations 3.1 and 3.2). Putative selective sweeps were identified as windows with $ZH_w \leq$ -3, because windows below this threshold represent the

extreme lower end of the distribution. Windows with $ZH_w \leq$ -3 were selected and the genes found with exonic SNPs were extracted for annotation analysis.

3.2.6 Annotation and Functional Analysis

Annotation analysis was performed using the Ensembl chicken gene set (Ensembl release 90) [84] and the NCBI chicken gene set (RefSeq release 86) [85] were downloaded and gene-based annotation of putative SNPs were performed using ANNOVAR (v 2017Jul16) [60]. Candidate genes and regions of putative sweeps identified were compared with chicken QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/GG/index) [86]. The genes were functionally annotated using DAVID Bioinformatics Resources software version 6.8 [87], PANTHER version 13.1 [88] and PATHRings[89].

3.3 Results and Discussion

3.3.1 Transcriptome-wide detection of SNPs

Each sequenced reads for our 56 bird samples were aligned to the chicken reference genome (*Gallus gallus* build 5.0) using the VAP workflow, on average 86% of the reads mapped to the genome spanning 16% of the reference genome across the read mapping tools (Table 3.2). SNPs identified were filtered based on the VAP filtering criteria, resulting in 652621 SNPs and 486968 SNPs in the Ross and Illinois line respectively. The Ross line is shown to have significantly (25%) more SNPs called compared to the Illinois line despite similar reference mapping coverage, this suggests larger variability in the Ross genome compared to Illinois (Table 3.3). Each SNP was subsequently grouped based on their variant allele frequency, with over 96%

having dbSNP information. As expected, a larger number of SNPs were heterozygous in Ross compared to Illinois, suggesting more loci in Illinois line have reached complete fixation compared to the Ross line (Table 3.4).

Table 3.2 Read mapping summary statistics from our VAP (Variant Analysis Pipeline) for the 56 pooled Ross and Illinois birds.

Line	**Tools**	**% reads mapped[a]**	**% reference covered[a]**
Ross	TopHat	87.9 (0.03)	16.4 (2.66)
	HiSAT	86.4 (0.02)	16.1 (2.62)
	STAR	88.1 (2.64)	16.6 (2.70)
Illinois	TopHat	86.7 (0.02)	14.9 (2.50)
	HiSAT	86.4 (0.02)	14.7 (2.46)
	STAR	86.0 (0.01)	15.1 (2.50)

Note: [a] Average (±SD) scores for each mapping tool.

Table 3.3 Quality control of identified SNPs.

Line	**Ross**	**Illinois**
Total SNPs	2139430	1693401
N3[1]	684366	552677
VF[2]	526025	440277
AR[3]	250994	194802
HS[4]	560	130
UN[5]	24864	18547
SNPs Used	**652621**	**486968**

Note: [1] SNPs not found in all three assemblers; [2]SNPs filtered using GATK Variant Filtration tool; [3] SNPs with less than 5 reads supporting alternate allele; [4]SNPs with VAF < 0.1; [5]non-autosomal SNPs on Galgal5.

Table 3.4 Comparison of SNPs identified as homozygous-alternate and heterozygous.

Line	VAF		# SNPs	Precision	ts/tv
Ross	Heterozygous	VAF < 0.99	503394	91.19%	3.02
	Homozygous Alternate	VAF ≥ 0.99	149227	97.03%	2.70
	Total		652621	94.13%	2.94
Illinois	Heterozygous	VAF < 0.99	345028	93.18%	3.02
	Homozygous Alternate	VAF ≥ 0.99	141940	96.91%	2.74
	Total		486968	94.26%	2.93

3.3.2 Nucleotide polymorphisms distribution

Comparing SNPs identified in both Ross and Illinois line, a large percentage of SNPs (55% in Ross and 74% in Illinois) were shared between both lines, while 294,182 SNPs and 128,529 SNPs were unique to Ross and Illinois lines respectively (Figure 3.1). The highest number of SNPs were observed in the first five chromosomes, while the lowest number of SNPs were observed on microchromosomes (Figure 3.2). This is consistent with the SNP distribution being a function of chromosome sequence length. Chromosomes 16 and 31 have been challenging to sequence due to the highly repetitive nature of these chromosomes, consequently SNP call to these have not been further evaluated (Figure 3.3).

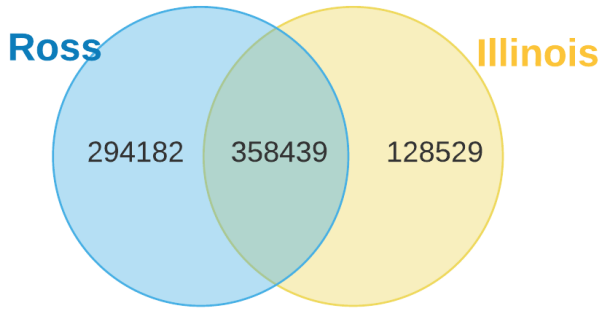


Figure 3.1 Venn diagram showing the number of shared and unique SNPs between Ross and Illinois lines.

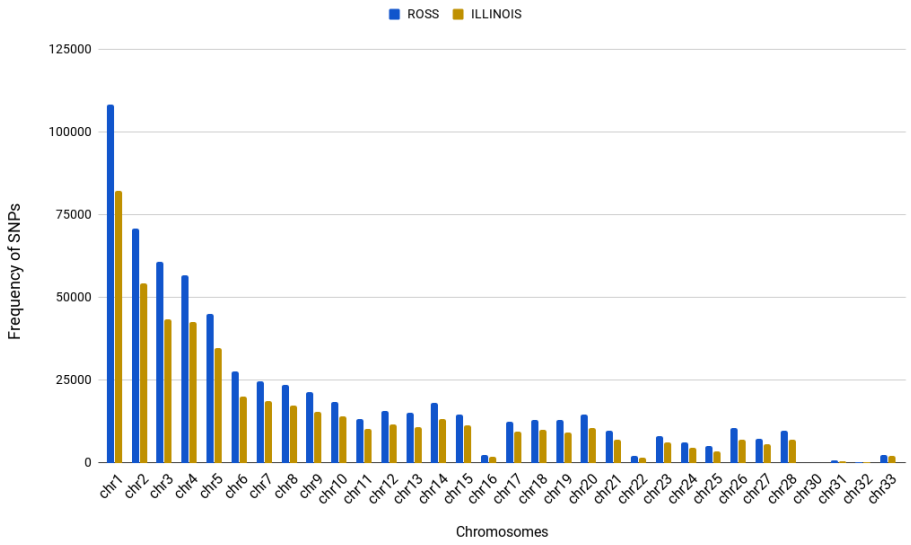


Figure 3.2 Histogram plot comparing the number of SNPs per chromosome between Ross and Illinois lines.

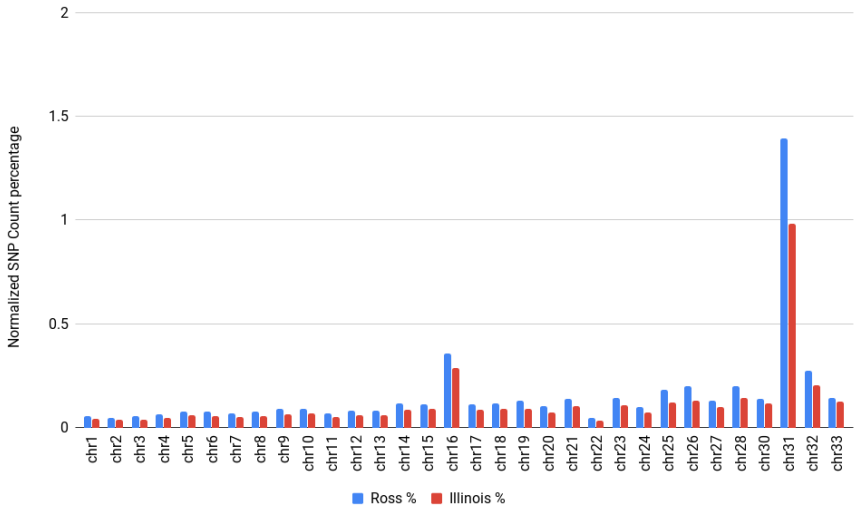


Figure 3.3 Histogram plot comparing percentage normalized SNPs per chromosome between Ross and Illinois lines.

3.3.3 Consequences of SNPs detected

17% of the SNPs were predicted to be within protein-coding regions when annotated against both the NCBI and ENSEMBL gene database for chicken in both lines; the remaining SNPs were found in non-coding or regulatory regions (Table 3.5).

Table 3.5 Classification of SNPs detected consequences.

Variant Consequences	Ross	Illinois	Overlap in Ross and Illinois
Intergenic	354937 (54.4%)	268469 (41.1%)	190075 (29.1%)
Up/downstream	32679 (5%)	24327 (3.7%)	19057 (2.9%)
Intronic	112037 (17.2%)	82274 (12.6%)	53651 (8.2%)
Exonic	108959 (17%)	79590 (12.2%)	67570 (10.4%)
Non-synonymous	26771 (4.1%)	18807 (2.9%)	15137 (2.3%)
Synonymous	82069 (12.6%)	60703 (9.3%)	52371 (8%)
Stop-gain/loss	119 (<1%)	80 (<1%)	62 (< 1%)
Splicing	19 (<1%)	15 (<1%)	10 (< 1%)
UTR3/UTR5	42016 (6.4%)	30838 (4.7%)	26865 (4.1%)
ncRNA	292 (<1%)	209 (<1%)	162 (< 1%)

3.3.4 Proof of Concept: Genes verified with Mutations arising from Domestication.

3.3.4.1 Yellow Skin Phenotype:

The yellow skin phenotype is found in most domesticated chickens and is expressed in homozygotes for the recessive allele. *Eriksson et al* identified genomic SNPs contributing to the reduced expression of the candidate gene, Beta-Carotene Oxygenase 2 (*BCO2* or *BCDO2*) [44]. We used the *BCDO2* locus to show that our approach can reveal an established sweep or fixation. In both the Ross and Illinois lines, we observed majority of the SNPs found have very high variant allele frequency

scores – homozygous of the mutation in the *BCDO2* locus (Table 3.6). This indicates, as proposed, complete or near fixation for the yellow skin allele.

Table 3.6 Partial list of SNPs found in the *BCDO2* locus (chr24:6110296-6131259).

Chromosomal Position	Mutation	Ross VAF	Illinois VAF
chr24:6113345	G/A	0.95	1.0
chr24:6113438	G/A	0.94	1.0
chr24:6113900	T/G	0.96	1.0
chr24:6116372	T/C	0.95	1.0
chr24:6116717	T/C	0.93	1.0
chr24:6121412	A/G	1.0	1.0

3.3.4.2 Domestication related mutation in the Thyroid Stimulating Hormone Receptor (*TSHR*):

TSHR is a significant gene that functions in the regulation of metabolic processes and reproduction. The *TSHR* mutation is a nonsynonymous substitution resulting in a glycine to arginine shift (Gly558Arg) [42]. In both lines, all SNPS found in the *TSHR* locus are homozygous for the alternate allele, including the candidate mutation (Table 3.7). However, due to the stringent filtering criteria, the causal SNPs were excluded in the variant filtering step of our VAP workflow; as a result of low read depth (DP < 10) and nucleotide positions with less than 5 alternative allele supporting reads. Nevertheless our data proves both lines carry the domestic TSHR allele [42]. The mutant *TSHR* gene is shown to affect reproductive traits resulting increase metabolic activity and growth [90].

Table 3.7 Exonic mutations identified at the *TSHR* locus (chr5:40811286-40858950).

Chromosomal position	Mutation	Consequence	Line
chr5:40857654	C/G	nonsynonymous	Ross(2); Illinois(2)
chr5: 40858336	***G/A***	***nonsynonymous***	***Ross(2); Illinois(1)***
chr5: 40858944	G/A	synonymous	Ross(2); Illinois(2)

Note: SNPs calls are excluded in variant filtering step as a result of (1) low read depth (DP < 10) or (2) nucleotide positions have less than 5 alternative allele supporting reads.

3.3.5 Selection Sweeps Detection.

Selection sweeps were detected using the SNPs variant allele frequencies to identify regions with high degree of fixation. We calculated the pooled heterozygosity in 20 kb window size with a 10-kb sliding step. To prevent windows containing very few SNPs from adding spurious fixation signals, we excluded windows with less than 10 SNPs. Following this criterion, 44% and 54% of the windows with SNPs < 10 were excluded from the Ross and Illinois line respectively. We observed a high degree of selection sweeps in the macrochromosomes of the *Gallus gallus* genome (Figure 3.4).

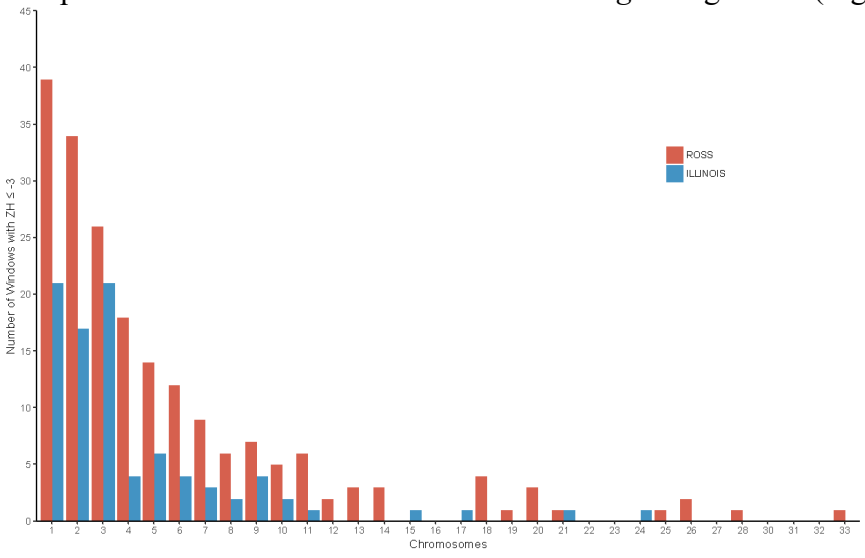


Figure 3.4 Histogram plot comparing the number of candidate regions under selection in Ross and Illinois lines.

3.3.5.1 Mean Genome Diversity

Genome heterozygosity levels were measured across both lines (Table 3.8). The highest level of heterozygosity was observed in the Ross (0.34 ± 0.095) line whereas it was lowest in the Illinois (0.32 ± 0.103) line. However, both lines reveal high variability in heterozygosity scores as opposed to prior studies, and this could possibly be as a result of the high amounts of homozygous alternate SNPs (having VAF ≥ 0.99) observed, especially in the Illinois line.

Table 3.8 Genome-wide pool heterozygosity statistics.

Line	**Ross**	**Illinois**
Windows	47183	50015
Total windows (SNPs ≥ 10)	26201	23091
Heterozygosity level	0.34 ± 0.095	0.32 ± 0.103
$ZH_w \leq -3$	272	134
Sweep regions	198	89
Genes identified	374	183

3.3.5.2 Candidate Sweep Regions

For the Ross line, out of the 26,201 windows analyzed only 272 of them passed the threshold of ≤ -3, identifying 198 candidate sweep regions. A total of 375 genes were identified in the 198 putative sweep regions (Table A.1). No significant peaks were observed on chromosomes 15, 16, 17, 22, 23, 24, 27, 30, 31 and 32 (Figure 3.4). The largest candidate sweep spanning 70-kb which is also the region with the lowest ZH_w scores (3.57± 0.025) was found on chr1:55.37Mb-55.44Mb. This region overlaps with the *IGF1*, *PMCH*, *PARPBP* genes, and has been detected to be under selection in prior studies [42,45]. This region is also known for the QTL affecting metabolic traits such as body weight, abdominal fat and muscle weight [91,92].

Overall the identified genes are well established as important enhancers to regulate growth, reproduction, energy balance, cell proliferation and cell death. Furthermore, the genes impacted by SNPs within coding regions in the candidate sweeps were the *insulin-like growth factor 1 (IGF1)* and the *PARP1 binding protein (PARPBP)* and were validated in prior studies (3.9), showing these may likely be the candidate genes under selection in this well-known QTL.

Table 3.9 Exonic nucleotide polymorphisms possibly influencing selection of candidate genes in putative sweep on chr1:55.37-55.44Mb.

Chromosomal position	Mutation	Gene Name	Consequence	VAF	dbSNP number
chr1:55374168	A/G	*IGF1*	synonymous	0.944	rs316492824
chr1:55428867	T/G	*PARPBP*	nonsynonymous	1	rs13869806
chr1:55428935	T/C	*PARPBP*	nonsynonymous	1	rs13869807
chr1:55458271	T/A	*PARPBP*	synonymous	1	rs315774625
chr1:55458294	C/T	*PARPBP*	nonsynonymous	1	rs317963948
chr1:55461367	A/G	*PARPBP*	nonsynonymous	1	rs13869828
chr1:55465970	G/T	*PARPBP*	nonsynonymous	1	rs315850110

Other candidate regions with ZH scores ≤ -3 include: a 60-kb region on chr1:54.61-54.67Mb with ZH_w score -3.08±0.018 overlapping the membrane transporter *SLC41A2* and thioredoxin reductase enzyme *TXNRD1*. One 50-kb region on chr8:1.26-1.31Mb (ZH_w=-3.20±0.06) overlapping *ENSGALG00000001983 (PRPF38B)*, *STXBP3*, *NR5A2* genes. Two candidate regions have 40-kb size each on chr2:6.41-6.45Mb (-3.34±0.15) consisting of 49 SNPs that impact the *PRKAG2* gene and chr3:64.19-64.23 (-3.32±0.07) overlapping *ENSGALG00000014979 (FRK)*, *FAM26E*, *HDAC2* genes. *PRKAG2* functions in regulating energy demand within cells and plays an important role in body weight, body weight gain, feed intake and feed conversion ratio traits [93], interestingly all the SNPs expect one synonymous SNP at

T6436892A (rs317797507), are located in introns of the *PRKAG2* gene, this can be as a result of low evolutionarily conservation of the gene and may reflect a potential isoform of the *PRKAG2* transcript [94]. 64 candidate regions have a 30-kb size each, and 129 have sizes of 20-kb (Figure 3.5).

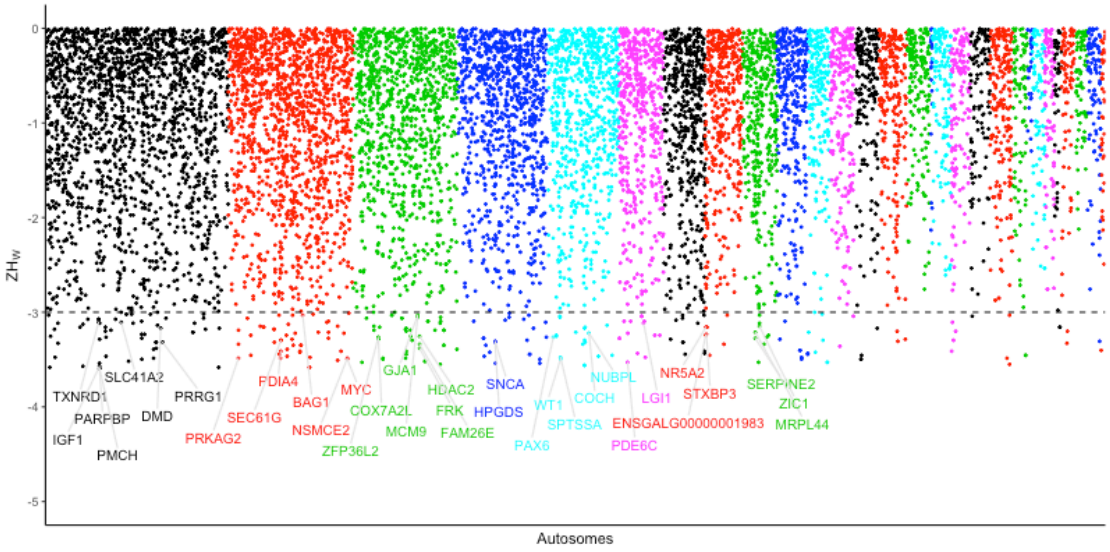


Figure 3.5 Selection sweeps analysis of the Ross line, including a partial list of candidate genes.

As for the Illinois line, out of the 23,091 windows only 134 of them passed the genome-wide threshold, defining 89 candidate sweep regions having 183 genes (Table A.1). No significant peaks were observed on chromosomes 12, 13, 14, 16, 18, 19, 20, 22, 23, 25, 26, 27, 28, 30, 31 and 32 (Figure 3.4). Analyzing candidate sweeps based on fragment size, the largest candidate sweep spanning 60-kb was found on chr3:62.42-62.48Mb with ZH_w scores -3.13±0.016. This region overlaps with the *gap junction protein alpha 1* (*GJA1)* and *minichromosome maintenance 9 (MCM9)* genes. *GJA1* polymorphisms have been found to be significantly associated with growth traits

in chickens, such as body weight and carcass weight [95], the *MCM9* functions as the *MCM8-9* complex in DNA maintenance and repair of interstrand crosslinks [96]. Other candidate regions with ZH scores ≤ -3 include: a 50-kb region on chr1:30.83-30.88Mb with ZH_w score -3.07±0.046 overlapping *ENSGALG00000033074 (ARID2)*, *DBX2* and *SCAF11* genes. Six candidate regions have 40-kb size each, 26 regions have a 30-kb size each, and 55 have size of 20-kb (Figure 3.6).

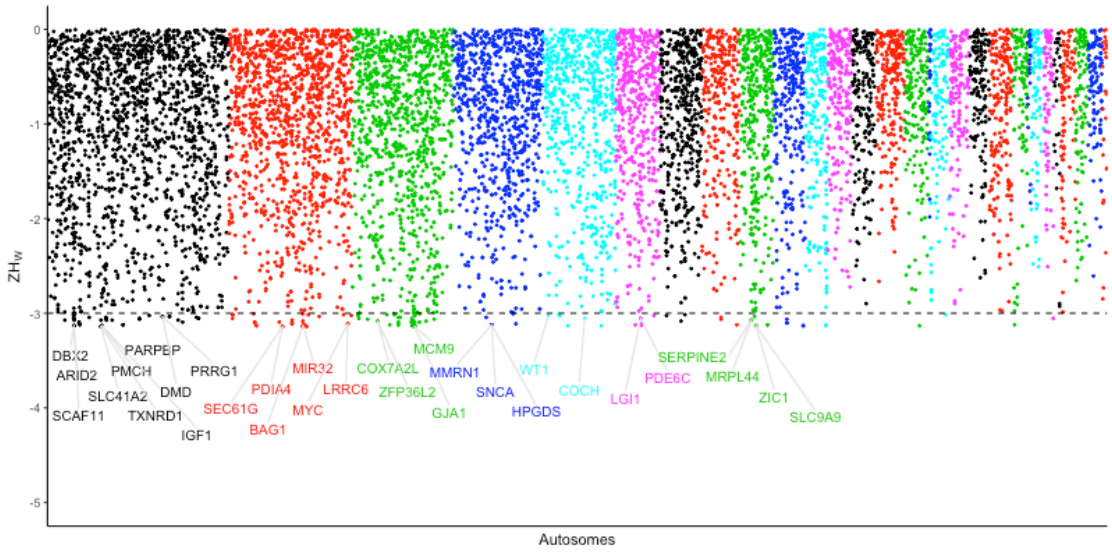


Figure 3.6 Selection sweeps analysis of the Illinois line, including a partial list of candidate genes.

3.3.5.3 Candidate Genes in Regions Detected across Populations

Examining the candidate regions under selection, we identified 14 shared regions in both Ross and Illinois lines. 33 genes were found overlapping the 14 shared regions found on chromosomes 1, 2, 3, 4, 5, 6 and 9. Notably, the majority of the shared regions are found on genes involved in growth traits, muscle and organ development (Table 3.10).

Table 3.10 Candidate Genes detected in the 14 shared regions.

Coordinates[a]	Ross		Illinois		Gene[d]
	Win[b]	ZHw[c]	Win[b]	ZHw[c]	
chr1:54.55 ~ 54.67	3	-3.08±0.02	2	-3.11±0.05	*TXNRD1* *SLC41A2*
chr1:55.37 – 55.44	4	-3.57±0.03	2	-3.13	*IGF1* *PARPBP* *PMCH*
chr1:114.18~114.50	1	-3.17	2	-3.034	*DMD* *PRRG1*
chr2:53.21~53.53	3	-3.33±0.07	2	-3.14	*PDIA4* *SEC61G*
chr2:84.58 – 84.63	1	-3.03	3	-3.07±0.06	*BAG1* *MIR32*
chr2:139.36 ~ 141.48	2	-3.49±0.01	2	-3.11	*MYC* [1]*NSMCE2* [2]*LRRC6*
chr3:24.06 ~ 24.14	2	-3.27±0.02	2	-3.08	*COX7A2L* *ZFP36L2*
chr3:62.42 – 62.75	2	-3.04±0.02	7	-3.10±0.05	*GJA1* *MCM9*
chr4:35.76 ~ 37.17	3	-3.38±0.13	2	-3.122	[2]*MMRN1* *HPGDS* *SNCA*
chr5:5.23 ~ 5.47	1	-3.26	1	-3.02	[1]*PAX6* *WT1*
chr5:34.25 ~ 34.60	2	-3.22	1	-3.05	*COCH* [1]*NUBPL* *SPTSSA*
chr6:20.32 – 20.34	1	-3.12	1	-3.05	*LGI1* *PDE6C*
chr9:8.36 ~ 8.88	2	-3.24±0.04	1	-3.08	*MRPL44* *SERPINE2*
chr9:11.80 ~ 12.13	1	-3.15	1	-3.13	[2]*SLC9A9* *ZIC1* [1]*MIR6611*

Note: [a] Chromosomal coordinates in megabases (Mb): '–' represent similar regions between Ross and Illinois, whereas '~' represent windows in close proximity between lines, such windows were merged. [b] Number of consecutive $ZH_w < -3$ windows that were merged. [c] average ZH_w (±SD) scores in region. [d] Gene(s) overlapping putative sweep regions: '[1]' were found only in Ross line, '[2]' were found only in Illinois line.

3.3.5.4 Candidate Genes found in Previous Studies

We compared the genes in the candidate selection regions with those from two previous studies on detecting selective sweeps in domesticated birds (broilers). In the Ross line, among the 374 genes identified in putative regions under selection, 41 genes were in genomic regions detected by Rubin et al (2010) and Elferink et al (2012). While in the Illinois line, 19 out of the 183 genes identified in our study were previously detected in prior studies. Table 3.11 lists the genes that overlap with previous studies.

Table 3.11 Genes that overlap with previous studies (Rubin et al., 2010 and Elferink et al., 2012).

Gene Name	Line	Chromosome	Reference
CHPT1	Ross	1	Elferink et al, 2012
DBX2	Illinois	1	Elferink et al, 2012
SIM2	Ross	1	Elferink et al, 2012
MMP2	Ross	1	Elferink et al, 2012
SLC41A2	Ross; Illinois	1	Elferink et al, 2012
TMEM18	Ross	1	Elferink et al, 2012
IGF1	Ross; Illinois	1	Rubin et al. 2010; Elferink et al, 2012
NUP37	Ross	1	Rubin et al. 2010; Elferink et al, 2012
PARPBP	Ross; Illinois	1	Rubin et al. 2010; Elferink et al, 2012
PMCH	Ross; Illinois	1	Rubin et al. 2010; Elferink et al, 2012
INHBA	Ross	2	Elferink et al, 2012
TFAP2A	Illinois	2	Elferink et al, 2012
MBOAT1	Ross	2	Rubin et al. 2010
SEC61G	Ross; Illinois	2	Rubin et al. 2010
TMX3	Ross	2	Rubin et al. 2010
TRIM55	Ross	2	Rubin et al. 2010
KIF6	Ross	3	Elferink et al, 2012
PRKN	Illinois	3	Elferink et al, 2012

Table 3.11(continued) Genes that overlap with previous studies (Rubin et al., 2010 and Elferink et al., 2012).

Gene Name	Line	Chromosome	Reference
ENSGALG00000014848	Illinois	3	Elferink et al, 2012
FRK	Ross	3	Elferink et al, 2012
NKAIN2	Illinois	3	Elferink et al, 2012
SP3	Ross	3	Elferink et al, 2012
CRIM1	Illinois	3	Rubin et al. 2010
ESR1	Ross	3	Rubin et al. 2010
ESRRG	Ross	3	Rubin et al. 2010
GPR137B	Ross	3	Rubin et al. 2010
CHRM3	Illinois	3	Rubin et al. 2010; Elferink et al, 2012
PSMD1	Ross	4	Elferink et al, 2012
PCDH10	Ross	4	Rubin et al. 2010
RAB28	Ross	4	Rubin et al. 2010
KLHL2	Ross	4	Rubin et al. 2010; Elferink et al, 2012
KATNBL1	Ross	5	Elferink et al, 2012
PAX6	Ross	5	Elferink et al, 2012
NUBPL	Ross	5	Rubin et al. 2010; Elferink et al, 2012
CH25H	Ross	6	Rubin et al. 2010
EXOC6	Ross	6	Rubin et al. 2010
PKP4	Ross	7	Elferink et al, 2012
CCDC93	Ross	7	Rubin et al. 2010
C8H1ORF146	Ross	8	Rubin et al. 2010
ENSGALG00000009458	Illinois	9	Elferink et al, 2012
ZIC1	Ross; Illinois	9	Elferink et al, 2012
AGTR1	Ross	9	Rubin et al. 2010
NCBP2	Ross	9	Rubin et al. 2010
FAM96A	Illinois	10	Elferink et al, 2012
TCF25	Illinois	11	Elferink et al, 2012
SPIRE2	Illinois	11	Elferink et al, 2012
FANCA	Illinois	11	Elferink et al, 2012
DEXI	Ross	14	Elferink et al, 2012
TVP23A	Ross	14	Elferink et al, 2012
NUBP1	Ross	14	Elferink et al, 2012
ENSGALG00000001031	Ross	19	Elferink et al, 2012
RAB22A	Ross	20	Elferink et al, 2012

Table 3.11(continued) Genes that overlap with previous studies (Rubin et al., 2010 and Elferink et al., 2012).

Gene Name	Line	Chromosome	Reference
SDHD	Illinois	24	Rubin et al. 2010; Elferink et al, 2012
CPAMD8	Ross	28	Elferink et al, 2012
MYO9B	Ross	28	Rubin et al. 2010

3.4 Conclusion

Using transcriptome sequencing (RNA-seq) data to determine regions of selection in the genome, we identified established mutations due to domestication (such as the *TSHR* and *BCDO2* locus) using the variant allele frequencies (VAF) derived from our variant calling pipeline. A significant contribution from our study is the successful application of RNA-seq to identify regions of recent selection in commercial birds using heterozygosity scores. Although our data consisted of over 20% SNPs with VAF $\geq$ 0.99 (i.e. homozygous to the alternate allele), we were able to detect well established selection sweeps such as the *IGF1*, *PARPBP* and *PMCH* locus, and also uncover novel candidate regions and possibly novel isoforms contributing to the recent selection in modern broilers such as the *PRKAG2* gene. Our findings suggest that selection in broiler birds has had a significant impact on the genes controlling traits related to growth and development, such as body weight, muscle mass and feed efficiency. Overall this study provides a useful benchmark that further analyses of identified genes under selection may reveal significant insights into understanding the traits that characterizes these distinct lines.

Chapter 4

TRANSATLASDB: AN INTEGRATED DATABASE CONNECTING EXPRESSION DATA, METADATA AND VARIANTS

(Adetunji, M. O., Lamont, S. J., & Schmidt, C. J. (2018). TransAtlasDB: an integrated database connecting expression data, metadata and variants. *Database : the journal of biological databases and curation*, *2018*, bay014.)

4.1 Abstract

High-throughput transcriptome sequencing (RNA-seq) is the universally applied method for target-free transcript identification and gene expression quantification, generating huge amounts of data. The constraint of accessing such data and interpreting results can be a major impediment in postulating suitable hypothesis, thus an innovative storage solution that addresses these limitations, such as hard disk storage requirements, efficiency and reproducibility are paramount. By offering a uniform data storage and retrieval mechanism, various data can be compared and easily investigated. We present a sophisticated system, TransAtlasDB, which incorporates a hybrid architecture of both relational and NoSQL databases for fast and efficient data storage, processing and querying of large datasets from transcript expression analysis with corresponding metadata, as well as gene-associated variants (such as SNPs) and their predicted gene effects. TransAtlasDB provides the data model of accurate storage of the large amount of data derived from RNA-seq analysis and also methods of interacting with the database, either via the command-line data management workflows, written in Perl, with useful functionalities that simplifies the complexity of data storage and possibly manipulation of the massive amounts of data

generated from RNA-seq analysis or through the web interface. The database application is currently modeled to handle analyses data from agricultural species, and will be expanded to include more species groups. Overall TransAtlasDB aims to serve as an accessible repository for the large complex results data files derived from RNA-seq gene expression profiling and variant analysis.

4.2 Introduction

RNA-seq provides a comprehensive view of the transcriptome, and can be used for abundance estimation, and identification of allele-specific expression profiles, alternative splicing, splice junction, novel transcripts and nucleotide polymorphisms [97,98]. The majority of studies adopt RNA sequencing for gene and transcript expression profiling between samples or single cells, by counting the number of mapped reads to a given gene or transcript as an estimation of expression levels [7,99–101]. While RNA-seq is primarily applied for gene expression analysis, RNA-seq is also a form of exome sequencing and recent studies have used RNA sequencing to detect sequence variation in genes expressed in the sample [11,51]. Several algorithms have been developed to estimate transcript-level expression and the widely accepted methodologies make use of the Tuxedo Suite of programs, which includes *TopHat* and *Cufflink*s [102,103], or the faster and more memory efficient, *HISAT* and *StringTie* [104]. *TopHat* [54,105] and *HISAT* [55] are reference RNA-seq read mapping algorithms, while *Cufflinks* [106] and *StringTie* [70] estimate abundance and differential expression from the alignment files. Another method for differential expression analysis is to count the number of reads overlapping genomic features of interest, using quantification programs like *featureCounts* [107] and *htseq-count* [108]

or pseudo-alignments programs like *kallisto* [109] or *Salmon* [110]. Read counts are required for a wide range of count-based statistical methods for differential expression or differential binding analyses such as *DESeq2* [111], *edgeR* [112]. Although RNA-seq is generally applied to gene expression analysis, recent studies have performed comparative analysis of RNA-seq with exome sequencing for variant detection analysis and the popularly used variant callers include *SAMtools* [113] and the *Genome Analysis Toolkit (GATK)* [59]. The different data files generated from RNA-seq analyses are typically large and complex, and can be a computational bottleneck, and expensive to store, especially with analyses that involve different sample groups [114].

Current storage programs involve centralizing publicly available datasets from related projects on the web. Such programs either entail a one-line summary of published projects, archives of actual files, or an integrative framework of various transcriptome analysis tools or biological databases [115–118]. Though these storage options provide a great resource for comparative analysis of related published works, they do not address the limitations most scientists working with "big data" experience, which is storage of the numerous, large analyses result files. Furthermore, assessing such data files in the near future will be a tedious process, leading to either reprocessing the data files or replicating the sample study, wasting time and effort. Thus, there is a need for an organizational framework allowing efficient storage of the different data results and uncomplicated access for retrieval of needed information in a meaningful way to answer biological questions. Given the lack of a uniform standard for data storage and management, resources and techniques for organizing and intelligently interpreting essential information are highly desirable.

We have created a sophisticated system, TransAtlasDB for efficiently storing, organizing and integrating the samples metadata, gene expression profiling and variant analyses results from sequenced samples. This system is a standalone database application that incorporates both classes of database technologies for both archiving and retrieving of various transcriptome analyses results. It serves as an organism-independent sample metadata and transcriptome analysis organizational framework and repository for gene-expression analysis and gene-associated variants, such as single nucleotide polymorphism (SNPs) and insertions and deletions (InDels). In addition, the application provides an extensive array of tools for uniform data storage and extraction mechanisms for convenient access to investigate potential patterns or research interests across different RNA-seq analysis. TransAtlasDB is designed for (i) archiving sample information; (ii) storing gene expression and variant analysis results; (iii) archiving metadata from the different analysis; (iv) validating data entry; (v) generating data tables for reporting; (vi) downloading viewed data tables; (vii) security and integrity of information; (viii) speed and performance in accessing large amounts of data; and (ix) uniform and lossless framework minimizing redundancy.

4.3 System Architecture

The main objective of this system is to create a platform for storing gene expression profiling and variant information from transcriptome analysis in a unified format, while maintaining data integrity and a consistent environment for data exploration.

The system is developed as a client-server architecture and implemented on a Unix/Linux system. As shown in Figure 4.1, the system architecture can be divided

into three layers: *User, Application, and Storage* Layer. The *User Layer* offers two modes of interacting with the databases: (i) A PHP interactive web environment with basic access to the databases through the hypertext transfer protocol, and (ii) A command-line Perl suite for interacting with the database. The interactive suite contains preconfigured queries of interest and allowances for custom queries as a print-out or export in user-friendly file formats. In addition, savvy users can interact directly with either database. The *Application Layer* is composed of a suite of Perl scripts and provides an abstraction with a set of procedures for the underlying complexities of parsing, validating, storing, and extracting data. This layer is composed of three major components: data validation, data deposition, and data extraction. The data validation component ensures all data files slated for storage or extraction are present, the data deposition component executes valid syntaxes for data storage, while the data extraction component functions as a post-processing service for data retrieval and fulfills requests from the User Layer.

The *Storage Layer* is responsible for storing and organizing data using design principles for databases with complex data. Similar to many biological web repositories, we applied a traditional relational data store, and due to its availability, simplicity, and flexibility, we chose the open source, SQL compliant relational database management system, MySQL (My Structured Query Language) [119]. This layer has been designed to organize data relationally, employing parent-child key relationships and enabling an efficient management of the stored data sets. Storing data in relational databases (RDBs) provide the convenience of maintaining consistency, data integrity and eliminating redundancy. However, data query performance of RDBs decreases with increased data storage. Given the large amounts

of data that transcriptome studies can generate, it is inevitable that query performance will degrade unless alternative solutions are identified. A potential solution involves partitioning the database across a set of machines, which requires often-expensive hardware and will ultimately be cumbersome and expensive to maintain, and, most importantly, not expedient to improving querying performance [120]. The requirements for our platform led us to implement a different type of database technology referred to as NoSQL. NoSQL (Not Only Structured Query Language) describes a class of technologies that provides an alternative approach to data storage compared to relational systems; most importantly they do not use relations (tables) as its storage structure and have a schema-less approach to handling large data. Due to the schema-less approach for data storage, NoSQL databases compromises on consistency within the database, and data duplication is allowed which threatens data integrity. Thus, some level of expertise and external protocols are required to adopt some form of data integrity. Regardless, NoSQL database solutions have shown significant advantages on indexing and querying performance with massive amounts of rapidly growing data compared to traditional RDB [18,121]. To ensure availability, simplicity, and accessibility for TransAtlasDB, we employed the open source FastBit NoSQL database technology. The FastBit NoSQL database implements a compressed bitmap indexing algorithm for efficient querying of large read-only datasets. It is an append-only and column oriented data store [122]. In addition, FastBit has a structured query language (SQL) interface, which guarantees synonymous access to stored data across both database systems. The FastBit SQL algebra provides a unique advantage for simultaneous data archival and retrieval of both the SQL-relational system and FastBit NoSQL system without having to learn another querying language and in turn,

the SQL-relational database will enforce data consistency. Consequently, the storage layer is a hybrid model that makes use of both MySQL and the FastBit NoSQL database technology.

TransAtlasDB hybrid architecture is the only independent database system that adopts a novel approach to database storage by the successful integration of both database technology types; RDB and NoSQL. This integration addresses the limitations of using either database systems for the organization and storage of big data, such as the decline in query performance with increasing data stored in relational databases stored is resolved using NoSQL fast access algorithm, and the lack of data integrity and structure with using NoSQL databases is resolved using relational database management systems. Thus, the database is designed to store smaller data sets such as sample information of transcriptome libraries and metadata details from transcriptome analysis in the relational database, while larger data sets such as the variants detected will be stored in the NoSQL system and both records can be queried interchangeably with the benefits of maintaining data integrity and not compromising on querying performance for massive data sets.

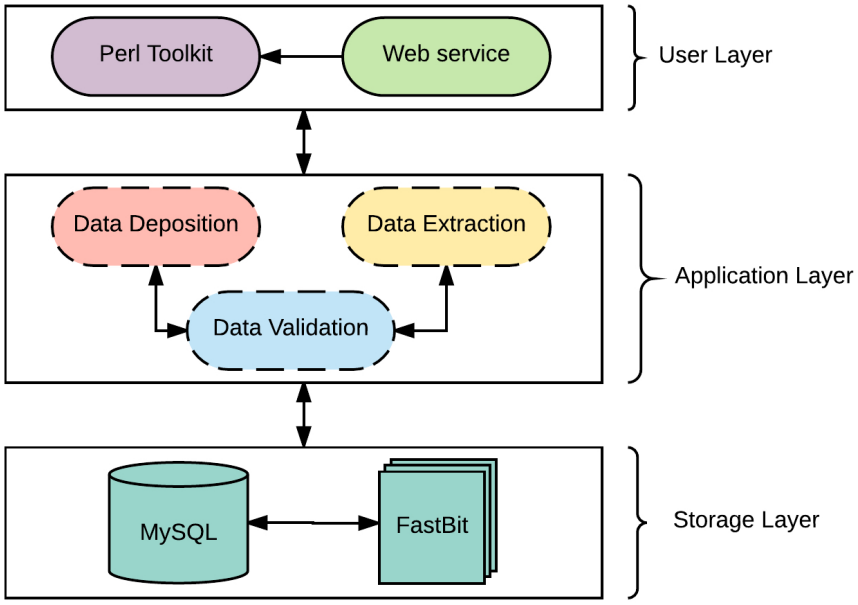


Figure 4.1 The Architecture of TransAtlasDB.

4.3.1 System Requirements

The TransAtlasDB database systems were developed using the MySQL Server (v5.5.53) and FastBit (v2.0.3), and designed to work on Unix/Linux operating systems. The software toolkit was written in Perl programming language (v5.18), with required modules listed and freely available on CPAN. The alignment/BAM file mapping parameters were reviewed using the SAMtools package (v1.3.1) [58].

The toolkit and database application were extensively tested by independent parties on the Linux Ubuntu Server (v14.04.5), Ubuntu Desktop (v16.04.2), and Mac OSX (v10.11) operating systems with the latest available MySQL server (v5.7.18), FastBit software (v2.0.3), SAMtools package (v1.5.1) and Perl programming language

(v5.22.1). The web interface was written in PHP (v7.1.4), developed using Apache (v2.4.18) and is compatible with most web browsers.

TransAtlasDB source code is available on GitHub at https://modupeore.github.io/TransAtlasDB, and detailed instructions on installation and execution are distributed with the source code.

4.4 Data Types

4.4.1 Input Data

TransAtlasDB accepts input data from the different software required for differential expression and variant detection analysis. The current version accepts outputs from the Tuxedo Suite – *TopHat2* or *HISAT2*, *Cufflinks* or *StringTie*, *kallisto*, *Salmon*, *htseq-counts* or *featureCounts*, *SAMtools/BCFtools* or *GATK*. Thus, the following information is required for successful utilization of TransAtlasDB; (a) Sample Information, (b) Alignment information, (c) Expression information and, optionally, (d) Variant information.

The *Sample information*, or metadata, is the reference point of the corresponding results from RNA-seq data and therefore important for data archival and retrieval of the various transcriptome analyses. TransAtlasDB preferably accepts the sample information using the FAANG (www.faang.org) sample submission spreadsheet template[1] to BioSamples (https://www.ebi.ac.uk/biosamples) as EMBI-

[1] https://www.ebi.ac.uk/seqdb/confluence/display/FAANG/Submission+of+samples+to+BioSamples

EBI BioSamples has the best support for sample archive. The FAANG sample submission spreadsheet template provides a detailed questionnaire for each sample and hence our database system was modeled to accept the FAANG excel template. However, the required fields in the spreadsheet are the *Animal* and *Specimen* sheets; with the *Animal*-'Sample Name', *Animal*-'Organism', *Specimen*-'Sample Name' and *Specimen*-'Organism Part' column filled. The database system also accepts a tab-delimited file with the minimum required columns of 'Sample Name', 'Derived from', 'Organism', and 'Organism Part', additional columns 'Sample description', 'First name', 'Middle Initial', 'Last name', and 'Organization' are also accepted. Definition of accepted columns is given in Table 4.1. Otherwise, the sample information can be manually inserted using SQL insert statements.

Table 4.1 Column names requirement status for Sample information tab-delimited file.

Header	**Status**	**Description**
Sample name	required	Sample identification number
Sample description	optional	Sample description
Derived from	required	Animal identification number
Organism	required	Organism name
Organism Part	required	Tissue name
First Name	optional	Person's first name
Middle Initial	optional	Person's middle Initial
Last Name	optional	Person's last name
Organization	optional	Organization

The *Alignment information* is comprised of the alignment BAM file, summary statistics file and optionally the bed files obtained from RNA-seq read mappers, TopHat2 or HISAT2. The *Expression information* consists of the genes normalized

abundance files generated using Cufflinks, Stringtie, Kallisto or Salmon containing either or both normalization procedures; FPKM (Fragments Per Kilobase of transcript per million) and TPM (Transcripts per million), and actual feature read counts using HtSeq-count, featureCounts or STAR quantMode option. The *Variant information* includes the variant VCF (variant call format) file [123] from variant callers, such as GATK [59], SAMtools[113], and many more. Table 4.2 provides an overview of applicable programs accepted in TransAtlasDB.

Table 4.2 List of programs accepted in TransAtlasDB.

Information	**Programs**
Alignment Information	
	TopHat2
	HiSAT2
	STAR
Expression Information	
	Cufflinks
	Strintie
	Kallisto
	Salmon
ReadCount information	
	htseqcount
	featureCounts
	STAR quantMode
Variant Information	
	GATK
	SamTools
Variant Annotation Information	
	VEP
	Annovar
Sequence files details (optional)	
	FastQC

Optionally, the functional annotations of variants predicted by different bioinformatics tools can also be provided in a tab-delimited format. TransAtlasDB currently accepts variant effect annotations from two annotation software; Ensembl Variant Effect Predictor, commonly known as *VEP* [124], and *ANNOVAR* [60]. The input data should be stored in a single folder for each sample (Figure 4.2).

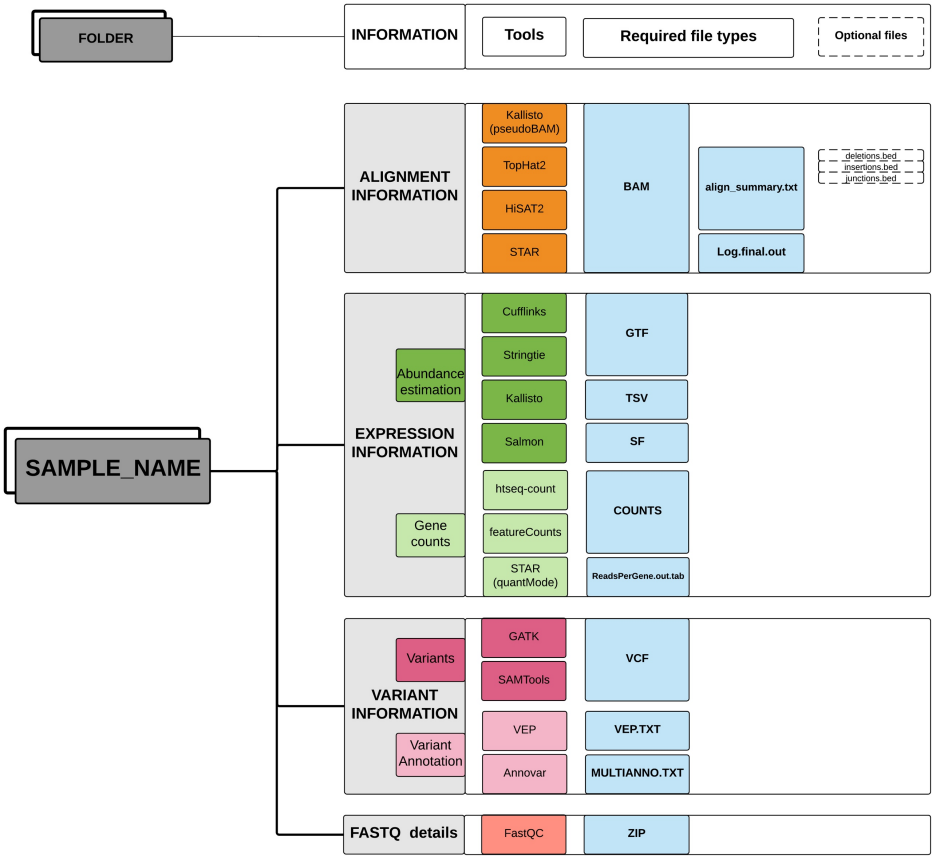


Figure 4.2 Directory structure layout for each sample. Output files (suffix) required from the specified software for the different RNA-seq analyses data types.

4.4.2 Output Format

TransAtlasDB outputs user-defined queries as a tab-delimited table. This table is the default output format which is accepted by most text editors or statistics tools such as Microsoft Excel, R and JMP software. Aside from the tab-delimited format for exporting results, the variant information can be generated as a VCF output. Predicted functional annotations and sample metadata are added in the INFO field of the VCF file, using the key "CSQ" and "MTD" respectively. Data fields are separated by "|"; the order of fields is written in the VCF header. VCFs produced by TransAtlasDB follow the standard VCF version 4 file format, and can be used for further downstream analysis or visualization using various variant viewers such as the University of California Santa Cruz (UCSC) Genome Browser [125], Integrative Genomics Viewer (IGV) [126], and other programs that accept VCF files.

4.5 Database Structure

The database system is structured using the RDB for the metadata information, alignment information, expression information and summary of the variant information, and NoSQL for the variant information. To maintain a coherent archive, the relational approach has been applied to design the basic database concept. The sample information is mapped into the relational table and the sample name will serve as the unique identification number (Id). The Id is used as the primary key for rapid indexing and enforcement of uniqueness, and the table data can be parsed using binary searching procedures across the different data types. The database design ensures mutual table relationships, and centralized checking of the foreign key constraints enforces the referential data consistency and integrity across tables. The parent-child

relationships are specified by matching the primary key of the parent table to the child tables.

4.5.1 Relational Database (MySQL) Schema

The database model (Figure 4.3) is divided into four sections corresponding to the different type of data required: sample information, alignment information, expression information and variant information. This forms a logical and simple organization of the data. The schema contains twenty-one (21) tables, six (6) views, and four (4) stored procedures, which are relevant to the organization of the different required data sets; the sample (*Sample* table) and additional information about the sample are stored in the Sample sub-tables. Transcriptome analysis results stored in the alignment summary and statistics (*MapStats* table), and the mapping metadata (*Metadata* table) are one-row descriptions for each sample and alignment details. The expression information summaries are stored in the *GeneStats* table, while the gene expression levels are stored in the NoSQL database. Variant details are organized in the *VarSummary*, *VarResult* and *VarAnnotation* tables. Table 4.3 provides a brief description of the TransAtlasDB RDB schema.

Table 4.3 Description of MySQL database schema. The MySQL schema consists of (A) 23 tables, (B) 6 views and (C) 4 stored procedures relevant for the organization of the different data sets generation from transcriptome analysis.

Attributes	**Description**
TABLES	
Animal	Animal information
AnimalStats	Additional information on Animal
Breed	Organism Breed
CommandSyntax	Analysis data commands
DevelopmentalStage	Organism developmental stage
GeneStats	Expression information summary
HealthStatus	Organism health status
MapStats	Alignment information and statistics
Material	Type of Sample
Metadata	Alignment information summary
Organism	Organism information
Organization	Organization of scientist
Person	Scientist information
ReadCounts	Raw counts details
Sample	Sample information
SampleOrganization	Cross reference of Sample and Organization
SamplePerson	Cross reference of Sample and Person
SampleStats	Additional information on Sample
Sex	Sex of Organism
Tissue	Organism part
VarAnnotation	Variant annotation information
VarResult	Variant information
VarSummary	Variant information summary
VIEWS	
vw_nosql	Sample details
vw_nosql	Prototype of NoSQL template
vw_sampleinfo	Summary analysis and statistics of each sample
vw_seqstats	Sequencing Metadata of all RNAseq analysis
vw_vanno	Variant annotation details
vw_vvcf	Prototype of VCF template
PROCEDURES	
usp_vall	Variants information in organism
usp_vchrom	Variants information of a chromosome
usp_vchrposition	Variants information of a chromosomal region
usp_vgene	Variants information of a gene

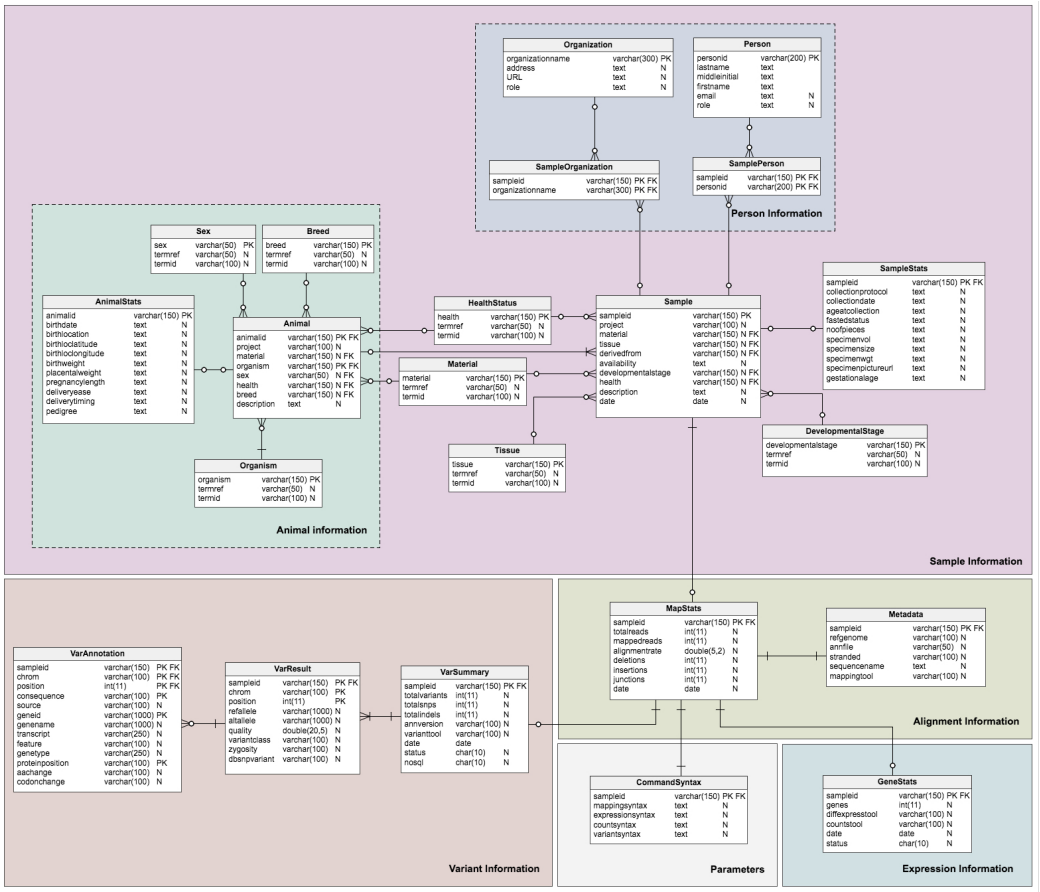


Figure 4.3 Schema of the TransAtlasDB Relational Database system. The MySQL tables are grouped by data stored (i.e. Sample Information, Alignment Information, Expression Information, and Variants Information).

4.5.2 Non-Relational Database (NoSQL) Schema

In order to prevent poor query performance in the RDB due to the large volumes of data stored, our current system implements the NoSQL database, FastBit, for archiving of both the gene-expression analysis and gene-associated variant analysis results, using custom transfer protocols from MySQL to the FastBit system. FastBit stores data as tables with rows and columns and makes an index for each column

instead of each row as in RDBs. Thus, the expression and variant information are organized with the same corresponding field names as depicted in the relational database (Table 4.4). This naming scheme allows a fluid interchangeable interaction with both the MySQL and FastBit platform.

Table 4.4 Fields in FastBit system for querying. FastBit fields are similar to the (A) variant information tables, (B) expression information tables and (C) gene counts information in the RDB, allowing synonymous access to queries data across both systems.

Fields		Type	Description
A.			
	sampleid	text	Sample Id
	chrom	key	Reference chromosome
	position	int	Reference Position
	refallele	char	Reference allele
	altallele	char	Alternate allele
	quality	double	Variant Quality
	dbsnpvariant	text	dbSNP membership number
	variantclass	key	Type of variant
	zygosity	key	Genotype
	source	text	Source of annotation
	consequence	text	Variant consequence
	geneid	text	Gene Id (from NCBI or Ensembl)
	genename	text	Gene short name
	transcript	text	Transcript Id (if provided)
	feature	text	Feature annotation
	genetype	text	Location of variant
	proteinposition	int	Relative position of aminoacid in protein
	aachange	text	Aminoacid change
	codonchange	text	Alternative codon with the variant
	organism	text	Organism name
	tissue	text	Tissue

Table 4.4 (continued) Fields in FastBit system for querying. FastBit fields are similar to the (A) variant information tables, (B) expression information tables and (C) gene counts information in the RDB, allowing synonymous access to queries data across both systems.

Fields		Type	Description
B.			
	sampleid	text	Sample Id
	chrom	key	Gene/Feature chromosome
	start	int	Gene/Feature start position
	stop	int	Gene/Feature end position
	genename	text	Gene short name (if available)
	geneid	text	Gene Id(s) associated with the gene/feature
	coverage	double	Estimated absolute depth of read coverage for the gene/feature
	tpm	double	TPM of the Gene/Feature
	fpkm	double	FPKM of the Gene/Feature
	fpkmconflow	double	The lower bound of the 95% confidence interval on the FPKM of the Gene/Feature
	fpkmconfhigh	double	The upper bound of the 95% confidence interval on the FPKM of the Gene/Feature
	fpkmstatus	char	Quantification status for the Gene/Feature
	genename	text	Gene short name
	tissue	text	Tissue
C.			
	sampleid	text	Sample Id
	genename	text	Gene short name (if available)
	readcount	Int	Read counts per Gene
	organism	text	Organism name
	tissue	text	Tissue

4.6 Package Toolkit

4.6.1 Package Components

TransAtlasDB system provides a command-line toolkit and can be used on diverse hardware systems where standard Perl modules and the Perl-DBD module are installed. The toolkit contains a suite of Perl scripts for handling the varied and large amounts of data generated from gene expression profiling and variant detection analysis. The suite serves several purposes: data entry into and data retrieval from the database(s); data browsing, double entry data validation, and verification of the different data files specified; completeness of data import; generation of complex reports and exports dynamic user-defined queries; and extracts subsets of data in tab delimited format. This suite provides semi-automated solutions that simplify the complexity of data storage and data querying methods by creating a user-friendly data management workflow. A brief outline of the package design is provided in Table 4.5.

Table 4.5 Scripts within the TransAtlasDB toolkit.

File Name	module	Description
INSTALL-tad.pL		Database installation module
connect-tad.pL		MySQL & FastBit re-connection application
tad-import.pl		
	metadata	Database sample import module
	data2db	Database import module
	delete	Database sample delete module
tad-interact.pl		Database interactive module with pre-configured database queries
tad-export.pl		
	query	User database queries
	db2data	Database retrieval module
example/		Folder with sample files

Additionally, the package offers a simple and quick installation procedure for setting up the database systems (via INSTALL-tad.pL). The detailed description of the source code and suite functionality with an example of usage are accessible via https://modupeore.github.io/TransAtlasDB/tutorial.html. The basic hardware and software requirements, short instruction of the installation and some test files are distributed within the package directory.

4.6.2 Toolkit Usage

TransAtlasDB Perl toolkit (Table 4.5) is a user-friendly framework that inputs, organizes, validates, archives and process complex transcriptome analyses data. The summaries of every transaction will be stored in log files for future reference. A pictorial representation of the procedures for data import and export are shown in Figure 4.4 and Figure 4.5 respectively.

4.6.2.1 Installation of TransAtlasDB

The TransAtlasDB database system and necessary components need to be installed to a local disk using *INSTALL-tad.pL*. The MySQL server and FastBit software should have been previously installed and added to the systems' or users' executable path. Only the '–password' argument is required if the user has admin privileges to the MySQL server. Otherwise, additional arguments, such as the '–username', '–databasename' will be needed. The NoSQL folder-name and location can be optionally specified; if not done, a default folder '*transatlasfb*' will be created in the working directory. The installation module needs to be carried out once per local disk to prevent database access conflict. However, if such conflict arises user settings can be viewed and, if needed, corrected (via *connect-tad.pL*).

4.6.2.2 Import Data using *tad-import.pl*

The sample information or sample metadata consists of the relevant details needed to uniquely identify each specimen used for RNAseq. The sample metadata can be imported (via the '–metadata' argument) from either the FAANG sample submission spreadsheet or a tab-delimited file (Figure 4.4[a]). The sample name must be unique for each sample and should follow the sample-naming-scheme of the FAANG BioSamples group – short species code, laboratory or institute short name, and alphanumeric sample ID – separated by underscore. For instance, the sample name, *GGA_UD_1004*, represents *Gallus gallus* species from University of Delaware with sample ID 1004. The sample information is also the reference point for the resulting data from transcriptome analysis with such sample.

After importing the sample metadata, the transcriptome analysis results can be inserted using the '–data2db' argument (Figure 4.4[b]). The transcriptome profiling data or variant analysis data can be imported together ('–all' flag) or separately ('–gene' or '–variant' flag) with data files in the directory structure presented in Figure 4.2. The variant functional-annotations predicted from either VEP or ANNOVAR can also be imported using additional flag ('–vep', '–annovar' respectively) and must be in their default tab-delimited format. If VEP, the filename should end with '.vep.txt', or else if ANNOVAR, the file having suffix '.multianno.txt' will be accepted.

Be aware that analysis results for a sample can only be imported once to ensure data integrity, nonetheless, previously imported data can be cautiously deleted using the '–delete' argument followed by sample name.

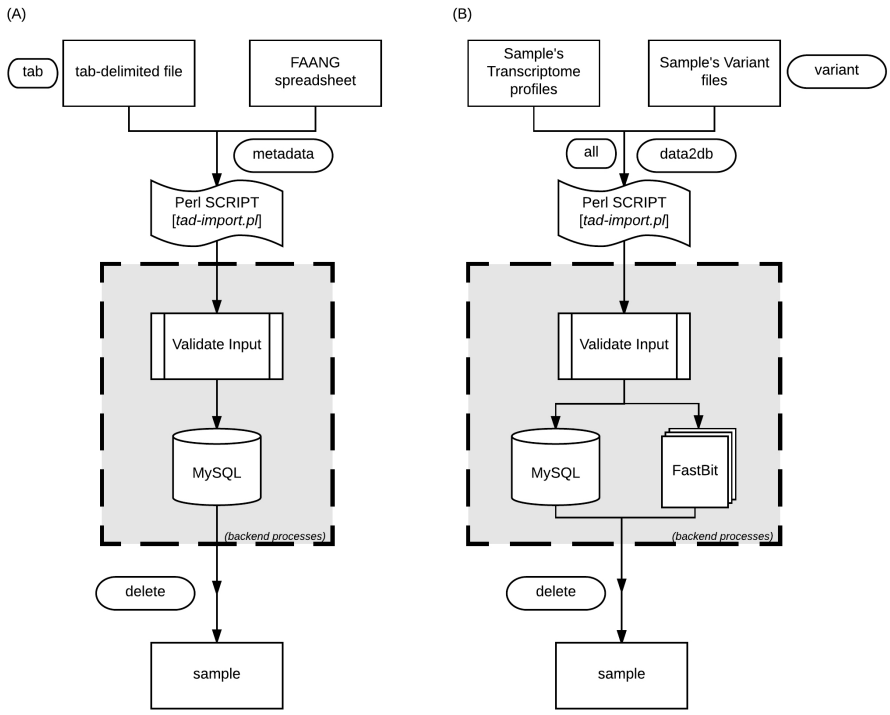


Figure 4.4 Data import procedure using tad-import.pl and available options for (A) samples metadata and (B) RNAseq data respectively.

4.6.2.3 Export Data using *tad-export.pl*

The export module offers two methods of extracting data from the database. One method allows users to execute direct data manipulation language (DML) SQL statements to the relational database (using the '–query' argument). For instance, executing the query '*show tables*' will retrieve all the rows currently in the database, which can be stored as a tab-delimited file.

The second method (via '–db2data' argument) consists of four options that are of research interest: (1) Average expression values of specified genes organized by the different tissues. (2) Gene expression profiles across the different samples of the same organism. Specific samples can be selected. (3) Variant distribution of all, or selected

chromosomes for individual samples in the database. (4) Variants and predicted functional annotations found in the organism or selected genes or chromosomes. The exported results can be written as a tab-delimited table or VCF output for variants (Figure 4.5).

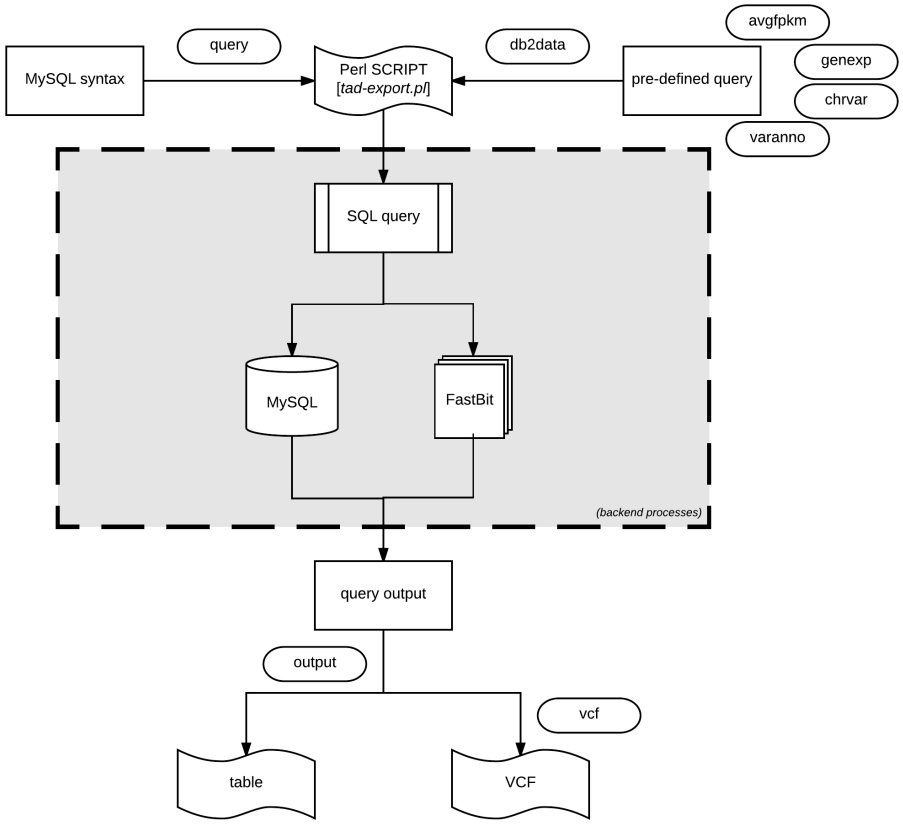


Figure 4.5 Data export procedure using tad-export.pl and available options either executing a MySQL query syntax or choosing from the four defined (avgfpkm, genexp, chrvar, and varanno) options.

If uncertain how to proceed with the export module, the interaction module (via *tad-interact.pl*) provides an easy-to-use menu-driven interface. The menu offers seven choices of exploratory research interest and provides a detailed description of

what can be done from the module. With little effort, it is self-explanatory to use. The interaction module only displays a small subset of results, nevertheless, further instructions on how to export the complete results will be displayed.

4.6.3 Web Portal & Use Cases

The PHP web environment provides another user-friendly access to the TransAtlasDB database system. The web portal provides detailed overview of the samples currently archived in the system, and options to query and export requested data from the database system. It relies on the perl command-line toolkit for interacting with the databases.

The use cases below are some examples of how biological inferences can be derived from the various RNAseq analysis data files stored in our TransAtlasDB system using the web environment. These examples can also be retrieved in the command-line toolkit provided.

TransAtlasDB web interface is comprised of five sections: (i) About page gives a summary of the samples archived in the database. (ii) Data Import page provides two methods of importing the samples metadata; either by uploading a sample file (FAANG spreadsheet or template tab-delimited file) or by manual entry. Storing the large data files such as the gene expression profiling and variants analysis results can only be done using the Perl toolkit as explained above. (iii) SQL Query page executes specified SQL queries to both databases. (iv) Metadata page displays the samples stored in the database and an overview of each sample storage-status as well as the analysis summary where applicable. The samples can be exported as a tab-delimited file. (v) Gene Expression page; the gene expression data can be viewed based on the individual sample-gene expression or average expression profiles of

multiple genes across all samples and tissues. By specifying one or more genes by their gene symbols, a fuzzy search is performed based on the characters specified. (vi) Variant page; variants can be viewed through querying gene symbols or chromosomal regions.

4.6.3.1 SQL queries

TransAtlasDB allows for the execution of SQL data query language (DQL) to both the MySQL relational database and FastBit NoSQL database using the appropriate SQL DQL syntaxes. This feature provides users unrestricted access to both database content without the limitation of having to interact with the command line. Select statements performed on either database will return a table of records based on the select expression. For instance, viewing the mapping metadata of all samples stored in TransAtlasDB; executing '*select * from Metadata*' will provide all records stored in the *Metadata* table (Figure 4.6[A]). FastBit provides an interactive bitmap index search (ibis) which are identical to SQL DQL statements but recognizes a limited number of attributes compared to the relational database. Select statements executed to the NoSQL directories will not require the FROM clause to be specified, rather the NoSQL directories can be selected from the drop-down menu provided. Figure 4.6[B] displays the *gene-information* records for this statement: '*select sampleid, chrom, start, stop, genename where organism like "Canis familia%" and genename != "NULL" order by chrom limit 10*'.

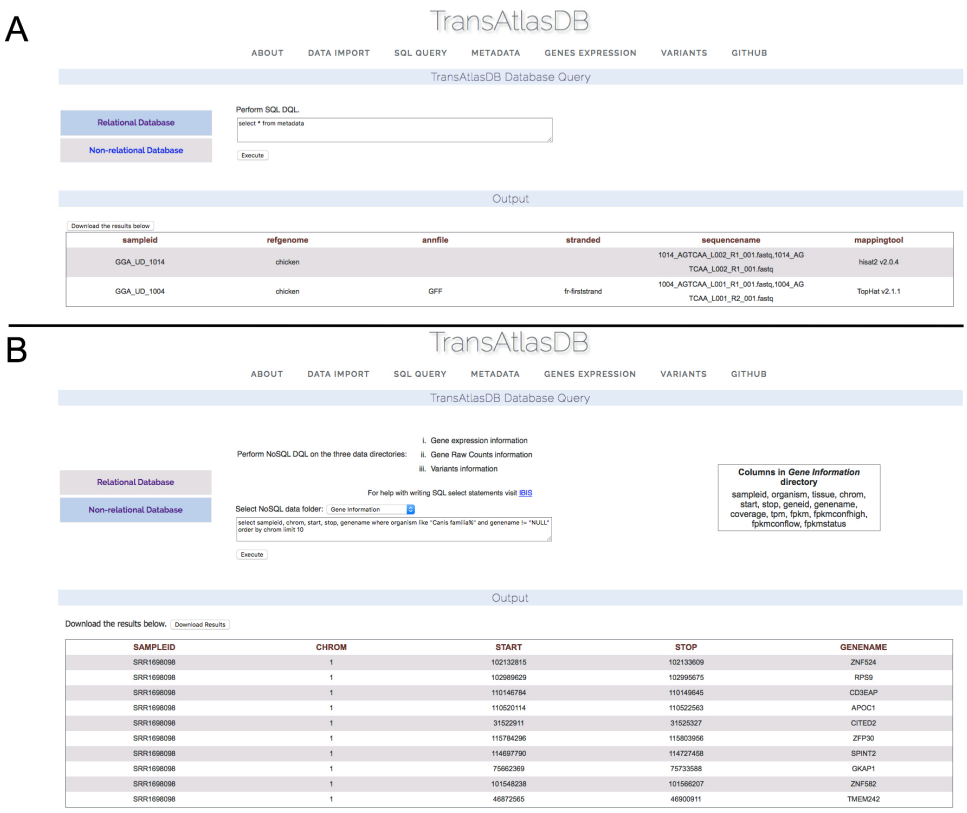


Figure 4.6 Performing SQL DQL via the web interface to the (A) relational and (B) nonrelational database.

4.6.3.2 Summary of TransAtlasDB content and Analyses metadata

Descriptive tables of the database content and status of the analyses data import can be displayed to provide users a way to quickly visualize samples already archived in the database and get the current status of all the samples archived in the database. Summary tables can be viewed in the About page (Figure 4.7), and Analyses data import status tables can be viewed in the Metadata page (Figure 4.8).

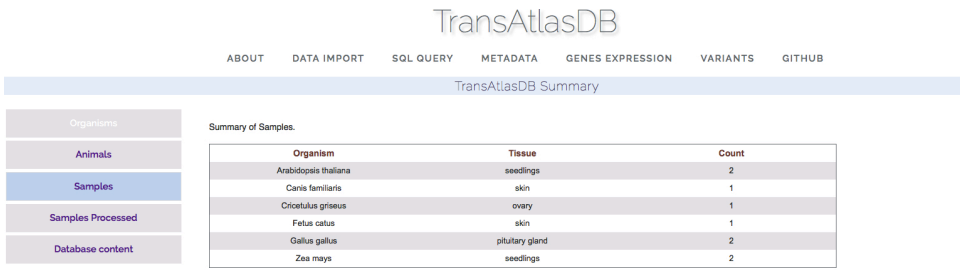


Figure 4.7 Various summary tables displaying database content in the About page.

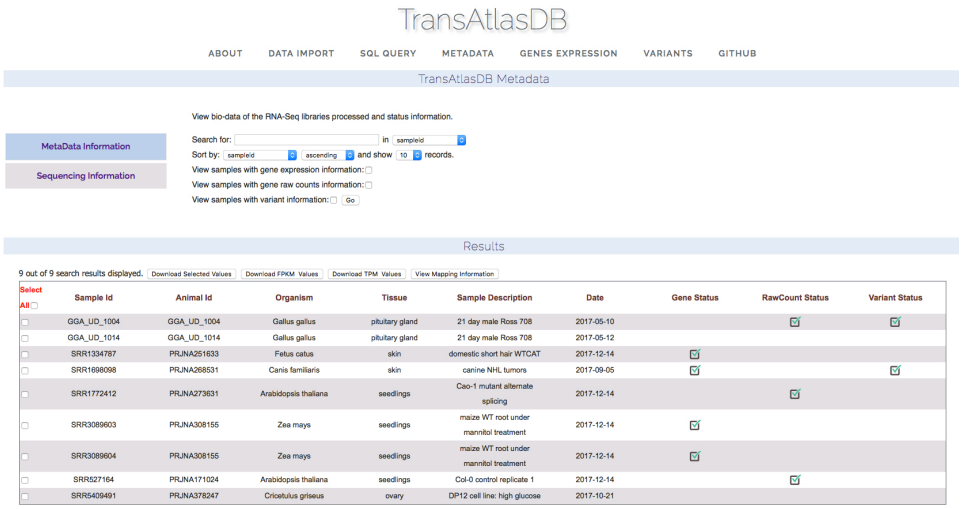


Figure 4.8 Various summary tables displaying database content in the About page.

4.6.3.3 Investigating gene expression levels & variants.

Based on the example data files provided, two Gallus *gallus* samples from the Pituitary gland, were previously imported into TransAtlasDB. Consider examining a summary of the Optineurin (OPTN) gene from the samples in the database (Figure 4.9[A]) on close inspection the summary of OPTN reveals identical minimum, average and maximum fpkm values indicating the gene may have identical fpkm values across all samples. Further exploration based on individual samples reveals OPTN may not be expressed in one of the samples, GGA_UD_1014, despite being obtained from the

same tissue, Pituitary gland (Figure 4.9[B]). These results can be exported as a tab-delimited file and adapted into statistical packages such as R or JMP for further analysis.

Unsure of the reason for different expression of the OPTN gene between the two samples, the variants can be examined by specifying the gene name (Figure 4.9[C]) or chromosomal region (Figure 4.9[D]). Multiple synonymous SNPs were found along the OPTN genomic region in sample GGA_UD_1014, while GGA_UD_1004 had only one synonymous SNP. The large number of SNPs in the GGA_UD_1014 sample, though synonymous, may have an effect on gene expression or mRNA stability and serves as a potential avenue for further analysis. These results can also be exported as a tab-delimited file for statistical analysis or as a VCF file to be used for downstream analysis or visualization.

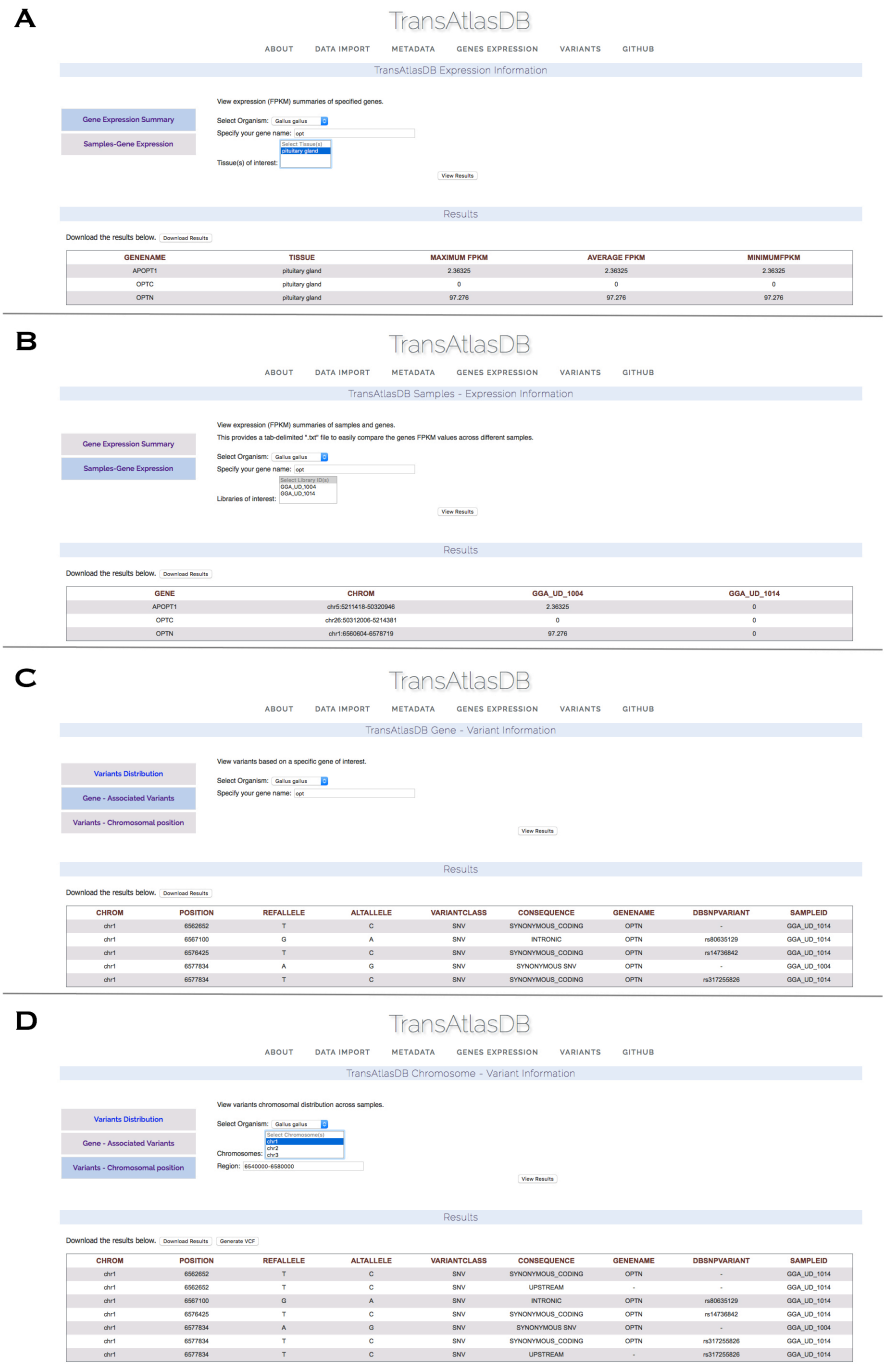


Figure 4.9 Use Cases via the web interface. (A) Genes summary expression levels across all samples. (B) Genes fpkm expression level for each sample. (C) Variants found in the OPTN gene. (D) Variants found around the chromosomal region of the OPTN gene.

4.7 Future Developments

The limitless resource potential of TransAtlasDB provides numerous options for expansion to integrate data files from other transcriptomic analyses studies and other next generation sequencing platforms like Exome sequencing. The database system is currently being extended to integrate analyses data files from human cancer studies.

4.8 Conclusion

The TransAtlasDB system should serve as a useful management platform for samples metadata and data derived from transcriptome analyses. Users can expertly store sample information and RNAseq analyses results and retrieve needed data based on specified query either using the Perl toolkit or web environment provided. TransAtlasDB provides the abstract layer with methods for data manipulation with minimum efforts to install a running system. TransAtlasDB adopts a hybrid infrastructure containing both types of database technologies; the relational database system maintains data in an organized form that eliminates data redundancy and enables efficient data management while the NoSQL system provides fast indexing and query performance that scales beyond the capabilities of relational databases. This makes TransAtlasDB a sophisticated database system capable of storing, organizing, and maintaining massive and complex transcriptome analyses data without compromise in performance whilst enforcing data integrity. The modular architecture of the system makes it possible to expand and integrate other analysis procedures potentially needed in the future. The database application is currently modeled to handle analyses data files from agricultural species and will be expanded to include

more species groups. A major advantage is that the platform can be installed locally, where users can personalize the hardware/software environment and data to import for storage, organization, access, and exchange of biological data. The modular architecture of the toolkit also enables addition of any extensions needed by the user. It is believed TransAtlasDB will be a useful and user-friendly environment for transcriptome analyses database storage.

4.9 Acknowledgement

This project was supported by Agriculture and Food Research Initiative Competitive Grant 2011-67003-30228 from the United States Department of Agriculture National institute of Food and Agriculture.

REFERENCES

1. Wolf JBW. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour. 2013;13: 559–572. doi:10.1111/1755-0998.12109

2. Cammen KM, Andrews KR, Carroll EL, Foote AD, Humble E, Khudyakov JI, et al. Genomic Methods Take the Plunge: Recent Advances in High-Throughput Sequencing of Marine Mammals. J Hered. Oxford University Press; 2016;107: 481–95. doi:10.1093/jhered/esw044

3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008/05/01. 2008;320: 1344–1349. doi:10.1126/science.1158441

4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. Nature Publishing Group; 2009;10: 57–63. doi:10.1038/nrg2484

5. Pereira MA, Imada EL, Guedes RLM. RNA-seq: Applications and Best Practices. Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health. InTech; 2017. doi:10.5772/intechopen.69250

6. Haas BJ, Zody MC. Advancing RNA-Seq analysis. Nat Biotechnol. Nature Publishing Group; 2010;28: 421–423. doi:10.1038/nbt0510-421

7. Oshlack A, Robinson MD, Young MD, Pan Q, Shai O, Lee L, et al. From RNA-seq reads to differential expression results. Genome Biol. BioMed Central; 2010;11: 220. doi:10.1186/gb-2010-11-12-220

8. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. SAGE Publications; 2015;9: 29–46. doi:10.4137/BBI.S28991

9. Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y. The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. BMC Genomics. BioMed Central; 2017;18: 690. doi:10.1186/s12864-017-4022-x

10. Sheng Q, Zhao S, Li C-I, Shyr Y, Guo Y. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. Genomics. Academic Press; 2016;107: 163–169. doi:10.1016/J.YGENO.2016.03.006

11. Piskol R, Ramaswami G, Li JB. Reliable Identification of Genomic Variants from RNA-Seq Data. Am J Hum Genet. 2013;93: 641–651. doi:10.1016/j.ajhg.2013.08.008

12. Neums L, Suenaga S, Beyerlein P, Anders S, Koestler D, Mariani A, et al. VaDiR: an integrated approach to Variant Detection in RNA. Gigascience. Oxford University Press; 2018;7: 1. doi:10.1093/gigascience/gix122

13. Chickerur S, Goudar A, Kinnerkar A. Comparison of Relational Database with Document-Oriented Database (MongoDB) for Big Data Applications. 2015 8th International Conference on Advanced Software Engineering & Its Applications (ASEA). IEEE; 2015. pp. 41–47. doi:10.1109/ASEA.2015.19

14. Lawrence R. Integration and Virtualization of Relational SQL and NoSQL Systems Including MySQL and MongoDB. 2014 International Conference on Computational Science and Computational Intelligence. IEEE; 2014. pp. 285–290. doi:10.1109/CSCI.2014.56

15. Gyorodi C, Gyorodi R, Pecherle G, Olah A. A comparative study: MongoDB vs. MySQL. 2015 13th International Conference on Engineering of Modern Electric Systems (EMES). IEEE; 2015. pp. 1–6. doi:10.1109/EMES.2015.7158433

16. Cattell R. Scalable SQL and NoSQL data stores. ACM SIGMOD Rec. 2011;39: 12. doi:10.1145/1978915.1978919

17. Atzeni P, Bugiotti F, Rossi L. Uniform Access to Non-relational Database Systems: The SOS Platform. Springer, Berlin, Heidelberg; 2012. pp. 160–174. doi:10.1007/978-3-642-31095-9_11

18. Parker Z, Poe S, Vrbsky S V. Comparing NoSQL MongoDB to an SQL DB. Proceedings of the 51st ACM Southeast Conference on - ACMSE ’13. New York, New York, USA: ACM Press; 2013. p. 1. doi:10.1145/2498328.2500047

19. Vilaça R, Cruz F, Pereira J, Oliveira R. An Effective Scalable SQL Engine for NoSQL Databases. Springer, Berlin, Heidelberg; 2013. pp. 155–168. doi:10.1007/978-3-642-38541-4_12

20. Fisher MR. A Sector Model--The Poultry Industry of the U.S.A. Econometrica. 1958;26: 37. doi:10.2307/1907383

21. Thornton PK. Livestock production: recent trends, future prospects. Philos

Trans R Soc Lond B Biol Sci. The Royal Society; 2010;365: 2853–67. doi:10.1098/rstb.2010.0134

22. AL-NASSER A, Al-KHALAIFA H, AL-SAFFAR A, KHALIL F, ALBAHOUH M, RAGHEB G, et al. Overview of chicken taxonomy and domestication. Worlds Poult Sci J. Cambridge University Press on behalf of World's Poultry Science Association; 2007;63: 285. doi:10.1017/S004393390700147X

23. Lamont SJ. Perspectives in Chicken Genetics and Genomics. Poult Sci. Oxford University Press; 2006;85: 2048–2049. doi:10.1093/ps/85.12.2048

24. Burt DW. Emergence of the Chicken as a Model Organism: Implications for Agriculture and Biology. Poult Sci. Oxford University Press; 2007;86: 1460–1471. doi:10.1093/ps/86.7.1460

25. Sawai H, Kim HL, Kuno K, Suzuki S, Gotoh H, Takada M, et al. The origin and genetic variation of domestic chickens with special reference to junglefowls Gallus g. gallus and G. varius. PLoS One. Public Library of Science; 2010;5: e10639. doi:10.1371/journal.pone.0010639

26. Garnham L, Løvlie H. Sophisticated Fowl: The Complex Behaviour and Cognitive Skills of Chickens and Red Junglefowl. Behav Sci (Basel). 2018;8: 13. doi:10.3390/bs8010013

27. Pitt J, Gillingham PK, Maltby M, Stewart JR. New perspectives on the ecology of early domestic fowl: An interdisciplinary approach. J Archaeol Sci. 2016;74: 1–10. doi:10.1016/j.jas.2016.08.004

28. Schmidt CJ, Persia ME, Feierstein E, Kingham B, Saylor WW. Comparison of a modern broiler line and a heritage line unselected since the 1950s. Poult Sci. Oxford University Press; 2009;88: 2610–2619. doi:10.3382/ps.2009-00055

29. Zuidhof MJ, Schneider BL, Carney VL, Korver DR, Robinson FE. Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 2005. Poult Sci. Oxford University Press; 2014;93: 2970–2982. doi:10.3382/ps.2014-04291

30. Resnyk CW, Carré W, Wang X, Porter TE, Simon J, Le Bihan-Duval E, et al. Transcriptional analysis of abdominal fat in chickens divergently selected on bodyweight at two ages reveals novel mechanisms controlling adiposity: validating visceral adipose tissue as a dynamic endocrine and metabolic organ. BMC Genomics. BioMed Central; 2017;18: 626. doi:10.1186/s12864-017-4035-5

31. Peters J, Lebrasseur O, Deng H, Larson G. Holocene cultural history of Red

jungle fowl (Gallus gallus) and its domestic descendant in East Asia. Quat Sci Rev. Pergamon; 2016;142: 102–119. doi:10.1016/J.QUASCIREV.2016.04.004

32. Moiseyeva IG, Romanov MN, Nikiforov AA, Sevastyanova AA, Semyenova SK. Evolutionary relationships of Red Jungle Fowl and chicken breeds. Genet Sel Evol. BioMed Central; 2003;35: 403. doi:10.1186/1297-9686-35-5-403

33. Qanbari S, Simianer H. Mapping signatures of positive selection in the genome of livestock. Livest Sci. Elsevier; 2014;166: 133–143. doi:10.1016/J.LIVSCI.2014.05.003

34. de Simoni Gouveia JJ, da Silva MVGB, Paiva SR, de Oliveira SMP. Identification of selection signatures in livestock species. Genet Mol Biol. Sociedade Brasileira de Genética; 2014;37: 330–342.

35. Cogburn LA, Porter TE, Duclos MJ, Simon J, Burgess SC, Zhu JJ, et al. Functional Genomics of the Chicken--A Model Organism. Poult Sci. Oxford University Press; 2007;86: 2059–2094. doi:10.1093/ps/86.10.2059

36. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. 2017; doi:10.1534/g3.116.035923

37. Davis RVN, Lamont SJ, Rothschild MF, Persia ME, Ashwell CM, Schmidt CJ. Transcriptome Analysis of Post-Hatch Breast Muscle in Legacy and Modern Broiler Chickens Reveals Enrichment of Several Regulators of Myogenic Growth. van Wijnen A, editor. PLoS One. Public Library of Science; 2015;10: e0122525. doi:10.1371/journal.pone.0122525

38. Stainton JJ, Charlesworth B, Haley CS, Kranis A, Watson K, Wiener P. Use of high-density SNP data to identify patterns of diversity and signatures of selection in broiler chickens. J Anim Breed Genet. John Wiley & Sons, Ltd (10.1111); 2017;134: 87–97. doi:10.1111/jbg.12228

39. Fleming DS, Koltes JE, Fritz-Waters ER, Rothschild MF, Schmidt CJ, Ashwell CM, et al. Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. BMC Genomics. BioMed Central; 2016;17: 812. doi:10.1186/s12864-016-3147-7

40. Li J, Li R, Wang Y, Hu X, Zhao Y, Li L, et al. Genome-wide DNA methylome variation in two genetically distinct chicken lines using MethylC-seq. BMC Genomics. BioMed Central; 2015;16: 851. doi:10.1186/s12864-015-2098-8

41. Su S, Miska KB, Fetterer RH, Jenkins MC, Lamont SJ, Wong EA. Differential

expression of intestinal nutrient transporters and host defense peptides in Eimeria maxima-infected Fayoumi and Ross chickens. Poult Sci. Oxford University Press; 2018;97: 4392–4400. doi:10.3382/ps/pey286

42. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. Nature Publishing Group; 2010;464: 587–591. doi:10.1038/nature08832

43. Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, Turner F, et al. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. PLoS One. Public Library of Science; 2012;7: e49525. doi:10.1371/journal.pone.0049525

44. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. Georges M, editor. PLoS Genet. Public Library of Science; 2008;4: e1000010. doi:10.1371/journal.pgen.1000010

45. Elferink MG, Megens H-J, Vereijken A, Hu X, Crooijmans RPMA, Groenen MAM. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PLoS One. Public Library of Science; 2012;7: e32720. doi:10.1371/journal.pone.0032720

46. Metzker ML. Sequencing technologies the next generation. Nat Rev Genet. Nature Publishing Group; 2010;11: 31–46. doi:10.1038/nrg2626

47. Tang X, Baheti S, Shameer K, Thompson KJ, Wills Q, Niu N, et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Res. Oxford University Press; 2014;42: e172. doi:10.1093/nar/gku1005

48. Oikkonen L, Lise S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome open Res. The Wellcome Trust; 2017;2: 6. doi:10.12688/wellcomeopenres.10501.2

49. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46: 912–918. doi:10.1038/ng.3036

50. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 2015;16: 195. doi:10.1186/s13059-015-0762-6

51. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. Futscher BW, editor. PLoS One. Public Library of Science; 2013;8: e58815. doi:10.1371/journal.pone.0058815

52. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. Oxford University Press; 2014;30: 2114–20. doi:10.1093/bioinformatics/btu170

53. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics. 2017;18: 80. doi:10.1186/s12859-017-1469-3

54. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. BioMed Central; 2013;14: R36. doi:10.1186/gb-2013-14-4-r36

55. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. Nature Research; 2015;12: 357–360. doi:10.1038/nmeth.3317

56. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. Oxford University Press; 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

57. Medina I, Tárraga J, Martínez H, Barrachina S, Castillo MI, Paschall J, et al. Highly sensitive and ultrafast read mapping for RNA-seq analysis. DNA Res. 2016;23: 93–100. doi:10.1093/dnares/dsv039

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Oxford University Press; 2009;25: 2078–9. doi:10.1093/bioinformatics/btp352

59. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20: 1297–303. doi:10.1101/gr.107524.110

60. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. Oxford University Press; 2010;38: e164. doi:10.1093/nar/gkq603

61. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The

Ensembl Variant Effect Predictor. Genome Biol. BioMed Central; 2016;17: 122. doi:10.1186/s13059-016-0974-4

62. Zhuo Z, Lamont SJ, Abasht B. RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. Sci Rep. Nature Publishing Group; 2017;7: 11944. doi:10.1038/s41598-017-12179-9

63. Zhuo Z, Lamont SJ, Lee WR, Abasht B. RNA-Seq Analysis of Abdominal Fat Reveals Differences between Modern Commercial Broiler Chickens with High and Low Feed Efficiencies. Brockmann GA, editor. PLoS One. Public Library of Science; 2015;10: e0135810. doi:10.1371/journal.pone.0135810

64. Zhou N, Lee WR, Abasht B. Messenger RNA sequencing and pathway analysis provide novel insights into the biological basis of chickens' feed efficiency. BMC Genomics. BioMed Central; 2015;16: 195. doi:10.1186/s12864-015-1364-0

65. Mutryn MF, Brannick EM, Fu W, Lee WR, Abasht B. Characterization of a novel chicken muscle disorder through differential gene expression and pathway analysis using RNA-sequencing. BMC Genomics. BioMed Central; 2015;16: 399. doi:10.1186/s12864-015-1623-0

66. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. BioMed Central; 2013;14: 59. doi:10.1186/1471-2164-14-59

67. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; doi:10.1186/s13756-018-0352-y

68. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

69. Yan Y, Yi G, Sun C, Qu L, Yang N. Genome-Wide Characterization of Insertion and Deletion Variation in Chicken Using Next Generation Sequencing. Wang J, editor. PLoS One. Public Library of Science; 2014;9: e104652. doi:10.1371/journal.pone.0104652

70. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. Nature Research; 2015;33: 290–295. doi:10.1038/nbt.3122

71. Kalari KR, Necela BM, Tang X, Thompson KJ, Lau M, Eckel-Passow JE, et al. An integrated model of the transcriptome of HER2-positive breast cancer. PLoS One. Public Library of Science; 2013;8: e79298. doi:10.1371/journal.pone.0079298

72. Frésard L, Leroux S, Roux P-F, Klopp C, Fabre S, Esquerré D, et al. Genome-Wide Characterization of RNA Editing in Chicken Embryos Reveals Common Features among Vertebrates. Gibas C, editor. PLoS One. Public Library of Science; 2015;10: e0126776. doi:10.1371/journal.pone.0126776

73. Kumar V, Shukla SK, Mathew J, Sharma D. Genetic Diversity and Population Structure Analysis Between Indian Red Jungle Fowl and Domestic Chicken Using Microsatellite Markers. Anim Biotechnol. Taylor & Francis; 2015;26: 201–210. doi:10.1080/10495398.2014.983645

74. Bakhtiarizadeh MR, Shafiei H, Salehi A. Large-scale RNA editing profiling in different adult chicken tissues. bioRxiv. Cold Spring Harbor Laboratory; 2018; 319871. doi:10.1101/319871

75. Paxton H, Tickle PG, Rankin JW, Codd JR, Hutchinson JR. Anatomical and biomechanical traits of broiler chickens across ontogeny. Part II. Body segment inertial properties and muscle architecture of the pelvic limb. PeerJ. PeerJ, Inc; 2014;2. doi:10.7717/PEERJ.473

76. Tickle PG, Paxton H, Rankin JW, Hutchinson JR, Codd JR. Anatomical and biomechanical traits of broiler chickens across ontogeny. Part I. Anatomy of the musculoskeletal respiratory apparatus and changes in organ size. PeerJ. PeerJ, Inc; 2014;2. doi:10.7717/PEERJ.432

77. Hedrick PW. Heterozygote Advantage: The Effect of Artificial Selection in Livestock and Pets. J Hered. Oxford University Press; 2015;106: 141–154. doi:10.1093/jhered/esu070

78. Fu W, Lee WR, Abasht B. Detection of genomic signatures of recent selection in commercial broiler chickens. BMC Genet. BioMed Central; 2016;17: 122. doi:10.1186/s12863-016-0430-1

79. Liu Z, Sun C, Qu L, Wang K, Yang N. Genome-Wide Detection of Selective Signatures in Chicken through High Density SNPs. Chaubey G, editor. PLoS One. Public Library of Science; 2016;11: e0166146. doi:10.1371/journal.pone.0166146

80. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. National Academy of Sciences; 2012;109: 19529–36.

doi:10.1073/pnas.1217149109

81. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. Public Library of Science; 2006;4: e72. doi:10.1371/journal.pbio.0040072

82. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. Cold Spring Harbor Laboratory Press; 2005;15: 1566–75. doi:10.1101/gr.4252305

83. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. 2013; doi:10.1146/annurev-genet-111212-133526

84. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. Oxford University Press; 2017;45: D635–D642. doi:10.1093/nar/gkw1104

85. Pruitt KD, Murphy TD, Thibaud-Nissen F, Kitts PA. P8007 RefSeq and Gene—NCBI resources to support comparative genomics. J Anim Sci. American Society of Animal Science; 2016;94: 183. doi:10.2527/jas2016.94supplement4183b

86. Hu Z-L, Park CA, Wu X-L, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. Oxford University Press; 2013;41: D871-9. doi:10.1093/nar/gks1150

87. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. Nature Publishing Group; 2009;4: 44–57. doi:10.1038/nprot.2008.211

88. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. Oxford University Press; 2017;45: D183–D189. doi:10.1093/nar/gkw1138

89. Zhu Y, Sun L, Garbarino A, Schmidt C, Fang J, Chen J. PathRings: a web-based tool for exploration of ortholog and expression data in biological pathways. BMC Bioinformatics. BioMed Central; 2015;16: 165. doi:10.1186/s12859-015-0585-1

90. Karlsson A-C, Fallahshahroudi A, Johnsen H, Hagenblad J, Wright D, Andersson L, et al. A domestication related mutation in the thyroid stimulating hormone receptor gene (TSHR) modulates photoperiodic response and reproduction in chickens. Gen Comp Endocrinol. Academic Press; 2016;228:

69–78. doi:10.1016/J.YGCEN.2016.02.010

91. Zhou H, Evock-Clover CM, McMurtry JP, Ashwell CM, Lamont SJ. Genome-Wide Linkage Analysis to Identify Chromosomal Regions Affecting Phenotypic Traits in the Chicken. IV. Metabolic Traits. Poult Sci. Narnia; 2007;86: 267–276. doi:10.1093/ps/86.2.267

92. Abasht B, Dekkers JCM, Lamont SJ. Review of Quantitative Trait Loci Identified in the Chicken. Poult Sci. Narnia; 2006;85: 2079–2096. doi:10.1093/ps/85.12.2079

93. Jin S, Moujahid EM El, Duan Z, Zheng J, Qu L, Xu G, et al. Association of *AMPK* subunit gene polymorphisms with growth, feed intake, and feed efficiency in meat-type chickens. Poult Sci. Narnia; 2016;95: 1492–1497. doi:10.3382/ps/pew081

94. Jo B-S, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. Korea Genome Organization; 2015;13: 112–8. doi:10.5808/GI.2015.13.4.112

95. Shahjahan M, Liu RR, Zhao GP, Zhang JJ, Zheng MQ, Li QH, et al. Polymorphisms in GJA1 and their association with growth traits in chicken. Genet Mol Res. 2015;14: 18839–18850. doi:10.4238/2015.December.28.33

96. Nishimura K, Ishiai M, Horikawa K, Fukagawa T, Takata M, Takisawa H, et al. Mcm8 and Mcm9 Form a Complex that Functions in Homologous Recombination Repair Induced by DNA Interstrand Crosslinks. Mol Cell. Elsevier; 2012;47: 511–522. doi:10.1016/J.MOLCEL.2012.05.047

97. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. BioMed Central; 2016;17: 13. doi:10.1186/s13059-016-0881-8

98. Todd E V., Black MA, Gemmell NJ. The power and promise of RNA-seq in ecology and evolution. Mol Ecol. 2016;25: 1224–1241. doi:10.1111/mec.13526

99. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. Cold Spring Harbor Laboratory Press; 2011;21: 2213–23. doi:10.1101/gr.124321.111

100. Schmdt CJ, Pritchett EM, Sun L, Davis RVNN, Hubbard A, Kniel KE, et al. RNA-seq: primary cells, cell lines and heat stress. bioRxiv. Cold Spring Harbor Laboratory; 2015; 013979. doi:10.1101/013979

101. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome

Med. BioMed Central; 2017;9: 75. doi:10.1186/s13073-017-0467-4

102. Ghosh S, Chan C-KK. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol. 2016;1374: 339–61. doi:10.1007/978-1-4939-3167-5_18

103. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. Nature Research; 2012;7: 562–578. doi:10.1038/nprot.2012.016

104. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. Nature Research; 2016;11: 1650–1667. doi:10.1038/nprot.2016.095

105. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. Oxford University Press; 2009;25: 1105–11. doi:10.1093/bioinformatics/btp120

106. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27: 2325–9. doi:10.1093/bioinformatics/btr355

107. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. Oxford University Press; 2014;30: 923–930. doi:10.1093/bioinformatics/btt656

108. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. Oxford University Press; 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

109. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. Nature Publishing Group; 2016;34: 525–527. doi:10.1038/nbt.3519

110. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. Nature Publishing Group; 2017;14: 417–419. doi:10.1038/nmeth.4197

111. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. BioMed Central; 2014;15: 550. doi:10.1186/s13059-014-0550-8

112. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Oxford University Press; 2010;26: 139–40. doi:10.1093/bioinformatics/btp616

113. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. Oxford University Press; 2011;27: 2987–93. doi:10.1093/bioinformatics/btr509

114. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics. Oxford University Press; 2015;14: 130–42. doi:10.1093/bfgp/elu035

115. Costa RL, Gadelha L, Ribeiro-Alves M, Porto F. GeNNet: an integrated platform for unifying scientific workflows and graph databases for transcriptome data analysis. PeerJ. PeerJ, Inc; 2017;5: e3509. doi:10.7717/peerj.3509

116. Elfilali A, Lair S, Verbeke C, La Rosa P, Radvanyi F, Barillot E. ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. Nucleic Acids Res. Oxford University Press; 2006;34: D613-6. doi:10.1093/nar/gkj022

117. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. Oxford University Press; 2014;42: D142–D147. doi:10.1093/nar/gkt997

118. Pimentel H, Sturmfels P, Bray N, Melsted P, Pachter L. The Lair: a resource for exploratory analysis of published RNA-Seq data. BMC Bioinformatics. BioMed Central; 2016;17: 490. doi:10.1186/s12859-016-1357-2

119. MySQL [Internet]. [cited 1 Jun 2017]. Available: http://www.mysql.com/

120. Schram A, Anderson KM. MySQL to NoSQL:Data modeling challenges in supporting scalability. Proc 3rd Annu Conf Syst Program Appl Softw Humanit - SPLASH '12. 2012; 191. doi:10.1145/2384716.2384773

121. Jatana N, Puri S, Ahuja M. A Survey and Comparison of Relational and Non-Relational Database. Int J …. ESRSA Publications; 2012;1: 1–5. Available: http://www.ijert.org/browse/august-2012-edition?download=622:a-survey-and-comparison-of-relational-and-non-relational-database

122. Wu K, Ahern S, Bethel EW, Chen J, Childs H, Cormier-Michel E, et al. FastBit: interactively searching massive data. J Phys Conf Ser. 2009;180: 012053. doi:10.1088/1742-6596/180/1/012053

123. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. Oxford University Press;

2011;27: 2156–8. doi:10.1093/bioinformatics/btr330

124. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. Oxford University Press; 2010;26: 2069–70. doi:10.1093/bioinformatics/btq330

125. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. Oxford University Press; 2016; gkw1134. doi:10.1093/nar/gkw1134

126. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. Oxford University Press; 2013;14: 178–92. doi:10.1093/bib/bbs017

127. Adetunji MO, Lamont SJ, Schmidt CJ. TransAtlasDB: an integrated database connecting expression data, metadata and variants. Database. 2018;2018. doi:10.1093/database/bay014

Appendix A

GENOMIC REGIONS IN CANDIDATE SWEEPS

Table A.1 Genomic regions identified as candidate sweeps in Ross and Illinois lines. Consecutive 10 kb sliding windows with ZH scores < -3 were merged.

Coordinates[a]	Line	Win[b]	SNPs[c]	ZH_w[d]	Het[e]	GENES[f]
chr1:1.18-1.20	Ross	1	53	-3.031	0.053	NET1, TUBAL3
chr1:3.45-3.47	Ross	1	18	-3.582	0.001	K123, MIR29A
chr1:4.10-4.12	Ross	1	44	-3.254	0.032	IL2RA, LOC419112
chr1:9.05-9.07	Ross	1	10	-3.18	0.039	SEMA3A, SEMA3E
chr1:11.38-11.40	Ross	1	10	-3.476	0.011	GNAI1
chr1:13.75-13.78	Ross	2	46	-3.116	0.025	ORC5, PUS7
chr1:18.74-18.76	Illinois	1	18	-3.073	0.007	*BRD1
chr1:30.63-30.66	Illinois	2	39	-3.024	0.009	*ENSGALG00000009638, DBX2, SCAF11
chr1:30.73-30.76	Illinois	2	29	-3.132	0.001	DBX2, SCAF11
chr1:30.83-30.88	Illinois	4	38	-3.024	0.001	*ENSGALG00000033074, DBX2, SCAF11
chr1:36.34-36.37	Illinois	2	34	-3.093	0.003	*ENSGALG00000010177, TBC1D15, TSPAN8
chr1:37.81-37.83	Illinois	1	33	-3.122	0.002	*KRR1, GLIPR1L
chr1:40.03-40.05	Ross	1	14	-3.158	0.041	*ENSGALG00000010939, MYF5
chr1:44.90-44.93	Ross	2	16	-3.127	0.002	CRADD, SOCS2
chr1:48.34-48.36	Ross	1	10	-3.074	0.049	EMP1
chr1:48.82-48.84	Illinois	1	12	-3.044	0.01	ATF7IP, MIR6581
chr1:54.55-54.58	Illinois	2	23	-3.073	0	SLC41A2, TXNRD1
chr1:54.61-54.67	Ross	3	98	-3.074	0.046	SLC41A2, TXNRD1
chr1:55.37-55.44	Ross	4	57	-3.539	0	*IGF1, *PARPBP, PMCH
chr1:55.41-55.44	Illinois	2	14	-3.132	0.001	*PARPBP, IGF1, PMCH
chr1:55.48-55.51	Ross	2	29	-3.116	0.029	*NUP37, CHPT1

chr1:57.85-57.88	Illinois	2	36	-3.044	0.009	*CHRM2, MIR490, PTN
chr1:58.72-58.74	Illinois	1	13	-3.073	0.007	C7orf73, DNM1L
chr1:60.04-60.07	Ross	2	26	-3.158	0.037	RAD52, WASHC1
chr1:60.39-60.42	Ross	2	36	-3.264	0.012	*ENSGALG00000012988, RAD52, WASHC1
chr1:64.44-64.46	Illinois	1	17	-3.073	0.007	CAPZA3, SLCO1C1
chr1:65.58-65.60	Ross	1	10	-3.19	0.038	MIR6608-1, RECQL
chr1:69.76-69.78	Ross	1	12	-3.074	0.049	PRR5
chr1:76.29-76.32	Illinois	2	16	-3.122	0.002	*ENSGALG00000031659, OVST, PHC1
chr1:76.49-76.52	Ross	2	14	-3	0.046	CD86, MIR6606
chr1:78.95-78.98	Ross	2	17	-3.021	0.054	*ENSGALG00000035228, HSD3B1, WDR3
chr1:79.97-79.99	Ross	1	10	-3.518	0.007	COX17, MIR6609
chr1:83.42-83.44	Ross	1	30	-3.021	0.054	CD200R1, CD200R1L
chr1:84.18-84.20	Ross	1	23	-3.116	0.045	ST3GAL6, TBX19
chr1:88.59-88.62	Illinois	2	16	-3.054	0.008	CD200L, MIR7446
chr1:92.43-92.46	Ross	2	23	-3.095	0.034	GJA5, GJA8
chr1:92.56-92.58	Ross	1	10	-3.275	0.03	*ENSGALG00000015494, GPR89B, POU1F1
chr1:92.85-92.87	Ross	1	24	-3.497	0.009	*POU1F1, CHMP2B
chr1:101.33-101.35	Ross	1	15	-3.233	0.034	C1H21ORF91, MIR155HG
chr1:104.95-104.97	Illinois	1	14	-3.024	0.012	*ENSGALG00000032882, C1H21ORF59, MIR1738
chr1:106.79-106.81	Ross	1	19	-3.021	0.054	*ENSGALG00000016044, CHAF1B, DYRK1A
chr1:109.20-109.23	Ross	2	14	-3.391	0.01	*ENSGALG00000016157, MX1, UBASH3A
chr1:110.26-110.28	Ross	1	15	-3.317	0.026	C1H21ORF33, PDXK
chr1:110.45-110.48	Ross	2	14	-3.317	0.02	ICOSLG, MIR222
chr1:111.11-111.14	Ross	2	32	-3	0.043	EFHC2, NDP
chr1:114.18-114.21	Illinois	2	39	-3.034	0.011	DMD, PRRG1
chr1:114.48-114.50	Ross	1	32	-3.169	0.04	DMD, PRRG1
chr1:118.54-118.57	Ross	2	17	-3.317	0.026	ACOT9, PHEX
chr1:121.38-121.40	Ross	1	22	-3.063	0.05	RBBP7, REPS2
chr1:123.28-123.31	Illinois	2	14	-3.034	0.011	EGFL6, TMSB4X
chr1:133.88-133.90	Illinois	1	23	-3.073	0.007	*IL1R1

chr1:137.47-137.49	Illinois	1	37	-3.093	0.005	*ENSGALG00000016828, LAMP1, TMCO3
chr1:163.97-164.00	Ross	2	13	-3.211	0.036	OLFM4, TDRD3
chr1:166.63-166.65	Ross	1	15	-3.116	0.045	DNAJC15, TNFSF11
chr1:167.85-167.87	Illinois	1	10	-3.044	0.01	*SPERT, LCP1
chr1:168.98-169.01	Illinois	2	18	-3.054	0.009	FNDC3A, RB1
chr1:173.38-173.41	Ross	2	14	-3.021	0.05	NBEA
chr1:175.13-175.15	Ross	1	15	-3.18	0.039	HMGB1, USPL1
chr1:187.82-187.84	Ross	1	11	-3.582	0.001	NOX4
chr1:188.84-188.86	Ross	1	10	-3.169	0.04	FZD4, MIR1664
chr2:1.21-1.23	Illinois	1	14	-3.054	0.009	ADCYAP1R1, GHRHR
chr2:1.74-1.77	Illinois	2	10	-3.015	0.013	VIPR1
chr2:6.41-6.45	Ross	3	49	-3.19	0.01	*PRKAG2
chr2:16.77-16.79	Ross	1	16	-3.063	0.05	PRTFDC1
chr2:20.06-20.09	Ross	2	15	-3.243	0.013	RSU1
chr2:20.72-20.74	Illinois	1	16	-3.005	0.014	ABCB1, RPP38
chr2:21.90-21.93	Illinois	2	21	-3.093	0.004	*ENSGALG00000009062, FZD1, MIR466
chr2:22.94-22.96	Illinois	1	20	-3.132	0.001	*ENSGALG00000009479, CDK6, VPS50
chr2:31.18-31.20	Ross	1	29	-3.148	0.042	*ENSGALG00000010949, IGF2BP3, NUPL2
chr2:34.23-34.26	Ross	2	19	-3.158	0.041	BTD, DPH3
chr2:45.31-45.34	Ross	2	17	-3.18	0.039	MIR1607, PDCD6IP
chr2:46.03-46.05	Ross	1	16	-3.317	0.026	ELMO1, MIR128-2
chr2:48.21-48.23	Illinois	1	31	-3.093	0.005	FKBP9, LSM5
chr2:50.00-50.02	Ross	1	11	-3.031	0.053	INHBA, VPS41
chr2:50.22-50.25	Ross	2	24	-3.444	0.014	INHBA, VPS41
chr2:52.29-52.31	Ross	1	11	-3.486	0.01	SEC61G
chr2:53.21-53.23	Ross	1	13	-3.412	0.017	PDIA4, SEC61G
chr2:53.29-53.32	Illinois	2	17	-3.141	0	PDIA4, SEC61G
chr2:53.50-53.53	Ross	2	13	-3.296	0.028	PDIA4, SEC61G
chr2:59.14-59.16	Ross	1	14	-3.243	0.033	MBOAT1, SOX4
chr2:60.32-60.35	Ross	2	29	-3.391	0.012	DEK, ID4
chr2:61.07-61.09	Ross	1	34	-3.127	0.044	JARID2, SIRT5
chr2:63.85-63.89	Illinois	3	41	-3.073	0.003	BLOC1S5, TFAP2A

chr2:65.11-65.13	Ross	1	32	-3.285	0.029	LY86
chr2:65.79-65.81	Ross	1	23	-3.031	0.053	LYRM4, TUBB2B
chr2:66.22-66.24	Ross	1	22	-3.349	0.023	LYRM4, TUBB2B
chr2:68.00-68.02	Ross	1	10	-3.222	0.035	*ENSGALG00000012873, OVALX, VPS4B
chr2:79.40-79.43	Ross	2	19	-3.412	0.016	CCT5, PAPD7
chr2:80.22-80.25	Ross	2	18	-3.349	0.023	NSUN2, ZPBP
chr2:84.49-84.51	Illinois	1	36	-3.141	0	*ENSGALG00000013138, BAG1, GALNT1, MIR32
chr2:84.58-84.62	Illinois	3	28	-3.005	0.002	BAG1, MIR32
chr2:84.61-84.63	Ross	1	18	-3.031	0.053	BAG1, MIR32
chr2:88.12-88.14	Illinois	1	23	-3.122	0.002	*ENSGALG00000013196, IRX1, MIR1816
chr2:90.21-90.23	Illinois	1	20	-3.141	0	CARMIL1
chr2:93.97-93.99	Illinois	1	26	-3.015	0.013	*RTTN
chr2:95.00-95.03	Ross	2	15	-3.582	0.001	CDH19, TMX3
chr2:96.75-96.77	Ross	1	15	-3.476	0.011	MC2R, MC5R
chr2:99.00-99.02	Ross	1	11	-3.021	0.054	*ENSGALG00000038316, LAMA1, TWSG1
chr2:103.53-103.55	Ross	1	21	-3.211	0.036	*ENSGALG00000015064, GATA6, MIR1597
chr2:105.33-105.35	Illinois	1	12	-3.034	0.011	CDH2, TTR
chr2:115.25-115.27	Ross	1	27	-3.106	0.046	*ENSGALG00000038395, MTFR1, TRIM55
chr2:115.92-115.94	Ross	1	11	-3.275	0.03	CPA6
chr2:116.39-116.42	Ross	2	13	-3.095	0.036	CPA6, MIR1569-2
chr2:127.81-127.84	Ross	2	43	-3.148	0.008	*ENSGALG00000034050, CPQ, RPL30
chr2:130.54-130.56	Ross	1	23	-3.412	0.017	*ENSGALG00000033520, DCAF13, DPYS
chr2:136.08-136.10	Illinois	1	17	-3.015	0.013	MIR1467-2, TNFRSF11B
chr2:136.52-136.54	Ross	1	14	-3.222	0.035	MAL2, NOV
chr2:139.36-139.39	Ross	2	17	-3.486	0.009	MYC, NSMCE2
chr2:141.45-141.48	Illinois	2	32	-3.112	0.003	LRRC6, MYC
chr2:142.05-142.07	Ross	1	23	-3.01	0.055	*WISP1
chr2:149.22-149.24	Illinois	1	14	-3.024	0.012	*ENSGALG00000039346, FK1L, PUF60
chr3:6.17-6.20	Illinois	2	11	-3.005	0.012	NRXN1, OTOR
chr3:8.48-8.50	Illinois	1	22	-3.054	0.009	FAM179A, YPEL5

chr3:9.03-9.05	Ross	1	11	-3.529	0.006	B3GNT2, PPP1R21
chr3:15.06-15.08	Ross	1	12	-3.063	0.05	BMP2, MIR1756B
chr3:16.46-16.48	Illinois	1	10	-3.063	0.008	EIF4A3, HNRNPLL
chr3:19.15-19.17	Ross	1	24	-3	0.056	LYPLAL1
chr3:19.96-19.99	Ross	2	26	-3.148	0.038	ESRRG, GPATCH2
chr3:24.06-24.09	Ross	2	43	-3.264	0.029	COX7A2L, ZFP36L2
chr3:24.11-24.14	Illinois	2	14	-3.083	0.006	COX7A2L, ZFP36L2
chr3:25.41-25.43	Ross	1	11	-3.158	0.041	*ENSGALG00000009967, PPM1B, THADA
chr3:29.13-29.16	Ross	2	22	-3.486	0.009	*ENSGALG00000010057, GLP1R, LOC421419
chr3:32.56-32.58	Illinois	1	12	-3.015	0.013	CEBPZ, CRIM1
chr3:37.34-37.36	Illinois	1	14	-3.122	0.002	*ENSGALG00000010812, CHRM3, MIR135B
chr3:38.05-38.08	Ross	2	14	-3.031	0.053	GPR137B, NTPCR
chr3:40.00-40.02	Ross	1	11	-3	0.056	ACTA1, EXOC8
chr3:40.24-40.26	Illinois	1	17	-3.005	0.014	*ENSGALG00000011101, ACTA1, PDCD2
chr3:40.44-40.46	Illinois	1	15	-3.015	0.013	*ENSGALG00000011111, ACTA1, PDCD2
chr3:45.27-45.30	Illinois	2	15	-3.122	0.002	*PRKN, IGF2R, QKI
chr3:45.39-45.42	Illinois	2	26	-3.044	0.005	*ENSGALG00000020003, IGF2R, QKI
chr3:46.33-46.36	Ross	2	11	-3.063	0.05	*ENSGALG00000012256, MIR1734, SF3B5
chr3:46.63-46.65	Illinois	1	23	-3.044	0.01	EPM2A
chr3:49.48-49.50	Ross	1	10	-3.18	0.039	*ENSGALG00000013505, *ENSGALG00000042638, ESR1, VIP
chr3:49.52-49.54	Ross	1	11	-3.486	0.01	*ENSGALG00000042638, ESR1, VIP
chr3:50.37-50.40	Ross	2	37	-3.169	0.037	RGS17, TIAM2
chr3:50.43-50.46	Ross	2	23	-3.412	0.014	RGS17, TIAM2
chr3:54.81-54.84	Ross	2	22	-3.19	0.032	IFNGR1, PERP1
chr3:55.98-56.00	Ross	1	15	-3.296	0.028	*ENSGALG00000013962, MYB, SGK1
chr3:59.12-59.15	Illinois	2	14	-3.093	0.005	ECHDC1, MIR1660
chr3:59.21-59.25	Illinois	3	23	-3.054	0.001	ECHDC1, MIR1660
chr3:60.05-60.07	Illinois	1	33	-3.132	0.001	HDDC2, NCOA7

chr3:61.07-61.11	Illinois	3	82	-3.083	0.003	*ENSGALG00000014848, FABP7, NKAIN2
chr3:62.42-62.48	Illinois	5	57	-3.102	0	GJA1, MCM9
chr3:62.72-62.75	Illinois	2	21	-3.024	0.012	GJA1, MCM9
chr3:62.72-62.75	Ross	2	21	-3.031	0.051	GJA1, MCM9
chr3:64.19-64.23	Ross	3	28	-3.254	0.02	*FRK, FAM26E, HDAC2
chr3:67.27-67.29	Illinois	1	15	-3.093	0.005	FIG4, MIR6699
chr3:67.51-67.53	Illinois	1	12	-3.005	0.014	FIG4, MIR6699
chr3:69.70-69.73	Ross	2	11	-3.518	0.007	ASCC3, HACE1
chr3:75.53-75.56	Ross	2	12	-3.095	0.047	EPHA7, LYRM2
chr3:76.69-76.71	Illinois	1	19	-3.083	0.006	*RARS2, ORC3
chr3:82.67-82.69	Ross	1	15	-3.19	0.038	*ENSGALG00000015951, B3GAT2, MIR30C2
chr3:88.47-88.49	Illinois	1	10	-3.063	0.008	*ENSGALG00000046637, ELOVL5, TINAG
chr3:92.38-92.41	Ross	2	25	-3.116	0.004	ACP1, TMEM18
chr3:93.90-93.93	Ross	2	11	-3.053	0.047	TMEM18, TSSC1
chr3:98.07-98.10	Ross	2	27	-3.349	0.023	E2F6, TRIB2
chr3:98.47-98.50	Ross	2	12	-3.19	0.038	DDX1, TRIB2
chr3:111.28-111.30	Ross	1	12	-3.391	0.019	PTCHD4
chr4:12.59-12.61	Ross	1	10	-3.095	0.047	MIR1573, UPRT
chr4:23.51-23.54	Ross	2	11	-3.275	0.03	ETFDH, NPY1R
chr4:24.00-24.02	Ross	1	16	-3.465	0.012	KLHL2
chr4:26.75-26.77	Ross	1	10	-3.349	0.023	LOC422426, PCDH10
chr4:31.63-31.65	Illinois	1	11	-3.044	0.01	LSM6, SLC10A7
chr4:32.16-32.18	Ross	1	30	-3.423	0.016	EDNRA
chr4:35.76-35.79	Illinois	2	16	-3.122	0.002	*ENSGALG00000010391, HPGDS, SNCA
chr4:36.24-36.27	Ross	2	17	-3.307	0.026	HPGDS, SNCA
chr4:37.15-37.17	Ross	1	28	-3.539	0.005	HPGDS, SNCA
chr4:47.50-47.52	Ross	1	14	-3.084	0.048	GPAT3, MIR1730
chr4:51.26-51.28	Ross	1	20	-3.053	0.051	*IL8L1
chr4:51.91-51.93	Ross	1	23	-3.01	0.055	CENPC
chr4:61.08-61.10	Illinois	1	30	-3.005	0.014	H2AFZ, NFKB1
chr4:63.49-63.51	Illinois	1	52	-3.112	0.003	CNOT7, MIR1605
chr4:66.63-66.65	Ross	1	11	-3.074	0.049	SGCB, TEC

chr4:69.49-69.51	Ross	1	13	-3.296	0.028	MIR6586, PDS5A
chr4:70.37-70.39	Ross	1	13	-3.031	0.053	*ENSGALG00000014337, RELL1
chr4:73.57-73.60	Ross	2	22	-3.444	0.014	CCKAR
chr4:78.60-78.62	Ross	1	15	-3.169	0.04	RAB28, WDR1
chr4:83.66-83.69	Ross	2	27	-3.497	0.005	LOC422894, MXD4
chr4:89.61-89.63	Ross	1	17	-3.106	0.046	HTR7L, MIR1684
chr4:90.02-90.04	Ross	1	16	-3.264	0.031	ADAM33, GFRA4
chr5:5.23-5.25	Ross	1	11	-3.264	0.031	PAX6, WT1
chr5:5.45-5.47	Illinois	1	11	-3.024	0.012	*WT1
chr5:8.84-8.87	Ross	2	17	-3.243	0.032	*ENSGALG00000005632, EIF4G2, LYVE1
chr5:11.87-11.89	Ross	1	31	-3.042	0.052	*ENSGALG00000006172, FTL, NUCB2
chr5:11.91-11.93	Ross	1	20	-3.053	0.051	*ENSGALG00000006172, FTL, NUCB2
chr5:12.10-12.13	Ross	2	36	-3.476	0.009	FTL, NUCB2
chr5:13.65-13.67	Illinois	1	11	-3.015	0.013	CD81, MIR6642
chr5:14.29-14.31	Illinois	1	33	-3.024	0.012	BRSK2, CTSD
chr5:21.21-21.23	Illinois	1	11	-3.132	0.001	API5, C5H11orf74
chr5:24.14-24.16	Ross	1	11	-3.211	0.036	SLC35C1, ZFYVE19
chr5:28.20-28.22	Ross	1	10	-3.338	0.024	ACTN1, EXD2
chr5:29.29-29.32	Ross	2	11	-3.561	0.003	GPHN
chr5:29.34-29.37	Ross	2	63	-3.169	0.028	BMF, GPHN
chr5:29.84-29.87	Ross	2	23	-3.158	0.026	KATNBL1, THBS1
chr5:34.25-34.27	Illinois	1	13	-3.054	0.009	COCH
chr5:34.57-34.60	Ross	2	34	-3.222	0.035	*ENSGALG00000009983, COCH, SPTSSA
chr5:43.98-44.00	Ross	1	14	-3.465	0.012	CALM1, TTC7B
chr5:48.59-48.61	Illinois	1	17	-3.132	0.001	YY2
chr5:50.34-50.36	Ross	1	10	-3.137	0.043	*ENSGALG00000011505, CKB, EIF5
chr5:54.32-54.34	Ross	1	10	-3.455	0.013	*ENSGALG00000011893, HIF1A, SIX1
chr6:0.69-0.72	Ross	2	14	-3.275	0.03	BICC1, PHYHIPL
chr6:6.60-6.62	Ross	1	30	-3.095	0.047	* PCDH15
chr6:9.52-9.54	Ross	1	20	-3.529	0.006	MINPP1, RNLS
chr6:9.60-9.62	Ross	1	16	-3.222	0.035	MINPP1, RNLS

chr6:11.23-11.25	Ross	1	13	-3.095	0.047	*ENSGALG00000004345, DNAJB12, P4HA1
chr6:17.19-17.21	Ross	1	10	-3.434	0.015	MIR6577, PAX2
chr6:17.45-17.47	Illinois	1	10	-3.122	0.002	NDUFB8, SCD
chr6:19.27-19.30	Ross	2	18	-3.042	0.052	CH25H, FAS
chr6:20.32-20.34	Illinois	1	13	-3.054	0.009	LGI1, PDE6C
chr6:20.32-20.34	Ross	1	19	-3.116	0.045	LGI1, PDE6C
chr6:20.58-20.60	Ross	1	10	-3.095	0.047	EXOC6
chr6:20.79-20.83	Illinois	3	96	-3.112	0.002	*ENSGALG00000038924, *MARCH5, KIF11
chr6:21.15-21.17	Ross	1	12	-3.476	0.011	BLNK, DNTT
chr6:28.51-28.53	Ross	1	22	-3.243	0.033	GFRA1, MIR1815
chr6:30.39-30.41	Illinois	1	21	-3.024	0.012	*ENSGALG00000009466, *MCMBP, FGFR2
chr6:31.87-31.89	Ross	1	10	-3.243	0.033	BUB3, OAT
chr7:5.80-5.82	Illinois	1	43	-3.024	0.012	*ENSGALG00000004196, *ENSGALG00000031157, MIR6691-1, USP40
chr7:6.13-6.15	Ross	1	13	-3.285	0.029	TWIST2
chr7:12.20-12.22	Ross	1	15	-3.084	0.048	FZD5, METTL21A
chr7:17.10-17.13	Ross	2	16	-3.031	0.05	OLA1
chr7:17.35-17.38	Ross	2	11	-3.317	0.026	PPP1R9B, SP3
chr7:17.96-17.98	Illinois	1	17	-3.083	0.006	DLX1, HAT1
chr7:18.64-18.66	Ross	1	16	-3.444	0.014	*ENSGALG00000035605, SP5, SSB
chr7:21.75-21.77	Illinois	1	13	-3.015	0.013	RBMS1
chr7:29.04-29.06	Ross	1	27	-3.116	0.045	*CCDC93
chr7:36.30-36.32	Ross	1	33	-3.031	0.053	KCNJ3, MIR6546
chr7:36.42-36.44	Ross	1	27	-3.412	0.017	*ENSGALG00000041257, KCNJ3, MIR6546
chr7:36.55-36.57	Ross	1	11	-3.434	0.015	ACVR1, PKP4
chr8:1.26-1.31	Ross	2	91	-3.158	0.033	*ENSGALG00000001983, *STXBP3, NR5A2
chr8:1.42-1.44	Ross	1	11	-3.233	0.034	*ENSGALG00000002145, NR5A2, STXBP3
chr8:1.78-1.80	Ross	1	11	-3.158	0.041	MIR181A1, NR5A2
chr8:3.75-3.77	Ross	1	14	-3.455	0.013	HSD17B7, RGS4
chr8:14.33-14.35	Ross	1	10	-3.338	0.024	C8H1ORF146, TGFBR3
chr8:15.73-15.76	Illinois	2	20	-3.044	0.008	LMO4, ZNF326

chr8:16.77-16.79	Ross	1	13	-3.031	0.053	MIR1761, SAMD13
chr8:28.44-28.46	Illinois	1	18	-3.122	0.002	*LEPR
chr9:8.36-8.38	Illinois	1	21	-3.073	0.007	MRPL44, SERPINE2
chr9:8.85-8.88	Ross	2	12	-3.211	0.03	MRPL44, SERPINE2
chr9:10.55-10.57	Illinois	1	38	-3.034	0.011	*TRPC1, MIR1458
chr9:10.79-10.81	Ross	1	11	-3.529	0.006	SLC9A9
chr9:11.80-11.82	Illinois	1	10	-3.132	0.001	SLC9A9, ZIC1
chr9:12.11-12.13	Ross	1	19	-3.148	0.042	MIR6611, ZIC1
chr9:12.42-12.45	Ross	2	45	-3.031	0.045	*AGTR1, NCBP2
chr9:15.26-15.28	Ross	1	19	-3.338	0.024	*ENSGALG00000007691, NCL, PSMD1
chr9:16.36-16.38	Ross	1	12	-3.264	0.031	B3GNT5, LAMP3
chr9:17.96-17.98	Ross	1	10	-3.243	0.033	PIK3CA, TBL1XR1
chr9:20.39-20.42	Illinois	2	13	-3.122	0.002	*ENSGALG00000009458, MIR551B, TERC
chr10:1.75-1.77	Ross	1	12	-3.285	0.029	ADPGK, MIR1623
chr10:4.06-4.08	Ross	1	19	-3.095	0.047	*ENSGALG00000003487, IDH3A, RPS27L
chr10:8.70-8.74	Illinois	3	92	-3.063	0.004	*ARPP19, *ENSGALG00000042839, *FAM214A, MYO5A
chr10:14.93-14.96	Ross	2	15	-3.402	0.018	NR2F2, RGMA
chr10:17.01-17.04	Ross	2	19	-3.476	0.009	MEF2A
chr10:19.72-19.75	Illinois	2	34	-3.132	0.001	FAM96A, MORF4L1
chr10:20.24-20.27	Ross	2	14	-3.434	0.014	*ENSGALG00000008425, CKMT1A
chr11:0.37-0.39	Ross	1	25	-3.211	0.036	SLC7A6OS, TLR21
chr11:2.42-2.44	Ross	1	11	-3	0.056	CBFB
chr11:4.28-4.30	Ross	1	17	-3.423	0.016	FTO, MMP2
chr11:11.45-11.47	Ross	1	14	-3.021	0.054	*ENSGALG00000033720, *NAE1, RRAD
chr11:14.04-14.06	Ross	1	11	-3.296	0.028	MIR6595, NUDT7
chr11:18.56-18.59	Ross	2	42	-3.529	0.006	*ENSGALG00000026534, BANP, CIDEC
chr11:19.05-19.08	Illinois	2	29	-3.122	0.001	*ENSGALG00000000521, *TCF25, FANCA, MC1R
chr12:5.20-5.22	Ross	1	11	-3.359	0.022	*ENSGALG00000044299, ACAD9, RAB43
chr12:16.50-16.52	Ross	1	21	-3.042	0.052	*ENSGALG00000007798, MIR1711, PPP4R2

chr13:6.15-6.18	Ross	2	18	-3.497	0.009	GABRG2, TENM2
chr13:7.77-7.79	Ross	1	20	-3.455	0.013	GABRB2, MIR146A
chr13:17.36-17.38	Ross	1	14	-3.169	0.04	ARHGAP26, FGF1
chr14:9.14-9.16	Ross	1	12	-3.201	0.037	*ENSGALG00000007216, DEXI, NUBP1
chr14:10.08-10.10	Ross	1	10	-3.307	0.027	*ENSGALG00000042981, PMM2, USP7
chr14:14.39-14.41	Ross	1	12	-3.285	0.029	*MCHR2
chr15:8.24-8.26	Illinois	1	21	-3.132	0.001	DDT, TBX6
chr17:8.48-8.50	Illinois	1	11	-3.015	0.013	LHX3, PMPCA
chr18:0.50-0.52	Ross	1	11	-3.412	0.017	MYH1A, MYH1G
chr18:0.80-0.82	Ross	1	14	-3.106	0.046	*ENSGALG00000032178, MIR1748, MYH1G, MYOCD
chr18:2.14-2.16	Ross	1	10	-3.296	0.028	STX8
chr18:3.84-3.86	Ross	1	14	-3.201	0.037	SEPT9
chr19:0.58-0.60	Ross	1	23	-3.042	0.052	*ENSGALG00000001031, MIR1698-2, SUPT4H1
chr20:10.48-10.50	Ross	1	24	-3.486	0.01	*BPIFB3, BPIFB2, MAPRE1
chr20:11.49-11.51	Ross	1	10	-3.031	0.053	PMEPA1, RAB22A
chr20:11.65-11.67	Ross	1	14	-3.55	0.004	PMEPA1, RAB22A
chr21:1.67-1.69	Illinois	1	15	-3.122	0.002	SKI
chr21:5.85-5.88	Ross	2	19	-3.455	0.013	*ENSGALG00000026658, EPHB2, GUCA2A
chr24:6.13-6.15	Illinois	1	53	-3.054	0.009	*IL18, *SDHD, HSPB2, NCAM1
chr25:2.43-2.46	Ross	2	14	-3.158	0.041	CRP, FCRL2
chr26:1.74-1.76	Ross	1	12	-3.328	0.025	*ENSGALG00000038399, MIR6618, SNRPE
chr26:1.94-1.96	Ross	1	26	-3.201	0.037	*NFASC
chr28:4.23-4.25	Ross	1	10	-3.307	0.027	*ENSGALG00000003742, MYO9B, TMEM38A
chr33:0.04-0.06	Ross	1	11	-3.402	0.018	MIR1668

Note: [a] Coordinates of region in megabases (Mb). [b] Number of consecutive $ZH_w < -3$ windows that were merged. [c] Number of SNPs identified in region. [d] The lowest ZH_w observed for a 10kb window in the region. [e] The lowest heterozygosity (H_w) observed for a 10kb window in the region. [f] Gene(s) overlapping putative sweep regions: *= SNP(s) impact the coding regions of the gene.

Appendix B

PERMISSIONS

Chapter 4 [127] was published in Oxford Database (Figure B.1). At the time of submission of this dissertation, Chapter 3 was under review with PLOS ONE (https://www.plos.org/editorial-publishing-policies). These journals do not require that authors receive permission before publishing the author's articles in their own dissertations. The research described in this dissertation if published in any journal, would be the property of the respective journal or press.

Figure B.1 Open access license for Chapter 4.