# Utilization and Discovery of Single Nucleotide Variations for Current and Future Genome-wide Applications

Scott Funkhouser

Michigan State University
Genetics Graduate Program
November 30th, 2015

Thesis Committee:
Catherine Ernst (Mentor and Program Chair)
Juan Pedro Steibel
Ron Bates
David Arnosti
Gustavo de los Campos
Yuehua Cui (Genetics Program Representative)

## ABSTRACT

Genome-wide sequence variation such as single nucleotide polymorphisms (SNPs) have been successfully utilized toward inferring genetic components of traits or diseases. Likewise, prediction of phenotypes from genomic data has been tested in human and livestock species with promising results. Our success with either inference or prediction is theoretically dependent on a number of biological factors, which if unaccounted for, can dampen our ability to detect quantitative trait loci (QTL), determine the observed heritability of a trait, or successfully predict a phenotypic outcome. These factors are especially apparent when the goal is to predict phenotypes among heterogeneous populations in which linkage disequilibrium between observed markers and underlying QTL, marker allele frequencies, and QTL effects can differ from subpopulation to subpopulation. Furthermore, recent advances in high throughput sequencing of transcriptomes have revealed that single nucleotide variants (SNVs) can exist within the transcriptome without existing in the genome, thus adding potential explanatory variables that go unrecognized if one simply uses sequence variation observed at the level of the genome to explain phenotypic variation. This is made possible by a mechanism known as RNA editing, in which the eukaryotic cell intentionally makes single nucleotide changes within premature RNA transcripts. For my thesis proposal, I will build and assess new tools and methodology to enhance genomic prediction among heterogeneous populations as well as characterize RNA editing patterns genome-wide to better understand the evolution of this phenomenon and the extent that it provides additional sources of genetic variation with which to utilize in genome-wide studies.

## TABLE OF CONTENTS

## SPECIFIC AIMS

I propose to establish new statistical methods to enhance predictive capacity of heritable traits in heterogeneous and admixed populations by incorporating estimates of ancestral composition with existing sources of genomic sequence variation. In addition, using bioinformatic approaches I aim to reveal new sources of genomic sequence variation in the form of RNA editing loci and explore the patterns and consistency of this phenomenon between individuals, tissues, and across species.

***Aim 1****) Develop and implement a new approach for genomic prediction, termed the "continuous model", to account for complex genetic heterogeneity.*

> ***1a****) Determine how prediction accuracy from the continuous model compares to prediction accuracy from previously developed "stratified" and "interaction" models.*
>
> ***1b)*** *Develop the means to perform hypothesis testing with the continuous model in order to infer significance of SNP effects.*

***Aim 2****) Investigate the function and impact of RNA editing throughout mammalian evolution and build novel tools for the detection, analysis, visualization of RNA editing data.*

> ***2a****) Build software specifically designed for the detection, analysis, and visualization of RNA editing data from multiple high-throughput technologies.*
>
> ***2b****) Test software using swine as a model for RNA editing. Long read and short read sequencing technologies will be compared for their ability to detect RNA editing events, and a subset of RNA editing candidates will be validated to assess detection accuracy.*
>
> ***2c)*** *Perform a comparative study in which RNA editing patterns are compared across a number of mammalian species harboring different genomic features.*

## BACKGROUND AND SIGNIFICANCE

### *Aim 1)*

*Using additive genetic models for prediction*

According to the classic quantitative genetics model, the regression of a quantitative phenotype onto allele content (dosage of a particular allele) provides a means to quantify the average effect of alleles (Falconer and McKay 1996). The sum of average effects for each allele at a particular locus is an estimate of the additive value or breeding value of the genotype. Additive values, suggested by their name, provide a best estimate of how a phenotype will be impacted by a particular genotype, assuming the alleles at the genotype act additively. This assumption has potential to work well within animal genetics, as selection of animals for breeding based on the sum of their additive values is projected to result in a 8-38% extra genetic gain beyond simply using pedigree information to inform selection decisions (Meuwissen and Goddard 1996). For certain traits, and for other applications including the genomic prediction of complex human disease, the additive model may not be sufficient because it ignores effects due to dominance (interactions between alleles at the same locus) and epistasis (interactions between loci) (Huang et al. 2012). However, the additive model continues to be useful for genomic selection and for the estimation of narrow-sense heritability of complex human traits (Yang et al. 2010).

The development of dense marker maps in various animal and plant species containing positions of single nucleotide polymorphisms (SNPs) prompted the application of the additive quantitative genetics model to many ( > 10,000) markers simultaneously (Meuwissen, Hayes and Goddard 2001). This approach can be represented with

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ represents the data vector of phenotypic measurements for each of the $n$ animals/observations, $\mathbf{1}\mu$ is a unit vector multiplied by the phenotypic mean, $\boldsymbol{X}$ is the $n \times m$ incidence matrix, with each of the $m$ markers represented as the dosage of a particular allele {0, 1 or 2}, $\boldsymbol{b}$ is a vector containing average effects for each SNP, and $\boldsymbol{\varepsilon}$ is a vector containing errors for each observation. Both $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ can be assumed to come from random distributions $N(0, \mathbf{I}\sigma_{\boldsymbol{b}}{}^2)$ and $N(0, \mathbf{I}\sigma_{\varepsilon}{}^2)$, respectively. Alternatively, an equivalent model can be expressed as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are the same as before, $\boldsymbol{Z}$ denotes an incidence matrix connecting additive values to phenotypes and $\boldsymbol{u}$ represents a vector of additive effects or breeding values, assumed to be a draw from the random distribution $N(0, \mathbf{G}\sigma_u{}^2)$, where $\mathbf{G}$ represents the genomic relationship matrix and is proportional to $\boldsymbol{X}\boldsymbol{X}'$, providing a measurement of genetic relatedness between all individuals contained in $\boldsymbol{X}$.

This model is appealing in that it is both simple and effective (provided $n$ is sufficiently large) at providing estimates of all SNP effects simultaneously by borrowing information for all individuals contained in $\boldsymbol{X}$. This means that if individuals in $\boldsymbol{X}$ come from differing subpopulations, our estimate of $\boldsymbol{b}$ could be confounded by the fact that estimated SNP effects may be different between subpopulations. Nevertheless, utilizing a heterogeneous training population to fit a model such as this consisting of both Jersey and Holstein cattle has shown success in enhancing genomic prediction of breeding values for purebred Jersey animals by up to 13% when compared to training the model on Jersey animals alone (Hayes et al. 2009).

*Challenges associated with SNP effect estimation among complex substructure*

When conducting genome-wide regressions, our parameters of interest - the so-called "SNP effects", are most-often not a product of the observed SNP itself. Instead, estimated SNP effects are a byproduct of the SNP being in linkage disequilibrium (LD) with the true quantitative trait loci (QTL). As in the cattle study, in instances where the training of a whole genome regression model using a heterogeneous population results in a gain in accuracy of SNP effects, it can be assumed that major QTL were in LD with the same SNP for both Jerseys and Holsteins, and that those QTL effects were the same between Jerseys and Holsteins. In practice, QTL will not always be in LD with the same SNP across subpopulations. Furthermore, the true underlying QTL effects may not be the same across groups due to the effects of dominance, epistasis, or potential epigenetic mechanisms that are group or environment specific.

In the context of genome-wide association studies (GWAS), in which the goal is to find SNPs significantly associated with a trait or disease, population stratification is considered a confounder that can be partly accounted for by using the top principal components (PCs) from the genomic data (Price et al 2006). While accounting for the top PCs can reduce the number of spurious genetic associations for the phenotype of interest, it is not necessarily satisfactory for improving the estimation of SNP effects that can be used for prediction. Principal components resulting from the vector decomposition of **G** have been included as fixed effects in whole genome regression models aimed at estimating breeding values among multi-

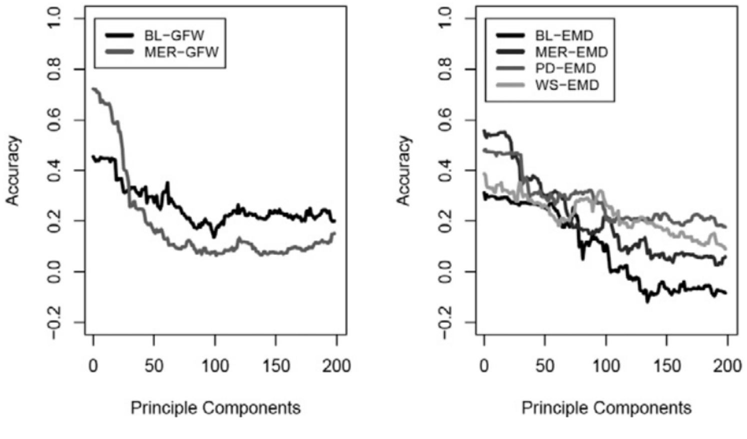

*Figure 1. The change in across breed genomic predictions (in which the breed in question was removed from a heterogenous reference panel, then predicted) for greasy fleece weight (GFW) and eye muscle depth (EMD) as an increasing number of PCs derived from G are included as fixed effects. BL – Border Leicester, MER – Merino, PD – Polled Dorset, WS – White Suffolk (Daetwyler et al 2012).*

breed sheep (Daetwyler et al 2012). As more principal components are included, across breed prediction accuracy was shown to decrease (Figure 1).

Rather than thinking of population substructure as a confounding variable that needs accounting for, genomic prediction can theoretically use substructure to decompose the SNP effect for site $j$ into those common to all groups $b_{0j}$, plus a group specific effect $b_{gj}$ (de los Campos et al. 2015). This so-called "*interaction model*" assuming the presence of 2 groups can be expressed as

$$\begin{bmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{1}\mu_1\\ \mathbf{1}\mu_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix}\boldsymbol{b}_0 + \begin{bmatrix}\boldsymbol{X}_1\\ \mathbf{0}\end{bmatrix}\boldsymbol{b}_1 + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{X}_2\end{bmatrix}\boldsymbol{b}_2 + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}, \tag{3}$$

which contains the same terms as (1), only the subscript $i$ denotes the group. For example, QTL could have the same effect between groups ($b_{0j}$) but the extent of LD between QTL and SNP marker $j$ could differ between groups, thus leading to a group specific SNP effect ($b_{gj}$). Together, $b_{0j} + b_{gj} = \beta_{gj}$, can provide the total SNP effect for marker $j$ for an individual belonging to group $g$. Alternatively, we can assume that there is no common component of the SNP effect between groups ($b_0 = 0$) using a so called "*stratified model*". This can be represented with

$$\begin{bmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{1}\mu_1\\ \mathbf{1}\mu_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{X}_1\\ \mathbf{0}\end{bmatrix}\boldsymbol{b}_1 + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{X}_2\end{bmatrix}\boldsymbol{b}_2 + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}. \tag{4}$$

These models provide a way to estimate SNP effects among a heterogeneous population with clearly defined clusters, but the question of how to best model heterogeneity containing complex substructure such as admixed individuals (hybrids or crossbred animals descending from 2 or more clusters) remains a question. Conceivably, the first approach to modeling admixed individuals could be to first quantify their ancestral composition, however whole genome regression models have not incorporated ancestral composition to enhance the accuracy of SNP effects.

### *Ancestral composition estimation*

Methods to estimate breed or ancestral composition have been devised that use group allele frequencies (Kuehn et al. 2011). In bulls, the following regression model was used

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{c} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{y}$ is a vector of a test bull's genotypes divided by 2 {0, 0.5, or 1}, $\boldsymbol{X}$ is a matrix of allele frequencies, with each row representing a different SNP, and columns representing a different purebred breed or group. $\boldsymbol{c}$ is a vector containing fixed effects for each group and $\boldsymbol{\varepsilon}$ is a vector of errors for each SNP. The vector $\boldsymbol{c}$ can represent relative composition estimates for the animal whose genotypes are represented in $\boldsymbol{y}$, however assuming this model is fit using an unbiased estimator such as ordinary least squares (OLS), the interpretation of each element of $\boldsymbol{c}$ may not be clear in cases where $b_j$ is less than 0 or greater than 1. In order to allow for the elements of $\boldsymbol{c}$ to represent group compositions or group proportions, the OLS equations can be solved using quadratic programming in such a way to force each element of $\boldsymbol{c}$ to be between 0 and 1 and all elements to sum to 1.

***Aim 2)***
*RNA Editing – A source of hidden sequence variation*

Hypothetically, for any given animal species, even the densest SNP panel can fail to “tag” all sequence variants by way of linkage disequilibrium, no matter how large the sample size. One reason for this is that mechanisms such as RNA editing permit the cell to make changes to RNA sequences without having to change the “hard wired” genomic sequence (Benne et al. 1986). In this way, the cell can make “soft” changes to the nucleotide sequence of transcribed genes without having to rely on permanent genomic mutations. These mechanisms result in single nucleotide variants (SNVs) that are unique to the transcriptome. Although current approaches for genomic prediction are not concerned with transcriptome-specific SNVs such as those caused by RNA editing, future efforts to combine genomic, transcriptomic, and epigenetic data for the purpose of predicting traits or diseases are increasingly immanent. Our ability to appropriately merge RNA editing data with other “omics” data types for the purpose of inference or prediction is dependent on our ability to both accurately identify RNA editing sites via high-throughput methods and understand their underlying functionality.

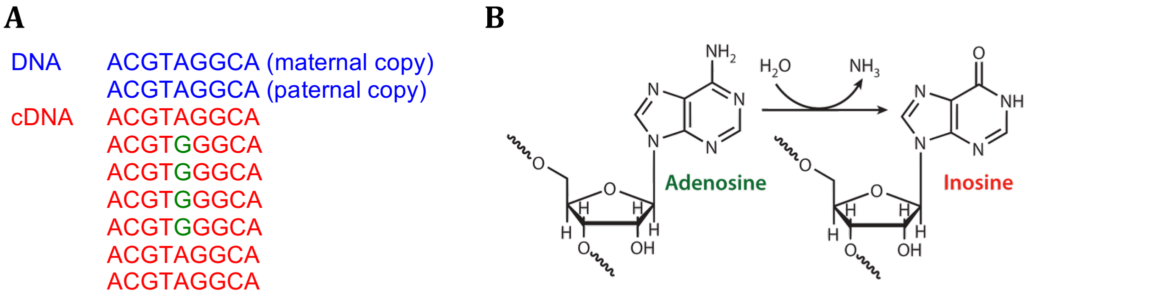

*Figure 2. (A) Hypothetical example of DNA sequence (blue) at a homozygous locus. Some proportion of RNA sequences transcribed from this locus will fail to match the sequence encoded in the genome due to RNA editing (green). (B) Mechanism catalyzed by ADAR, an Adenosine to Inosine (A-to-I) deamination along RNA transcripts, observed as an Adenosine to Guanine (A-to-G) after reverse transcriptase is used to make cDNA (Modified from Nishikura 2010).*

*ADAR biology and common targets*

In mammals, RNA editing is thought to take a predominant form transcriptome-wide, an adenosine to inosine deamination catalyzed by adenosine deaminase acting on RNA (ADAR) (Figure 2). Catalytic activity of ADAR is dependent on a double-stranded RNA (dsRNA) substrate, whereby conversion from adenosine to inosine on one strand typically results in unwinding of double stranded RNA species (Bass and Weintraub 1988). ADAR function is spread across three distinct genes, ADAR1, ADAR2, and ADAR3. ADAR1 is ubiquitously expressed and under the control of three promoters, two of which are constitutive while the third is interferon induced (George CX et al 1999). ADAR2 is primarily expressed among cells of the central nervous system (Melcher T et al 1996), while less is known about ADAR3, which is primarily expressed in the brain (Chen et al 2000).

Knockout experiments have shown that ADAR and RNA editing is essential for life. *ADAR1*$^{-/-}$ mice leads to embryonic lethality due to widespread apoptosis (Wang et al 2003), while *ADAR2* $^{-/-}$ mice die young from seizures resulting from under-editing at the Q/R site within GluR-B receptor pre-mRNA, required for glutamate AMPA receptor ion channel function. This phenotype can be rescued by introducing a GluR-B$^{R}$ genomic mutation, which mimics the edited form of GluR-B (Higuchi et al 2000).

Although RNA editing of coding regions such as GluR-B is essential for life, this phenomenon is suggested to rarely result in the recoding of protein sequences (Kleinman et al. 2012). Alternatively, high-throughput approaches to identify candidate RNA editing sites have shown that a substantial amount of RNA editing events reside within non-coding, repetitive regions. In humans, the primate specific Alu retrotransposon has been shown to attract most A-to-I editing transcriptome wide (Athanasiadis et al 2004; Blow et al. 2004; Levanon et al. 2004; Eisenberg et al. 2005; Bazak et al. 2014). Few comparative studies have been done to determine if retrotransposons in other species can attract as much editing activity. Many speculate that widespread A-to-I editing is specific to primates, due to Alu elements' relatively low divergence and high copy number, which increases the amount of dsRNA in the primate transcriptome relative to other species (Eisenberg et al. 2005, Neeman et al. 2006).

The initial sequencing of the human genome revealed a preference for Alus to be near and within genic regions (Lander et al. 2001). As a result, human RNA editing sites are typically clustered in introns, 3'UTRs, and gene-proximal regions (Athanasiadis et al 2004, Figure 3). The function of these non-coding RNA editing sites remains largely a mystery, but instances of Alu exonization modulated by RNA editing have been documented, in which either splice sites or splicing enhancers are created upon A-to-I editing (Lev-Maor et al. 2007, Sela et al. 2009). Additionally, RNAi pathways have been shown to be antagonized by RNA editing, as dsRNAs that are extensively edited become resistant to Dicer processing (Scadden and Smith 2001).

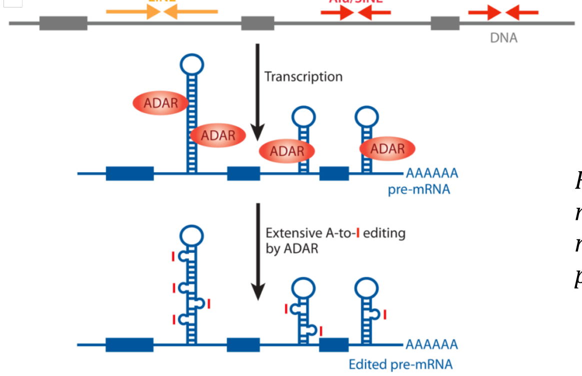

*Figure 3. Retrotransposons such as LINEs (long interspersed nuclear repeats) and SINEs (short interspersed nuclear repeats) direct the editing of pre-mRNAs in non-coding portions of genes (Modified from Nishikura 2010).*

Although editing of coding regions is rare and not marked by retrotransposons, it is nevertheless site-specific as in the case of the Q/R site within GluR-B. One theory is that non-Alu editing sites are dependent on nearby edited Alu sites, which attract ADAR in enough density to catalyze the editing of short, nearby double stranded stretches of RNA in coding regions (Ramaswami et al. 2012, Figure 4). Using in-vitro experiments, editing of either human or mouse *NEIL1*, *GLI1*, and *ZFP14* within their respective coding regions has been shown to be affected by the presence of flanking human Alu elements (Danial et al. 2014). It has yet to be determined whether non-primate SINE elements can act in the same way as the human Alu to enhance the editing level of nearby coding regions.

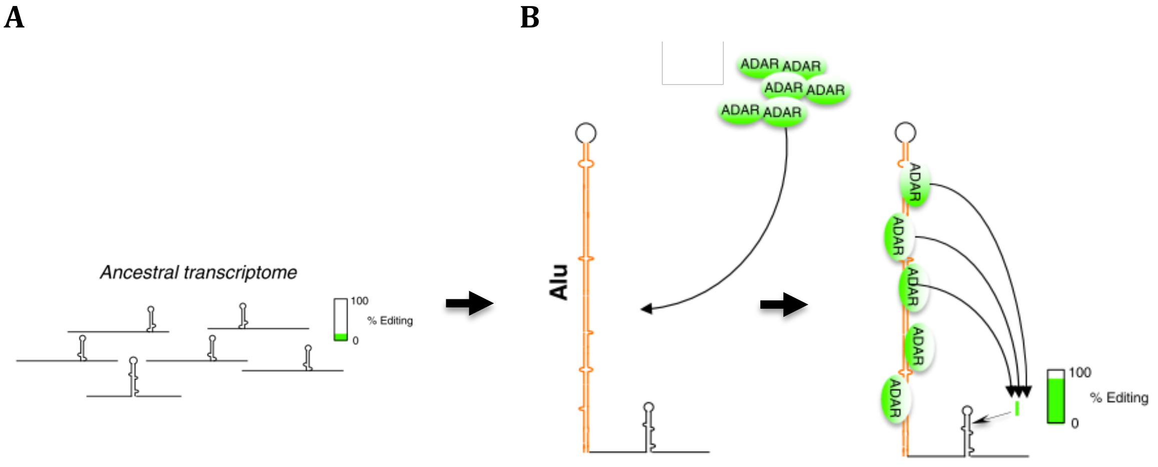

*Figure 4. Model for RNA editing of coding regions. (A) The ancestral transcriptome is thought to contain low levels of editing in coding regions, maintained by short hair-pin loops. (B) The arrival of the primate specific Alu has enhanced the length of hair-pin loops throughout the transcriptome, recruiting ADAR in high enough density to increase the editing of nearby coding regions (Modified from Daniel et al. 2014)*

### *Bioinformatic approaches to identify RNA editing sites*

No technology exists to sequence or genotype the "editome" (collection of all editing levels throughout the transcriptome) of any species. Instead, the computational methodology needed to identify instances of RNA editing using existing technology such as whole genome sequencing (WGS) and RNA-Seq is still being developed. In conducting these studies, the goal is to identify with high confidence, loci in which the genome sequence fails to match the corresponding transcriptome sequence. A-to-G (DNA-to-RNA) mismatches of these kind provide evidence of ADAR catalyzed A-to-I editing, because inosine is converted to guanine during reverse transcription. For any given mismatch, the proportion of reads containing guanine provides an estimate of the "editing level" for that site. Initial studies to incorporate WGS and RNA-Seq into an RNA editing detection pipeline have suggested widespread differences between DNA and RNA sequence consisting of substantial DNA-to-RNA mismatches of all types (many non A-to-G), transcriptome-wide (Li et al. 2011). Such findings were met with criticism, noting that non A-to-G mismatches have no biological explanation, and are suspiciously associated with the ends of sequencing reads while canonical A-to-G mismatches presumably resulting from ADAR editing are found to be uniformly distributed along the lengths of reads (Pickrell et al. 2012, Lin et al. 2012). Conservative pipelines for RNA editing detection that minimize the number of errors due to sequencing and mapping have since been the norm (Ramaswami et al. 2012, Chen et al. 2014, Fresard et al. 2015). As focus has continued on the human editome, most pipelines have shifted to utilizing RNA-Seq data alone to detect putative editing sites (Lee et al. 2013, Zhang and Xiao 2015). While this may be an attractive and cost-effective approach to identify and catalog human RNA editing sites, it may not be suitable for all species because these methods rely on extensive knowledge of genomic SNPs. Furthermore, these methods rely on multiple SNVs to be present in the same read to discern RNA editing from unknown SNPs.

### *Significance*

As data-driven observational studies continue to embody modern biology, new approaches for prediction and inference are needed to surmount challenges associated with estimating genome-wide parameters. Precision management of animals, along with precision medicine in humans, demand that genome-wide prediction can accurately model heterogeneous populations with complex substructure. As sequencing costs continue to decline, the feasibility to incorporate sequencing data into genome-wide regressions is greater and will be needed to gain prediction accuracy beyond what panels of common SNPs can provide. Even still, a complete genome sequence cannot contain all the information about a single individual or animal, as complex transcriptional processes such as RNA editing add complexity to the way in which genetic information is expressed. Research is needed to understand why RNA editing exists and how it effects the expression of genes and modification of phenotypes.

## RESEARCH METHODS

***Aim 1****) Develop and implement a new approach for genomic prediction, termed the "continuous model", to account for complex genetic heterogeneity*

In order to accomplish this aim, I propose a new algorithm outlined in the following steps:

1 - Methods such as k-medoids clustering are used to identify main groups among the population studied, using the genomic relationship matrix **G**.
2 - The medoids from each group are used as the "center" of each cluster, with individuals taken around each center within a Euclidian radius $r$ to use as reference individuals for cluster composition estimation (3).
3 - Cluster composition estimation is performed for all individuals, where estimates lie on a continuum from 0 to 1 for each cluster.
4 - A novel whole genome regression is performed (6), in which SNP effects are decomposed into the number of groups present plus a common effect, and the estimation of SNP effects is influenced by the cluster composition of each individual.

In *step 1*, the genomic relationship matrix between all individuals in the heterogeneous population is calculated by **G** = $\frac{\boldsymbol{XX'}}{p}$, where the genotype matrix $\boldsymbol{X}$ is first centered and scaled to a sample mean 0 and unit variance for each SNP. Clustering of **G** can be done using partitioning around medoids (P.A.M.), in iterative process that converges upon finding medoids, the observations from each cluster with minimal dissimilarity to all other observations in its cluster. This algorithm requires pre-specifying the number of clusters $k$, so first decomposing **G** into PCs and plotting the first 2 eigenvectors can help one estimate the number of clusters present. Alternatively, a silhouette analysis can be used to estimate $k$ a priori. In *step 2*, observations within Euclidean distance $r$ of each medoid are used to compute allele frequencies for each cluster for each SNP. The distance $r$ may vary depending on cluster size, but should be restricted to encapsulate < 25% of the observations from the cluster. In *step 3*, the allele frequencies computed in step 2 are used as the design matrix $\boldsymbol{X}$ from (3), which is used to predict each individual's genotypes in the dataset. For *step 4,* the resulting group composition estimates $\boldsymbol{c}$ are used to fit the following model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_0 + \boldsymbol{D}_1\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{D}_2\boldsymbol{X}\boldsymbol{b}_2 + \cdots + \boldsymbol{D}_k\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{k}} + \boldsymbol{\varepsilon} \tag{6}$$

where

$$\boldsymbol{D}_{\boldsymbol{i}} = \begin{bmatrix} \boldsymbol{c}_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{c}_{in} \end{bmatrix} \tag{7}$$

and $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ assume the same roles as in equation 1. $\boldsymbol{b}_0$ is a vector of SNP effect components common to all groups (main SNP effects), while $\boldsymbol{b}_i$ is a vector of SNP effect components for the $i^{\mathrm{th}}$group out of $g$ groups. Together, $b_{0\mathrm{j}} + b_{1\mathrm{j}} + b_{2\mathrm{j}} + \cdots + b_{\mathrm{kj}} = \beta_j$ represent the total SNP effect of the $j^{\mathrm{th}}$SNP. For "pure" individuals who have a group composition of 1 for group $i$, the effect of the $j^{\mathrm{th}}$ SNP becomes $b_{0\mathrm{j}} + b_{ij}$. Fitting of (4) can be accomplished with "Baysian Ridge Regression", as implemented in the R package BGLR (Perez P. and de los Campos G. 2014). Prior densities for each $\boldsymbol{b}_i$ are thereby considered Gaussian, with a scaled-inverse Chi-squared distribution used as the prior for the error.

To illustrate steps 1-3 of this approach more carefully, suppose there are two groups, in which case (4) simplifies to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_0 + \boldsymbol{D}_1\boldsymbol{X}\boldsymbol{b}_1 + (\mathbf{1} - \boldsymbol{D}_1)\boldsymbol{X}\boldsymbol{b}_2 + \boldsymbol{\varepsilon}.$ A wheat data set from CIMMYT, which has previously been used to demonstrate methods for prediction/selection using both the "stratified" and "interaction" models (Lehermeier et al. 2015, de los Campos et al. 2015) consists of 599 pure wheat lines genotyped for 1279 DArT markers (Triticarte Pty. Ltd., Canberra, Australia). We have used this data set to 1) form groups, 2) identify medoids and surrounding points for group composition estimation, and 3) estimate group composition for all wheat strains, while constraining group composition estimates to be between 0 and 1 (Figure 5).

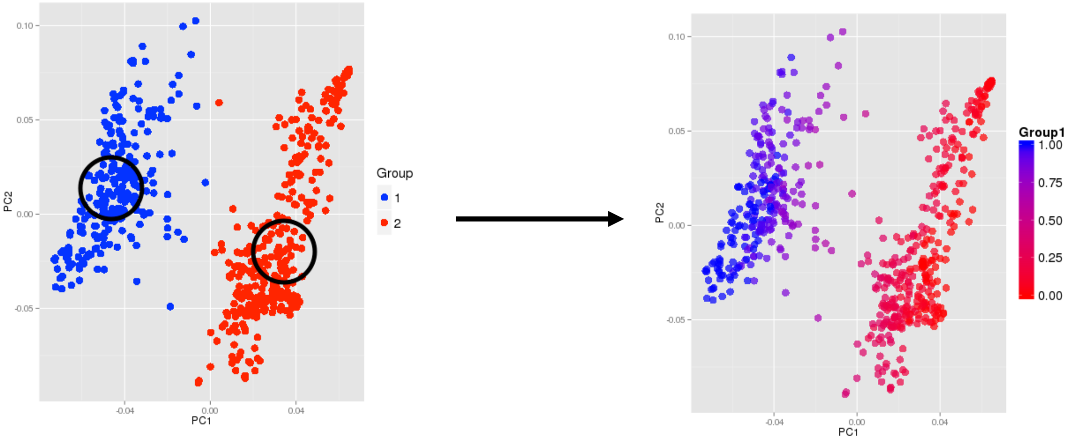

*Figure 5. Graphical representation of steps 1 through 3 of proposed algorithm using CIMMYT data. Center of black circles represent approximate locations of medoids of each identified cluster using P.A.M. and the radius of the circle encapsulates those observations used as reference in cluster composition estimation (left). Applying cluster composition estimation to all strains results in quantifying cluster composition on a continuous scale (right).*

For datasets such as the CIYMMT wheat strains featured in Figure 5, we do not expect the proposed model will enhance the estimation of SNP effects or the ability to perform accurate prediction. That is because these are pure strains coming from recombinant inbred lines, in which cluster composition is estimated to be nearly pure (100% group 1 or 100% group 2) for most individuals. In this case, we suspect the stratified model could outperform either the continuous model or the interaction model since most individuals fall into distinct genetic groups, speculated to have distinct SNP effects. However, I propose that prediction among a different dataset that exhibits more complex genetic substructure such as admixture could be made more accurate from the proposed continuous model (Figure 6).

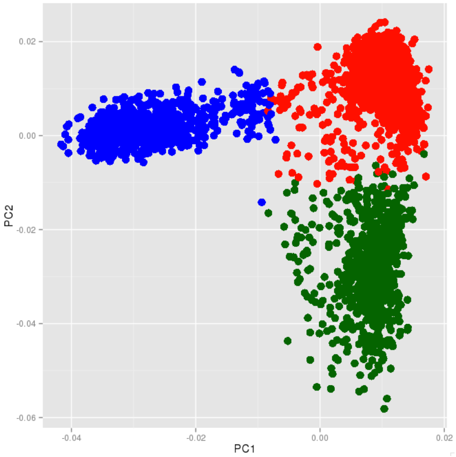

*Figure 6. Another example dataset, from the Pig Improvement Company (PIC), in which 3534 pigs from a single PIC nucleus line were genotyped for 50,436 SNPs. The **G** matrix was decomposed into the first two eigen vectors, and the first two PCs are plotted. Colors represent distinct clusters identified using K-medoids clustering.*

***Aim 1a**) Determine how prediction accuracy from the continuous model compares to prediction accuracy from previously developed "stratified" and "interaction" models.*

In order to assess prediction accuracy of the continuous model, I will compare the performance of all three models – the continuous, the interaction, and the stratified, using high density SNP datasets that exhibit varying degrees of substructure complexity (Table 1).

*Table 1. Potential sources of to use in the estimation of continuous model prediction accuracy.*

| **Dataset** | ***n* (individuals)** | ***p* (Genotyped SNPs)** | **# traits phenotyped** | **Origin** |
|---|---|---|---|---|
| CIYMMT | 599 | 1,279 | 1 trait measured under 4 environments | Wheat commercial population (Figure 5) |
| PIC | 3,534 | 50,436 | 3 | Swine commercial population (Figure 6) |
| MSUPRP | 948 | 8,848 | 2 | F2 – Decsending from Pietrain and Duroc swine breeds |

CIYMMT, PIC, and MSUPRP datasets are all of those that I currently possess and am available to utilize. Beyond the datasets provided in Table 1, I will seek to obtain new SNP datasets in which there are at least 1,000 individuals genotyped at > 10,000 SNPs and phenotyped for at least 1 trait. Exact origins and species can vary, but the amount of "cluster admixture" measured by the proportion of individuals that have a cluster composition of less than 75% for any one cluster, will ideally be greater than 0.25 in order to maximize the benefits of the continuous model.

Prediction accuracy will be compared between the three models by first dividing each dataset into training (TRN) and testing (TST) components. TRN datasets will be used to fit the models expressed in equations (3), (4), and (6). The accuracy of predictions for each model will be assessed in TST individuals. 50 repetitions of prediction accuracy assessment will be done by randomly assigning 150-900 individuals to the TST component (depending on total sample size *n*), leaving the remainder to the TRN component. For each dataset, prediction accuracy will be measured as the correlation between predictions and observed phenotypes after 50 trials.

We hypothesize that no one model – stratified, interaction, or continuous will outperform the others for all datasets. Instead, the optimal model to use to obtain maximum prediction accuracy will likely depend on the data and its substructure, which can be characterized by parameters including the aforementioned "cluster admixture". Using a dataset where cluster admixture is roughly zero, meaning that all individuals belong to a cluster with a cluster composition > 0.75, to fit either the interaction or continuous model will likely result in over fitting because such datasets are less likely to contain "main effects" $\boldsymbol{b}_0$ common between groups in addition to group specific effects.

***1b**) Develop the means to perform hypothesis testing with the continuous model in order to infer significance of SNP effects.*

In addition to proposing the continuous model as one for prediction, I will assess the continuous model's ability to infer QTL. Whole genome regressions have potential to infer which genetic loci have an impact on phenotype after considering all polygenic background effects. The continuous model has potential to not only aid in prediction in situations where complex substructure warrants the decomposition of SNP effects into various components, but to enable the detection of average SNP effects among complex populations and determine their significance. To enable hypothesis testing for significant QTL using the continuous model, I will fulfill the following objectives:

1. Develop the means to test the null hypothesis $H_0\colon \boldsymbol{b}_{0j} + \boldsymbol{b}_{1j} + \boldsymbol{b}_{2j} + \cdots + \boldsymbol{b}_{gj} = \boldsymbol{\beta}_j = 0$, against the alternative hypothesis $H_A\colon \boldsymbol{\beta}_j \neq 0$. In this case, $\boldsymbol{\beta}_j$ indicates the total effect of the jth SNP when considering its effects across all groups.
2. Determine the null distribution needed to obtain p-values.

Although computationally onerous, 2) may be established by using permutations. Upon establishing 1) and 2), additional tests will be implemented, including:

1. $H_0\colon \boldsymbol{b}_{0j} + \boldsymbol{b}_{ij} = \boldsymbol{\beta}_{ij} = 0$ against $H_A\colon \boldsymbol{\beta}_{ij} \neq 0$. $\boldsymbol{\beta}_{ij}$ provides an estimate of the jth SNP effect in the ith group.
2. $H_0\colon \boldsymbol{b}_{ij} - \boldsymbol{b}_{i'j} = 0$ against $H_A\colon \boldsymbol{b}_{ij} - \boldsymbol{b}_{i'j} \neq 0$. $\boldsymbol{b}_{ij} - \boldsymbol{b}_{i'j}$ provides an estimate of the difference in SNP effects between groups $i$ and $i'$.

How the continuous model is fit can have a dramatic impact on the estimates of $\boldsymbol{\beta}_j, \boldsymbol{\beta}_{ij}$, and $\boldsymbol{b}_{ij} - \boldsymbol{b}_{i'j}$**.** After establishing the means to conduct the hypothesis tests above, I will test these methods with the aforementioned datasets by utilizing a number of procedures to fit the continuous model. These will include Bayesian versions of ridge regression (Hoerl and Kennard 1970) in which all SNP effects $\boldsymbol{b}_{ij}$ are shrunk toward 0, the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani 1996; Bayesian LASSO, Park and Casella 2008) a variable selection procedure in which null SNP effects are shrunk to 0 and removed from the model, or the Bayesian methods from Meuwissen et al, which non-null SNP effects are assumed to come from a different distribution from null SNP effects. For each dataset and for each Bayesian estimator, significance of SNP estimates will be compared across the continuous, interaction and stratified models to determine how QTL inference may be affected by the choice of model or the choice of estimator.

***Aim 2****) Investigate the function and impact of RNA editing throughout mammalian evolution and build novel tools for the detection, analysis, and visualization of RNA editing data.*

Based on the observation that RNA editing of NEIL1, GLI1, and ZFP14 are influenced by the positioning of nearby Alu elements (Daniel et al. 2014), and that differences in transcriptome-wide editing between human and mouse have been attributed to specific properties of the species' repetitive elements (Neeman et al. 2006), I am primarily interested in conducting a comparative RNA editing study involving multiple mammalian species harboring distinct repetitive elements. This study will utilize novel software developed specifically for the purpose of high-throughput RNA editing analysis. I will also assess how RNA editing detection is affected by the use of long-read sequencing technology vs. short-read sequencing technology.

***2a****) Build software specifically designed for the detection, analysis, and visualization of RNA editing data from multiple high-throughput technologies.*

While multiple approaches to identify instances of RNA editing have been developed, the code used in RNA editing analyses is rarely published in full. It is possible that discrepancies in RNA editing results are at least in part due to poor practices in reproducible research. In order to address these issues, I will write software for the R (R Core Team, 2015) framework intended to analyze large bioinformatic files, detect candidate RNA editing sites, and visualize RNA editing data in novel ways. The finished R package will be submitted to BioConductor while an open-source version will continually be made available on GitHub so that developers may contribute or modify the package for their own use. No such software currently exists and is made widely available specifically for the purpose of RNA editing analysis within the R framework, which is becoming an increasingly popular framework for computing and graphics within the biological community.

This software, (tentatively named "editTools") is aimed at implementing methods for:

1. The detection and analysis of RNA editing data from Variant Call Format (VCF) and Sequence Alignment/Map (SAM) files containing whole genome and whole transcriptome sequencing data.
2. Merging RNA editing data with other annotation files, including those for genes/transcripts, repetitive elements, miRNAs, and miRNA target sites.
3. Visualizing merged datasets by way of plotting discrete count data (e.g. Counts of A-to-G mismatches sites across tissue types, repetitive element types, etc.), plotting continuous data (e.g. editing levels), and positional data (interactive chromosome maps to show the position of candidate RNA editing sites relative to genes, repetitive elements, etc.)
4. The ability to process data from both short and long read sequencing technology, where differences in technology warrant alternate mapping and variant calling techniques.

Accomplishing implementations 1 & 2 requires surmounting limitations to the R language, namely, the ability to process large bioinformatic files efficiently. To accomplish this, C++ source code can be used to build the editTools package, providing increased performance for the user. I have already written a basic C++ library to analyze VCF files, and existing C++ libraries to handle SAM/BAM files such as BamTools (Barnett et al. 2011) can be incorporated into the editTools package. To accomplish all visualization implementations, base plotting tools in addition to the ggplot2 package (H. Wickham, 2009) will be used to plot discrete count data and continuous data. Plotting positional data by way of interactive chromosome maps can be accomplished with BioJS, a javascript infrastructure for visualizing and interacting with biological data.

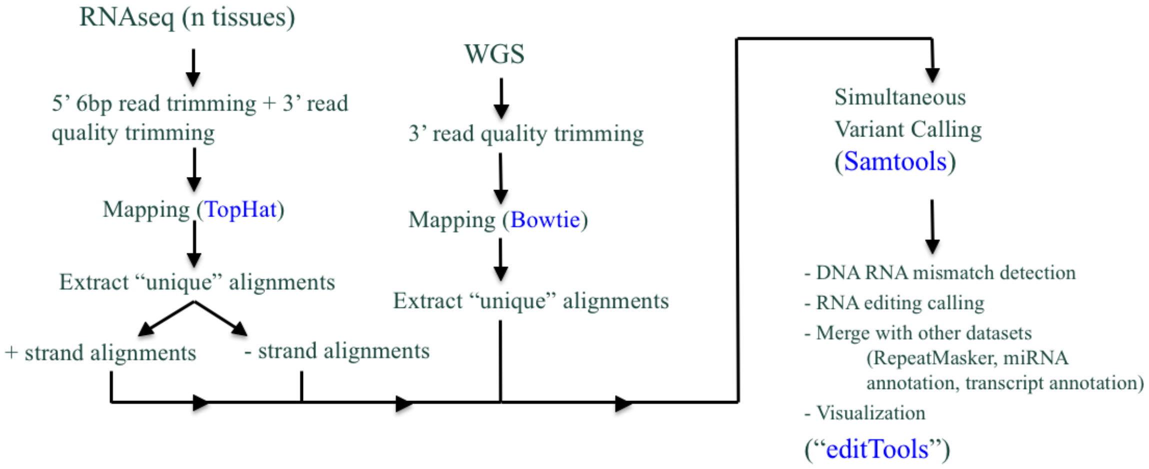
RNAseq (n tissues)
5' 6bp read trimming + 3' read quality trimming
Mapping (TopHat)
Extract "unique" alignments
+ strand alignments
- strand alignments
WGS
3' read quality trimming
Mapping (Bowtie)
Extract "unique" alignments
Simultaneous Variant Calling (Samtools)
- DNA RNA mismatch detection
- RNA editing calling
- Merge with other datasets (RepeatMasker, miRNA annotation, transcript annotation)
- Visualization
("editTools")

*Figure 7. One proposed pipeline for editTools usage to analyze a single individual with n tissues, in which VCF files (output of Samtools) are used as input for editTools.*

Pre-processing (trimming, mapping, etc.) of raw sequencing files will be required for editTools usage. editTools is currently in development, and is capable of the pipeline depicted in Figure 7. Example usage of the pipeline and editTools is outlined here:

1. Raw sequencing reads from RNA-seq (cDNA reads) and WGS (DNA reads) are trimmed for quality at their 3' ends. In addition, cDNA reads are trimmed 6bp at their 5' ends to eliminate misidentification of DNA RNA mismatches due to artifacts associated with the use of random hexamers during cDNA library prep (Lin et al. 2012).
2. cDNA reads are mapped to the reference genome using a splice-aware aligner such as TopHat (Trapnell et al. 2009), while DNA reads are mapped to the reference genome using a similar algorithm such as the one in Bowtie (Langmead and Salzberg 2012). Parameters for mapping may effect downstream RNA editing results (Lee et al. 2013).
3. Uniquely mapped DNA and cDNA alignments are kept to be used in downstream analysis. "Uniquely mapped" can take on various interpretations (Chen et al. 2014), but conservatively can be thought of as those reads that contain only one alignment in a sequence alignment file. In conducting this step, the goal is to limit artificial DNA-RNA mismatches due to multi-mapped reads caused by homologous sequences and imperfect mapping algorithm design.
4. In order to distinguish plus-strand transcripts from minus-strand transcripts, which in turn is needed to distinguish A-to-G (DNA-to-RNA) mismatches from T-to-C mismatches, cDNA reads coming from plus-strand transcripts are separated from those coming from minus-strand transcripts.
5. Variant calling algorithms to produce variant calling format files (VCF) such as Samtools and Bcftools (Heng L 2011) are used to process DNA and cDNA alignments simultaneously. If Samtools is used, the argument "-t DP, DV, SP" is required for downstream usage with editTools. These arguments force the resulting VCF file to include per sample read depths, variant read depths, and strand bias p-values for each variant site.
6. Analysis of VCF files is done using editTools, which by default, identifies candidate RNA editing sites where 1) the genotype is homozygous according to 95% of the DNA reads, 2) at least 10 DNA reads were used to derive the genotype, 3) at least 5 cDNA reads from the same tissue disagree with the genotype call, and 4) these cDNA reads must have a strand-bias P-value of 0.1 or greater.

The above pipeline uses VCF files as input for editTools, which requires extra work on behalf of the user to accomplish steps 3-5 with external software. Once editTools incorporates the ability to handle SAM/BAM files, steps 3-5 will become unnecessary for the user since editTools will be able to handle them internally.

***2b**) Test software using swine as a model for RNA editing. Long read and short read sequencing technologies will be compared for their ability to detect RNA editing events, and a subset of RNA editing candidates will be validated to assess detection accuracy.*

I will utilize WGS and RNA-Seq data from 2 Yorkshire male pigs from the Functional Annotation of Animal Genomes project (FAANG) to apply the existing editTools pipeline (Figure 7) and examine the previously unstudied swine "editome". Liver, fat, muscle, spleen, cortex, cerebellum, hypothalamus, and lung were collected from each animal. The Illumina TruSeq Stranded Total RNA Sample Preparation protocol has been used to prepare RNA libraries to sequence at a depth of 100 million reads per tissue using the Illumina HiSeq 2000. Likewise, genomic DNA from each animal will be purified from muscle tissue using the Invitrogen Purelink Genomic DNA Mini Kit and library preps generated with the Illumina TruSeq Nano DNA Library Preparation Kit HT, used to sequence the genome with roughly 30X coverage using the Illumina HiSeq 2500.

With data from 8 tissues and two biological replicates, I will analyze the differences in editing patterns between the tissues and determine if differential ADAR expression can explain the variation that is observed in the number of editing loci and average editing levels. Canonical (A-to-G) editing sites within coding regions (expected to be less than 100) will be validated using Sanger sequencing. A subset ( < 30) of non-canonical candidate editing sites will also be selected for Sanger sequencing. I will determine if there are any patterns among those that can and cannot be validated, so as to improve the predictive accuracy of the current pipeline.

Using "3rd generation" long-read sequencing platforms developed by Pacific Biosciences (PacBio), we will also apply editTools to determine how long-read sequencing technology compares to the "2nd generation" short-read technology of Illumina in conducting genome-wide RNA editing scans. Potentially, using long-read technology can overcome challenges associated with mapping reads to the reference genome and could improve the accuracy of RNA editing detection (Roads and Au 2015). From USDA-MARC, we will obtain PacBio Single Molecule, Real-Time (SMRT) genome sequence, along with PacBio full-length Iso-Seq transcriptome sequences from the same animal for a number of swine tissues including hypothalamus, spleen, thymus, and small intestine. Results in the form of DNA-RNA mismatch counts, editing levels, and positions of mismatches will be compared when using PacBio long-read cDNA reads against Illumina short-read cDNA reads from the same animal and tissues. In both instances, SMRT DNA reads will be used to determine the genotypes of candidate RNA editing loci. Although PacBio sequencing is known to have a higher error rate, we propose that the drawbacks to sequencing error can be overcome with sufficient depth. That is why we will require at minimum 30X coverage of genome and transcriptome from both SMRT and Illumina technologies.

***2c****) Perform a comparative study in which RNA editing patterns are compared across a number of mammalian species harboring different genomic features.*

As RNA editing studies are becoming increasingly human-centric, I will reassess the "editability" of other mammalian genomes again using the functionality of editTools and short-read sequencing technology. We believe that the human Alu may not be the only SINE element to influence the editome of a mammal in significant ways. The swine specific PRE1 element resembles the human Alu in many ways including copy number, length, and divergence. If editability is of a transcriptome is dependent on these factors, then we would expect the swine transcriptome to be as edited as human. Through collaborators and available online datasets, we will obtain WGS and liver RNA-Seq data from five different representative species (Table 2), each possessing a different SINE repertoire.

*Table 2. Representative species/SINE elements for comparative RNA editing study*

| **SINE element** | **Length** | **Copy number** | **Structure***  | **Species** |
|---|---|---|---|---|
| Alu | 282bp | 1.1x10$^6$ | 7SL-7SL | Primates |
| B1/B2/B4 | 135/185/278bp | 8x10$^3$ – 6.5x10$^5$ | 7SL | Rodents |
| PRE1 | 246bp | 1x10$^6$ | tRNA-?? | Swine |
| Bov-tA | 204bp | 2x10$^5$ | tRNA-LINE | Cattle, goats, sheep |
| Fc-1 | 105bp | 4-6x10$^4$ | tRNA | Cat, dog, panda |

* Structure denotes the ancestral origins and structure of each element. For example, the primate Alu element is made up of two components each derived from a 7SL RNA. "??" denotes an unknown structural component.

We will require 2 biological replicates from each – human or macaque, mouse, pig, cattle or sheep, and cat or dog. For each animal/individual, genome sequencing depth must be at least 30X, and liver transcriptome sequencing must consist of at least 90M cDNA reads. For swine/PRE1, the two Yorkshire animals used in aim 2b will be reused.

In this observational study, the editomes of each species will be compared and the level of conservation of RNA editing patterns will be elucidated. Total numbers of editing loci, their positions, and their editing levels will be analyzed to determine if widespread RNA editing is truly a unique feature to humans and other primates. The editability of each SINE element and each SINE subfamily for each species will be evaluated. Editability can be strictly defined as the fraction of the number putative edits / (number of putative edits + adenosines) along a particular SINE element. For each candidate RNA editing site in a repetitive region, I will calculate the distance to the nearest inverted repetitive region of the same type, since only those SINE elements that are able to form dsRNA should be editable. Lastly, to assess the theory that RNA editing in coding regions is influenced by nearby SINE elements, I will identify putative A-to-I editing sites in coding regions for each species and assess their distance to nearby edited SINE elements. If across species, conserved RNA editing sites in coding regions are associated with nearby SINE elements, then this can provide evidence that SINE elements, regardless of their origin, may have positional constraints to preserve the editing of particular coding regions.

## PROSPECTIVE TIMELINE

| | Aim 1 | Aim 2 |
|---|---|---|
| Spring 2016 | Seek additional SNP datasets for prediction analysis. | Locate additional sequencing datasets for comparative study. Enhance the functionality of editTools to take SAM/BAM as input. |
| Summer 2016 | Determine prediction accuracy of continuous model. Begin theoretical work needed for inference. | Begin preparing editTools for Bioconductor submission. RNA editing analysis of 8 swine tissues. |
| Fall 2016 | Test inference abilities of continuous model. | Finalize editTools submission. Assess impact of RNA editing detection using long vs short reads. |
| Spring 2017 | Finalize prediction and estimates estimates with any additional datasets. | Conduct comparative RNA editing study using multiple species. |
| Fall 2017 | Finalize Aims 1 and 2. Refine editTools to keep pace with other software, writing. | |
| Spring 2017 | Writing, defense, corrections. | |

## POTENTIAL FUNDING SOURCES

Funding from USDA and NSF will be sought for additional resources toward whole-genome sequencing and whole-transcriptome sequencing of swine to aid in aims 2b and 2c.

## REFERENCES

Athanasiadis, A., Rich, A., & Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biology, 2(12). doi:10.1371/journal.pbio.0020391

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P., & Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics, 27(12), 1691–1692. doi:10.1093/bioinformatics/btr174

Bass, B. L., & Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell, 55(6), 1089–1098. doi:10.1016/0092-8674(88)90253-X

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., … Levanon, E. Y. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Research, 24(3), 365–76. doi:10.1101/gr.164749.113

Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., & Tromp, M. C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell, 46(6), 819–826. doi:10.1016/0092-8674(86)90063-2

Chen, C. X., Cho, D. S., Wang, Q., Lai, F., Carter, K. C., & Nishikura, K. (2000). A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA (New York, N.Y.), 6(5), 755–767. doi:10.1017/S1355838200000170

Chen, J.-Y., Peng, Z., Zhang, R., Yang, X.-Z., Tan, B. C.-M., Fang, H., … Li, C.-Y. (2014). RNA editome in rhesus macaque shaped by purifying selection. PLoS Genetics, 10(4), e1004274. doi:10.1371/journal.pgen.1004274

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., & Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science, 90(10), 3375–3384. doi:10.2527/jas.2011-4557

Daniel, C., Silberberg, G., Behm, M., & Ohman, M. (2014). Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biology, 15(2), R28. doi:10.1186/gb-2014-15-2-r28

de los Campos, G., Veturi, Y., Vazquez, A. I., Lehermeier, C., & Pérez-Rodríguez, P. (2015). Incorporating Genetic Heterogeneity in Whole-Genome Regressions Using Interactions. Journal of Agricultural, Biological, and Environmental Statistics. doi:10.1007/s13253-015-0222-5

Eisenberg, E., Nemzer, S., Yaron, K., Rotem, S., Gideon, R., & Levanon, E. Y. (2005). Is abundant A-to-I RNA editing primate-specific? Trends in Genetics, 21(2), 73–77. doi:10.1016/j.tig.2004.12.004

Falconer, D. S., and T. F. S. Mackay (1996). An Introduction to Quantitative Genetics. Longman Group, Essex, UK

Frésard, L., Leroux, S., Roux, P.-F., Klopp, C., Fabre, S., Esquerré, D., … Pitel, F. (2015). Genome-Wide Characterization of RNA Editing in Chicken Embryos Reveals Common Features among Vertebrates. PloS One, 10(5), e0126776. doi:10.1371/journal.pone.0126776

George, C. X., & Samuel, C. E. (1999). Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. Proceedings of the National Academy of Sciences of the United States of America, 96(8), 4621–4626. doi:10.1073/pnas.96.8.4621

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution, 41(1), 51. doi:10.1186/1297-9686-41-51

Higuchi, M., Maas, S., Single, F., & Hartner, J. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature, 406(July), 1998–2001. Retrieved from http://www.nature.com/nature/journal/v406/n6791/abs/406078a0.html

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55–67. doi:10.1080/00401706.1970.10488634

Janss, L., G. de los Campos, N. Sheehan, and D. Sorensen. 2012. Inferences from Genomic Models in Stratified Populations. Genetics 192(2):693-704.

Kleinman, C. L., Adoue, V., & Majewski, J. (2012). RNA editing of protein sequences: A rare event in human transcriptomes. Rna, 18(9), 1586–1596. doi:10.1261/rna.033233.112

Kuehn, L. a, Keele, J. W., Bennett, G. L., McDaneld, T. G., Smith, T. P. L., Snelling, W. M., … Thallman, R. M. (2011). Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. Journal of Animal Science, 89(6), 1742–50. doi:10.2527/jas.2010-3530

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. doi:10.1038/35057062

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–9. doi:10.1038/nmeth.1923

Lee, J., Ang, J. K., & Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants, (Strategy 1), 725–732. doi:10.1261/rna.037903.112.Park

Lee, J., Ang, J. K., & Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants, (Strategy 1), 725–732. doi:10.1261/rna.037903.112.Park

Lehermeier, C., Schon, C.-C., & de los Campos, G. (2015). Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. Genetics, 201(1), 323–337. doi:10.1534/genetics.115.177394

Lev-Maor, G., Sorek, R., Levanon, E. Y., Paz, N., Eisenberg, E., & Ast, G. (2007). RNA-editing-mediated exon evolution. Genome Biology, 8(2), R29. doi:10.1186/gb-2007-8-2-r29

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., … Jantsch, M. F. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nature Biotechnology, 22(8), 1001–1005. doi:10.1038/nbt996

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England), 27(21), 2987–93. doi:10.1093/bioinformatics/btr509

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., & Cheung, V. G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. Science (New York, N.Y.), 333(6038), 53–8. doi:10.1126/science.1207018

Lin, W., Piskol, R., Tan, M. H., & Li, J. B. (2012). Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome." Science, 335(6074), 1302–1302. doi:10.1126/science.1210624

Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M., & Seeburg, P. H. (1996). RED2, a brain-specific member of the RNA-specific adenosine deaminase family. J Biol Chem, 271(50), 31795–31798. doi:10.1074/jbc.271.50.31795

Meuwissen, T. H. E., and M. E. Goddard, 1996 The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28: 161–176.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829. doi:11290733

Neeman, Y., Levanon, E. Y., Jantsch, M. F., & Eisenberg, E. (2006). RNA editing level in the mouse is determined by the genomic repeat repertoire. RNA (New York, N.Y.), 12(10), 1802–1809. doi:10.1261/rna.165106

Park, T., & Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482), 681–686. doi:10.1198/016214508000000337

Perez, P., & de los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics, 198(2), 483–495. doi:10.1534/genetics.114.164442

Pickrell, J. K., Gilad, Y., & Pritchard, J. K. (2012). Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome." Science, 335(6074), 1302–1302. doi:10.1126/science.1210624

Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. a, & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), 904–9. doi:10.1038/ng1847

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C., & Li, J. B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nature Methods, 9(6), 579–581. doi:10.1038/nmeth.1982

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics, (August). doi:10.1016/j.gpb.2015.08.002

Scadden, a. D. J., & Smith, C. W. J. (2001). RNAi is antagonized by A -> I hyper-editing. EMBO Reports, 2(12), 1107–1111. doi:10.1093/embo-reports/kve244

Sela, N., Mersch, B., Hotz-Wagenblatt, A., & Ast, G. (2010). Characteristics of transposable element exonization within human and mouse. PLoS ONE, 5(6), e10907. doi:10.1371/journal.pone.0010907

Tibshirani, R. (1994). Regression Selection and Shrinkage via the Lasso. Journal of the Royal Statistical Society B. doi:10.2307/2346178

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England), 25(9), 1105–11. doi:10.1093/bioinformatics/btp120

Vassetzky, N. S., & Kramerov, D. a. (2013). SINEBase: A database and tool for SINE analysis. *Nucleic Acids Research*, *41*(D1), 83–89. doi:10.1093/nar/gks1263

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi:10.1038/ng.608

Zhang, Q., & Xiao, X. (2015). Genome sequence–independent identification of RNA editing sites. Nature Methods, 12(4). doi:10.1038/nmeth.3314