Smith, E. J., Henshall, J.M. (2009). Variability in the Distributions of Single Nucleotide Polymorphism Effects in Livestock Populations. *Proceedings of the 18th Conference of the Association for the Advancement of Animal Breeding and Genetics, 18*, 64-67.

***Background***

* SNP is a DNA sequence variation when a *single nucleotide* (A,T,C,G) in the genome differs between members of a biological species or paired chromosomes in humans.
* Compare two sequenced DNA fragments from different individuals: AAGCCTA, AAGCTTA
  + Differs by a single nucleotide and is an example of two alleles.
  + *Allele* is an alternative form of the same gene that can result in different observable *phenotypic traits*, such as pigmentation.
  + Almost all common SNPs have only two alleles.
* SNPs are assigned a *minor allele frequency* (the lesser of the two frequencies) within a population.
  + An SNP allele common in one geographic or ethnic group may be much rarer in another, demonstrating genetic variations between individuals.
  + Such information is most useful in DNA fingerprinting, disease detection and treatment.
* *SNP density* is affected by the following factors:
  + *Genetic recombination* – new combinations of alleles, encoding a novel set of genetic information
  + *Mutation rate –* measured in units per gamete
* *Genome size* is total amount of DNA contained within one copy of a single genome.
* *Quantitative trait loci* (QTL) is a region of DNA associated with a particular phenotypic trait, often found on different chromosomes.
  + *Phenotypes* can be modelled as the sum of genetic and environmental effects.
  + *Heritability* reflects all genetic contributions to a population’s phenotypic variance including *additive*, dominant andmaternal/paternal effects.
  + *Additive variance* is the variance due to the average (additive) effects of the alleles.
* Past research on the distribution of QTL effects suggest more rigorous and robust analyses required.
  + Hayes and Goddard (2001) found QTL effects on pig and dairy data displayed a skewed distribution with a few QTL of large effect.
  + Mackay (2004) found that homozygous QTL exhibited an exponential distribution, with most of the variation between parental lines attributable to larger effects.
  + Roff (2007) highlights the need to study the distribution of QTL effects with greater statistical precision.

***Outline***

* *Aim:* To identify factors that influence the distribution of SNP effects.
* *Scope:*
  + Bayesian methods used in association studies of dense SNP and phenotype data, rely on assumptions about the distribution of SNP effects.
  + Obtaining reliable estimates for the *true and unknown* distribution of SNP effects is hindered by limited data.
  + Simulation was used to accommodate for this lack of data.
* *Method:*
  + Five simulations of livestock populations were performed given the following parameters:
    - #SNPs 🡪 number of SNPs
    - SNP / cM 🡪 SNP density
    - Dams 🡪 number of female parents
    - Sires 🡪 number of male parents
    - U 🡪 distribution of sampled SNP effects
* *Model:*
  + paired allele frequencies, where
  + *mutation rate* = 3.1 x 10-4 per gamete
  + - where is SNP effect size and
    - is the uniform distribution
  + 🡪 environmental variance (constant)
  + 🡪 genetic variance (target)
  + 🡪 absolute value of the SNP effect sizes
  + 🡪 additive variance
* *Specifications:* 
  + was adjusted to account for when SNP was fixed at
  + was not simulated so only the narrow-sense definition of heritability is adopted, meaning only was modelled.
  + “Assortative” mating system was simulated to account for SNP transmission between animals, mutation and recombination effects.
  + There is algorithm convergence since the results of three repeated runs of year periods of simulated data were all similar, indicating stabilised simulations.
* *Analysis:* 
  + *SNP effects*
  + *QTL effects* 
    - 🡪 phenotype
    - 🡪 controlled and set to zero in a planned experiment.
    - 🡪 broad definition of heritability
    - 🡪 narrow definition (additive variance only)
    - 🡪
* *Results:* 
  + Findings do not support the assumption that SNP effect distributions follow an exponential function where .
  + Frequency histograms suggest that was not strictly greater than zero, with S1,S3,S4,S5 containing:
    - inflexion points where
    - concave down regions where
  + S2 is clearly not exponential and displays a uniform (rectangular) distribution instead.
  + Such results indicate that may depend on the population parameters used in the simulation, rather than obeying an exponential function.
  + >6 or *large effect observations* in S5 suggest there is an upper limit to the effect size for mutations that can survive in a population.
  + Analysis of the *additive variance* () as a function of *allele frequency (*), suggest that the different (uniform) distributions of *sampled* SNP effects is another factor that influences.
    - S1 and S5 have the same simulation parameters except for the width of the sampling interval and yet their distributions are significantly different.
* *Conclusions:*
  + Use of particular distributions (like the exponential) as priors for Bayesian analyses of SNP effects is invalidated.
  + SNP effect distribution was found to be influenced by genome size, SNP density, population size and the distribution of sampled SNP effects.
* *Comments:*
  + Are Bayesian prior assumptions of exponentiality based on QTL effects rather than on SNP effects (due to lack of data)? In other words, is QTL a proxy variable for SNP?
  + How were the interval values for uniform distribution chosen?
  + Use of MCMC methods to approximate distribution of realised simulations for more accurate inferences, in addition to histograms.
  + Simulate genetic variance in order to model broad-sensedefinition of heritability, rather than “tuning” the simulation to keep consistent with observed heritabilities. This is to reduce underestimation bias since.
  + Alternative approaches to genomic simulation explored by more recent studies below.
* *Recent studies:*
  + The following two papers are co-authored by UNE Postdoctoral Fellow at the School of Environmental and Rural Sciences, Dr John Hickey.
    - Daetwyler et al. (2013) “Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking”
    - Hickey & Gorjanc (2012) “Simulated Data for Genomic Selection and Genome-Wide Association Using a Combination of Coalescent and Gene Drop Methods”