Question and Comments From Andy Baumgarten on "Utilizing evolutionary conservation information to improve prediction accuracy in genomic selection"

- **GERP-Score:** I would be interested to understand how much the GERP-score is adding to the predictive power over the benefits of just fitting the haplotype effects. One potential approach would be to assign the same weights to all haplotypes in the additive model, run the predictive power analysis, and compare the results to the approach considering GERP. The same approach could be used for the dominance model where a value of 1 would be assigned to heterozygote and non-reference SNP calls and 0 to everything else.

  Another approach to this is to run the same analysis using just SNP calls. This would be a bit of pain given the number of SNPs, but you could down-sample the SNPs and rerun the same analysis.

- **Circular Shuffling:** This is really cool approach to randomizing the data. One question/comment is whether the 1 Mbase shift if enough to negate LD patterns within the peri-centromeric (large amounts of sequence vs. little recombination) regions of the genome. I only bring this up as a potential explanation of the ability of the permuted data to predict the different traits so well. The majority of the inbreds in the study are Pioneer lines or selfs out of Pioneer hybrids. All of the inbreds involved have 2 - 4 cycles of breeding within Pioneer and would have defined heterotic group assignment. Furthermore, there would be defined LD patterns within a heterotic group. Shifting 1 Mb in the peri-centromeric regions may still allow for LD within the haplotypes, explaining the predictive ability of the shuffled data.

- **BayesC:** If I remember correctly, Jinliang ran the BayesC approach using a pi-value around 0.99. This high of pi-value will potentially assign effects to a small set of regions across all iterations, potentially indicating regions of the genome that are driving predictive power. It would be interesting to examine the % of iterations where a haplotype block was present in the analysis, the average effect of variation of the haplotype block across iterations, and average % variation explained. This could be examined for additive and dominance models and used to investigate interesting heterotic trends or compare whether high effect haplotypes found in the dominance model do indeed fall in low recombination regions. It would also allow you to examine the correlation of these summaries with the GERP scores for given regions.

- **GCA/SCA and Heterosis:** I think it be useful to include the GCA (general combining ability)/SCA (specific combining ability) modelling directly within the ASReml phenotypic analysis. I would recommend fitting terms for male parent, female parent, and their interaction as random terms in this Asreml analysis. This will result in the additive effect of each inbred across hybrids as well as the deviation of a hybrids vs. the additive effects of the two parents. My experience is that you will gain some power by running this analysis in one step instead of two.

To do this, I would use the exact same Asreml model but with a male parent, female parent, and hybrid term. Asreml then has a linear "predict" statement that allows you to assemble BLUPs across the lines using linear combination of effects found in the model. The predict statement I would use to estimate a GCA-based value for the hybrid would be:

Predict Male Female !present Male Female

- This gives the hybrid value based entirely and parental values estimated from hybrid testing

I would compare this to this statement:

Predict Male Female Hybrid !present Male Female Hybrid

- This gives the hybrid value based on the combination of parental and hybrid effects estimated by the model.

The difference between these two BLUPs is the SCA component. With this difference, you could compare:

- Correlation between your BPH terms and the SCA term.
- Correlation between additive and dominance model GERP scores and SCA and GCA BLUPs.
- Whether SCA and BPH follow similar trends for specific hybrid crosses. For example, is there more SCA between SS x NSS crosses. If you are interested, I can help you find publically available information concerning what is a SS and a NSS.

Let me know if you need some help with ASReml syntax and modeling.

- **GERP/SCA/GCA Modelling:** I really enjoy this statistical modeling step because it represents a true fusion between quantitative and population genetics. It would be interesting to compare the change in log-likelihood between each of these models. Particularly, I would be interested in the change of log-likelihoods between model 2(GERP breeding value), model 3 (SCA added), and model 4 (SCA + GERP breeding value). No significant change between adding the GERP breeding value and the SCA term would indicate that you are capturing the same variability. It would also be interesting to list and compare the variance components that were developed in each of these models to examine if the GERP score and SCA term capture the same variability. Again, this could be fit directly within the phenotypic analysis.