Machine learning for genomic prediction in *Lolium perenne*

Nastasiya F. Grinberg, Alan Lovatt, Matt Hegarty, Andi Macfarlane, Tina Blackmore, Kirsten Skøt, Ian Armstead, Ross King, Wayne Powell and Leif Skøt

PAG XXIII, January 2015

Data for genomic predictions

- Training population
  - Intermediate flowering generations – F11, F12, F13, F14 (370 genotypes)
  - Late flowering – F5 (170 genotypes)
  - Phenotypic data of half sib progeny in plot trials
  - Genotypic data from 4K SNP Array
  - Traits – ground cover \( y_2 \), Yield \( y_1 \) & \( y_2 \), digestibility, water soluble carbohydrates, n

Methods tried

1. rrBLUP
2. Random forest (rf)
3. LASSO
4. Gradient boost machine

Results

1. Correlation between observed and predicted values
   \[ R^2 \]
   2. \( R^2 \) (error related to how well the model predicts actual value compared to observed values. The further away from observed values you go, the bigger the mean squared error is (which is used to calculate \( R^2 \)).
3. It actually tells you how variation in predicted values compared to that in observed values. Hence, we can talk about ‘variance explained by the model’.

Broad sense heritabilities

<table>
<thead>
<tr>
<th>Population</th>
<th>gc_y2</th>
<th>Yield_y1</th>
<th>Yield_y2</th>
<th>DMD</th>
<th>WSC</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int F13</td>
<td>0.22</td>
<td>0.71</td>
<td>0.67</td>
<td>0.82</td>
<td>0.73</td>
<td>0.67</td>
</tr>
<tr>
<td>Late F5</td>
<td>0.15</td>
<td>0.08</td>
<td>0.57</td>
<td>0.39</td>
<td>0.61</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Population structure

![Population structure graph](image)

Linkage disequilibrium

![Linkage disequilibrium graph](image)

Conclusions

- Close relationship between training and test population improves correlation
- Training population size improves correlation
- Traits related to forage quality perform better than yield and ground cover
- rrBLUP performs as well as or better than ML methods
- Maybe very large training populations will improve performance of ML