NIAB Diverse MAGIC: dissecting trait genetic architecture across 70 years of wheat breeding21 January 2021

— — —

Q&A session

Presenter: Michael Scott, Postdoctoral Research Associate, University College London, UK

The webinar recording is available on the IWGSC YouTube channel at https://youtu.be/5lPyjn2NOL0

Q; What tool did you use for variant detection?

Answer Given (timestamp 48:20). In short, we used a GATK pipeline for capture data and then used some filters commonly applied to wheat. Thank you for your question!

Q: How the genome wide QTLs are identified in multiple years study?

In general, we didn't combine the data for the same phenotype in different years. In the case of protein-yield deviation, the manhattan plot shows composite p-values across both years obtained after using Fisher's method for combining p-values. In general, we performed QTL detection at the level of both SNPs and founder haplotypes and we used permutation to determine genomewide significance thresholds. Thank you for your question!

Q; What is the minimum value of genomic prediction to be considered good?

Answer Given (Timestamp 45:45). One thing I will add to my answer here is that I gave prediction accuracy as the correlation coefficient between predicted values and phenotype. Sometimes this is divided by the sqrt(heritability) of the trait because the maximum prediction accuracy is the variance explained by genetic versus environmental factors (which is the heritability). If much of the phenotypic variation is explained by the environment, genomic prediction won't ever be able to predict the phenotype with high accuracy. This might be another factor to consider when deciding on an acceptable genomic prediction accuracy. Thank you for your question!

Q: How to determine which model to use for prediction?

We used three models for genomic prediction: ridge regression, elastic net, and LASSO, but I only showed results for LASSO. LASSO and elastic net has very similar prediction accuracies but LASSO used fewer SNPs in the prediction model. On average, the prediction accuracies were higher for LASSO (and elastic net) than for ridge regression. However, ridge regression may have been slightly more suitable for highly polygenic traits where there was no genomewide significant QTLs detected. You can see this in more detail in the preprint. Thank you for your question!

Q: Why is Maris Huntsman not among the founders, it has been such a corenerstone cultivar for wheat breeding

That is true! As I explained in response to another question, the founder selection preceeded my involvement in the project. However, they followed a selection algorithm based on genotypes from a

panel of varieties that was available at the time. The algorithm was designed to capture as much genetic diversity as possible. Therefore, I assume that Maris Huntsman was either not in the panel or not chosen by the algorithm. Perhaps a variety that is in the pedigree of many other varieties is unlikely to be selected for inclusion because a large fraction of its genetic diversity is likely to be captured by varieties that descend from it. Thank you for your question!

Q: Maybe provocative: could you just go ahead and keep in improving this population in recurrent selection scheme and still achieve progress and select better cultivars, or in other word why do breeders still perform 100s of crosses every year?

I think it would be very interesting to perform recurrent selection in this population. However, it would probably take several years to catch up to current elite varieties in terms of yield. Performing selection recurrently within this population would still require crosses to be made between lines. One approach I think would be interesting here would be to target the founders of future generations such that you capture complementary (different) beneficial alleles, such that you retain as much functional variation as possible. To the extent that you trust the genomic prediction models to capture truly functional variation, this allows improvement to continue from within this population for as long as possible. Thank you for your question! I hope that I have understood and answered correctly.

Q: You mentioned you tested your genomic prediction on an independent set of lines. Can you give more information about these lines, have they been evaluated in the same environments as the MAGIC (training) or different ones?

Answer Given (timestamp 53:20).

Q; Cross validation within a same dataset are always more optimistic. I have been criticized for having done this way (but many do!)

Yes, we expect genomic prediction to be most accurate in this case. Of course, the appropriate method depends on the goals. Here, I mainly argued that genomic prediction was a tool to understand the segregating variation in the population itself, rather than advocating directly for selection in a different population based on the genomic prediction models in this population.

Q: How were the haplotypes called at the gene region?

Answer Given (Timestamp 43:15). In short, this was done by complete-link-clustering of the SNP genotypes for the founders within each gene.

Q: Can you explain how you make difference between HC and LC colocalizations QTL with know genes?

Answer Give (timestamp 56:30). In short, this is a manual description of our ability to overlap the previous reports with our physical map locations.

Q: Could you explain more how you guess the missing sequences of the progenies without those of the founders?

Answer Given (timestamp 54:50). In short, we used STITCH software, which is described in doi:10.1038/ng.3594

Q: When you select the founders, do you have some traits already in your mind or just take random cultivars regardless of specific traits/genes they are known for?

Answer Given (Timestamp 44:30) In short, the founders were selected for genetic diversity and not with traits in mind. However, common variation for key traits is captured, such as vernalisation requirement. Thank you for your question!

Q: I thought the opposite of the result you presented about the trade between yield and protein in current cultivars. Because these days farmers sell wheat grain based on the protein premium of their produce.

This is an interesting thought, but I don't know enough about the dynamics between breeders and farmers to comment confidently. To speculate: it would be interesting if farmers chose to grow older varieties for their protein % rather than more modern ones! However, I suspect that some modern varieties have been bred for protein content and they just haven't been included in our sample. Thank you for your question! Please let me know if you have any further insight into this!

Q: Would future phenotyping necessitate using all 500 RILs? Could one use a subset of 150 lines? Have they been grown anywhere outside the UK?

Answer Given (timestamp 51:30). In short, there is a trade-off between phenotyping effort and power. 150 lines means that you would expect <10 to have ancestry from each founder at each locus. This might be too low to confidently assess alleles that are private to one founder, but you might have more success if the causal allele is shared between a few founders. Detection probability also depends on the size of the phenotypic effect of any causal allele, which is unknown at the beginning of most studies. In a power analysis, you could decide on an acceptable effect size that you are looking for based on biological relevance. Thank you for your question!

Q: Would it be possible to work with this newly develop 16 parents MAGIC populations? If so, what will be the procedure?

Yes, the germplasm and data are freely available if you wish to work with this population. Please visit the website for information: http://mtweb.cs.ucl.ac.uk/mus/www/MAGICdiverse/index.html . For germplasm, please contact James Cockram at NIAB.

Q: which approach is Powerful for loci locating between GWAS and MAGIC population approach ?

One advantage of a MAGIC population over a GWAS is that you avoid issues of confounding population structure, which may weaken your analysis. Ignoring this issue, we generally expect GWAS with similar sample sizes to have greater mapping resolution due to the natural recombination events captured. Whether you have more detection power between MAGIC and GWAS (again ignoring population structure) depends on the population allele frequency. As I mentioned, no alleles are rare in MAGIC populations and this slightly increases the probability of detecting a QTL for a rare causal allele. However, it is slightly more likely that you will not capture a common variant at high frequency in a MAGIC population versus a GWAS panel, so the detection probability is slightly lower for common variants. That rough description is based on the power analysis I showed where I assume you choose founders or GWAS samples randomly from a source population of interest. However, collections used for GWAS often have biased sampling and MAGIC population founders may also be chosen with some bias in mind (in our case maximising diversity). For more discussion you could look at our review of multi-parent populations: https://doi.org/10.1038/s41437-020-0336-6 . Thank you for your question!

Q: Could you explain a bit more about the part about breaking off the trade-off traits?

I showed a negative relationship between yield and protein % in both the founders and inbred MAGIC lines. We were interested in identifying genetic variation that tends move away from this trend line. Therefore, we created a composite measurement, Protein-Yield-Deviation (PYD). PYD is given by the perpendicular distance away from a symmetrical (Thiel-Sen) regression between protein content and yield. This means high PYD lines tend exceed the regression and have higher yield than would be generally expected given their protein content and vice versa. We performed genomic prediction on this composite PYD trait. We also noticed that awned lines tend to have higher PYD. The awns locus also reaches genomewide significance for affecting PYD in QTL mapping, but only if we combine p-

values for PYD from both years. In general, the idea is to target selection at both traits simultaneously. Thank you for your question!

Q: can you explain the heritablity and R2 plot a bit more?

Answer Given (timestamp 46:45). In short, heritability is a measure of the total fraction of phenotypic variance that is attributable to genetic variation versus environmental variation. The R2 is calculated from the genomewide significant QTLs only.

Q: How you made genetic prediction models?

I partly answered this question in a response above. In short, we used three models for genomic prediction: ridge regression, elastic net, and LASSO, but I only showed results for LASSO. We implemented the genomic prediction models using the glmnet R package. Thank you for your question!

Q: why mostly heritable variation(like height) can not be measured by QTL?

This reflects the fact that many loci are apparently affecting the trait but only by a small amount. This means that they won't be detected by the QTL analysis because there isn't sufficient power to detect alleles with small effects on the phenotype. Genomic prediction is one approach to try and estimate the effect of genomewide loci of small effect outside of the QTL that reach genomewide significance. Thank you for your question!

Q: Was the goal to find markers linked to complex traits or develop lines for breeding?

Answer Given (timestamp 49:58).