Novel design of imputation-enabled SNP arrays for dual hybridization of wheat and barley

15 April 2021

— — —

Q&A session

Presenter: Gabriel Keeble-Gagnère (Agriculture Victoria Research, Australia)

The webinar recording is available on the IWGSC YouTube channel at https://youtu.be/A1Qvl8lleVU

Q: Let's say I would like to use your chip, how much will it cost me per genotype or assay more or less?"

Please contact the following Illumina representatives for details:

- America: Kahlil Lawless klawless@illumina.com
- EU, Middle East and Africa: Andre Eggen aeggen@illumina.com
- China and Taiwan: Youpei Cao ycao1@illumina.com
- Asia-Pacific and Japan: Terry Loo tloo@illumina.com

Q: Very interesting to see the concordance in results between single and double hybs. Did you need to standardise the DNA concentration of the samples combined or was there a degree of tolerance of different DNA concentrations between the samples

Timestamp: 45'25"

Most of the barley and wheat SNP in dual hybridisation assays produced scorable cluster patterns. Two factors were important to achieve this. First, we empirically determined that 800 ng genomic DNA (rather than the standard 200 ng) was required for the assay, with the input DNA for each sample matching the ratio of the genome size for each species; e.g. 200 ng barley DNA and 600 ng wheat DNA. And second, the Illumina supplied species-specific SNP manifest file should be used to create the GenomeStudio project for SNP clustering and genotype calling. This meant for a dual hybridsation assay, the creation of two GenomeStudio projects was required, one for wheat using the wheat SNP manifest file, and the other for barley using the barley-specific SNP manifest file. Using this approach, we have successfully run dual sample hybridization assays across tens of thousands of samples.

Q: I want to study diversity of Ethiopian durum wheat landraces for drought tolerance using SNP markers. which genotyping platform more reliable for my study? also I want to incorporate GWAS study.

Timestamp: 45'55"

We were previously involved in an international collaboration that used the Illumina wheat 90K SNP array (Wang et al. 2014 Plant Biotech J) to investigate global diversity among 1,856 tetraploid wheat accessions including wild emmer from the North Eastern and Southern Levant Fertile Crescent, domesticated emmer wheat, domesticated emmer wheat from Ethiopia, durum wheat landraces and

durum wheat cultivars. The results of this study were reported in Maccaferri 2019 Nat Genet. The wheat-barley 40K SNP array contains the subset of SNP on the 90K wheat array – 2,609 in SNP in total – that differentiated: 1) the top 2% Fst values in Maccaferri et al. (2019) between the four subgroups of tetraploid species: wild emmer, domesticated emmer, domesticated wild emmer, durum landraces and durum cultivars; 2) subgroup-specific private SNP that showed a MAF ≥ 0.1 in one of the subgroups and were either monomorphic or showed a MAF <0.05 in the other subgroups; 3) subgroup-specific high MAF SNP that were present at ≥0.3 MAF in any one of the subgroups; and 4) neutral SNP that did not show any signatures of selection, were polymorphic in all subgroups and showed an overall MAF of ≥0.4. These 2,609 SNP differentiate the tetraploid species subgroups as efficiently as the Infinium wheat 90K array. In this context, we believe that the wheat-barley 40K SNP array provides near identical resolution for investigating genetic diversity within Ethiopian durum landraces as the wheat 90K SNP array.

Q: How is this chip better than GBS and how did you make sure you limited ascertainment bias?
Also, In diversity panels, how well were you able to impute MAFs?

Timestamp: 48'50"

The wheat-barley 40K SNP array is better than GBS for several reasons. First, the array typically produces >95% call rate and has exceptionally high calling accuracy for heterozygous loci. The array is therefore well suited for early generation material. Second, the assay turnaround time is three days and is highly scalable (i.e. can run from 24 or 96 samples to thousands of samples at a time), which makes it suitable for implementation in breeding, as well as research applications.

To minimize ascertainment bias, the selection of SNP for inclusion as content on the wheat-barley 40K SNP array was based on SNP discovered in maximally diverse and representative worldwide germplasm that included landraces, varieties, and novel trait donor and historical breeding lines. In addition, the algorithm used to select tSNP for imputation was MAF agnostic. This is reflected in the very similar MAF distribution profiles for the tSNP, impute SNP and wheat exome and barley whole genome sequence SNP (Slide 21). While this process does not completely remove ascertainment bias, it helped to minimize it and therefore provide the best compromise between a SNP array and GBS assay, which is typically ascertainment bias free.

Using 100-fold cross validation, we were able to achieve >97% accuracy (as measured by both correlation and concordance between the imputed and actual SNP genotypes) for imputing the set of SNP tagged at $r2$ ≥0.9 (inclusive of heterozygous calls) in both globally diverse wheat and barley germplasm. Importantly, imputation accuracy was also high for the set of SNP tagged at $r2$ ≥0.5.

Q: Apart from manual scoring, which scoring algorithm did you use for allele discrimination?

Timestamp: 50'24"

SNP clustering and allele calling was performed using GenomeStudio Polyploid software (Illumina Ltd) using the Illumina supplied wheat or barley SNP manifest file. A modified version of the custom genotype calling pipeline described in Maccaferri et al. 2019 Nat Genet was also used.

Q: With dual Hyb, is there a line effect? i.e. Do certain cultivars/landraces of one species "ruin" calling for the mate in the pair?

Timestamp: 52'00"

We have performed tens of thousands of assays and have not observed this to be an issue. Given the SNP probe sequences were selected to align uniquely to the target genome and not align to the other genome – i.e. a wheat SNP probe had to align uniquely to the wheat genome and not to the barley genome, and vice versa – we don't not expect this to occur frequently. We would expect problems, if for example, a dual sample hybridization assay was performed using a barley sample and wheat-barley chromosome addition line since the presence of the barley chromosome in the latter sample would confound the signal for barley SNP targeting that chromosome.

Q: In SNP chip genotyping data how to consider gap or null allele (represented as __)? How to proceed after Identifying a particular SNP if we want to get the details of the loci?

Timestamp 52'50"

As SNP on the wheat-barley 40K array were selected to be single-locus, they typically behave as single dose markers having cluster positions at Theta values of about 0 and 1 (Slide 36). For single-dose marker, null alleles (when present) are detected as 'no signal' or failed assays. When null alleles occur at low frequency in a population, they can be difficult to discern from failed reactions. Typically, evidence for a null allele must be considered in the context of its frequency in the population and surrounding SNP. In the case of larger sized presence-absence structural variations, several SNP will be detected as 'no signal' if multiple SNP tag the structural variant. The detection of null alleles on wheat-barley 40K SNP array is further aided by the availability of the physical position of the targeted SNP locus in the wheat or barley genome.

Q: How does the array work on introgressions from wild species (ventricosa, thynopyrum, and others)?

Timestamp: 54'00"

One of the advantages of the Infinium assay chemistry is that it is based on the hybridization and extension of a 50-mer probe sequence to a target SNP locus. The combination of probe length and relatively low hybridization temperature used in the assay means probes can hybridise to related loci including homoeologs, paralogues and orthologues (Wang et al. 2014 Plant Biotech J). When extended, the SNP probes create signal which can be scored. Hence, we would expect a subset of the SNP on the wheat-barley 40K SNP array to generate scorable cluster patterns in wild wheat species. However, we would also expect strong ascertainment bias since the array content did not include SNP discovered from wild species. That said, the wheat-barley 40K array would be particularly useful for tracking introgressions from wild species into wheat or barley since in this context ascertainment bias is not an issue.

Q: Since LD is population dependent what do you think is the drawback of utilizing this approach in diverse panels such as landraces which were not part of the design to start with?

Timestamp: 55'05"

Yes, LD is population dependent and can be a drawback for the approach we described to select SNP for the wheat-barley 40K SNP array. Being aware of this limitation, we ensured we used SNP discovered in geographically diverse wheat and barley germplasm that included landraces, varieties, synthetic derivatives and novel trait donor lines.

Q: Is there a chance the multi-species pulse chip will be made available to the public?

Timestamp: 57'54"

The multi-species pulse 30K SNP array is currently not publicly available for purchase from Illumina. Organisations interested in accessing the array for research-purposes should contact Matt Hayden, matthew.hayden@agriculture.vic.gov.au

Will the exome data for the reference panel be made available to the public for imputing new lines genotyped with this array?

Timestamp: 57'25"

The wheat-barley 40K SNP array can be directly purchased from Illumina. The wheat and barley exome SNP data underpinning the described imputation capability of the array is already in the public domain. The associated SNP genotype calls, along with genotype calls for the value-add SNP content, will be published in the associated journal manuscript which is currently submitted.