SURVEYING WHEAT HAPLOTYPE DIVERSITY FOR TARGETED BREEDING

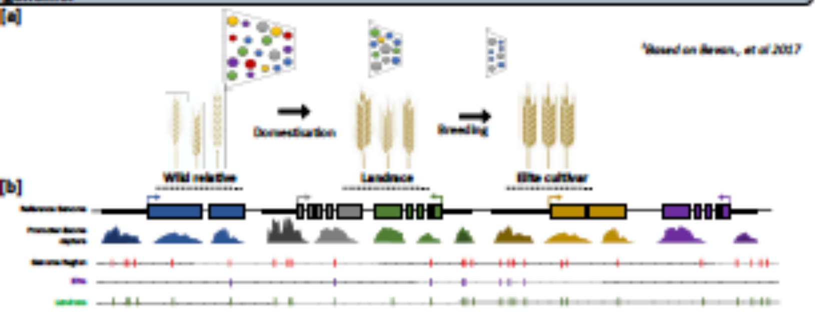
Jesús Quiroz-Chávez , Ricardo Ramírez-González, Kumar Gaurav, Brande Wulff, Cristóbal Uauy

John Innes Centre, Norwich Research Park, Norwich, UK.

Rationale

Often, wheat breeding programs exploits limited genetic variation in elite cultivars which restrains gains achieved for yield. Landraces is a source of unexplored allelic variation. Most of wheat genetic variation has been studied within gene-coding regions or specific known SNP polymorphisms which are usually employed individually for genome-phenotype association analysis. The combination of two or more of these variants inherited together as haplotypes can strengthen association analysis. Methods for haplotype building rely on the use of SNPs mainly identified by alignment methods. The use of *k*-mers as a variant calling have the potential of capture structural variations in addition to SNP and detect the real state of a haplotype.

Fig. 1. Reduced genetic variation in commercial wheat and haplotypes distribution across the genome.



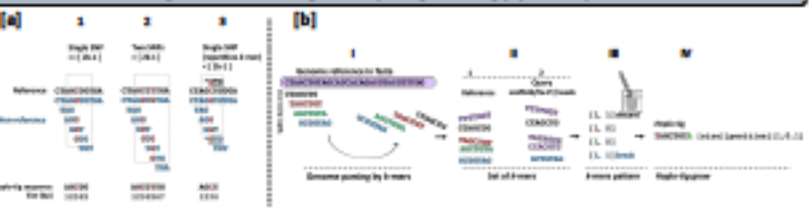
¹Based on Bevan,, et al 2017

[a] Allele diversity (colored circles) was reduced during wheat domestication. A subset of this diversity remains in landrace collections while limited variation is exploited in elite cultivars. [b] Hypothetical genetic variation across the wheat genome.

We hypothesize that:

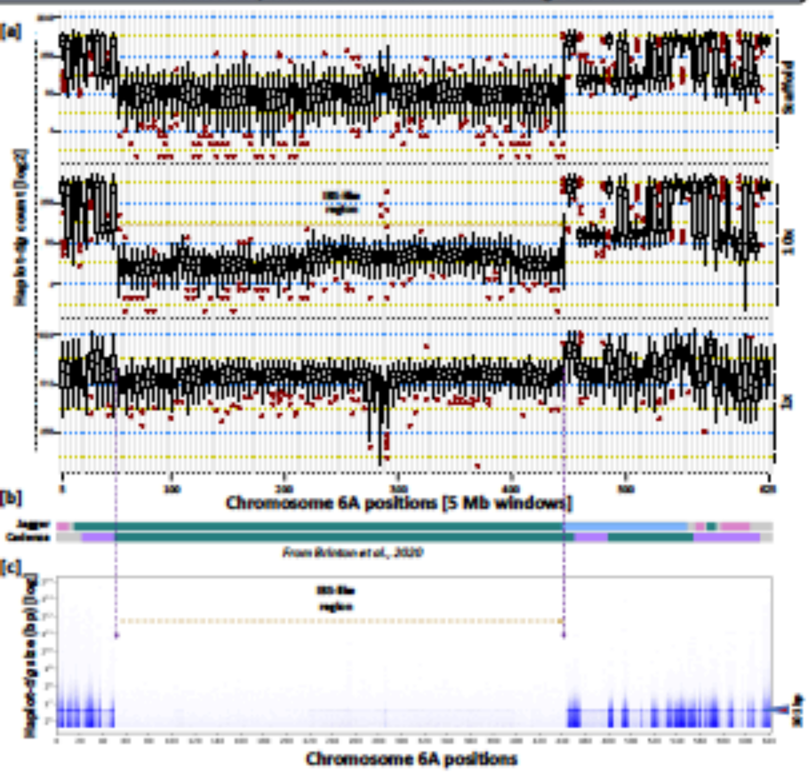
1. The use of *k*-mers for variant calling and haplotype building will allow capture allelic variation in difficult to align genome regions such as non-coding gene sequences.
2. A haplotype database representative of UK varieties can be created to impute haplotypes from skim sequencing.
3. Unexploited haplotypes represent a valuable source of variation for breeding.

Fig. 3. Variant calling and haplo-*tig* building pipeline by *k*-mers



In this project we define haplo-tig as a representation of a variant. This variant can encompass a single or multiple SNPs, InDels or a combination of both. [a] Possible haplo-tig size outputs regarding variation type. (1), ==2k-1, when a single SNP is present, (2) > 2k-1 when two or more SNPs occur within the interval of the *k*-mer size (k), and (3) < 2k-1 when a repetitive *k*-mer from a different region is mapped into the reference sequence before the haplo-tig is expanded. Additional larger and shorter haplo-tigs are generated by InDels or missing assemblies in the reference. [b] *k*-mer sequence parsing and haplo-tig elongation along the genome. (I) The genome reference is split in *k*-mers of *k* size. After, (II) *k*-mers of the reference and query samples are compared for presence [1] absence [0]. If a *k*-mer is present in the two samples, it is discarded, otherwise kept for haplo-tig extension until a *k*-mer in both samples are matched across the genome reference. High haplo-tig count indicates high variation between samples. Conversely, low haplo-tig count, low variation.

Fig. 4. Results. *k*-mer pipeline is capable to detect Identical By State (IBS)-like regions using either chromosome, scaffold assemblies as well as using RAW reads.

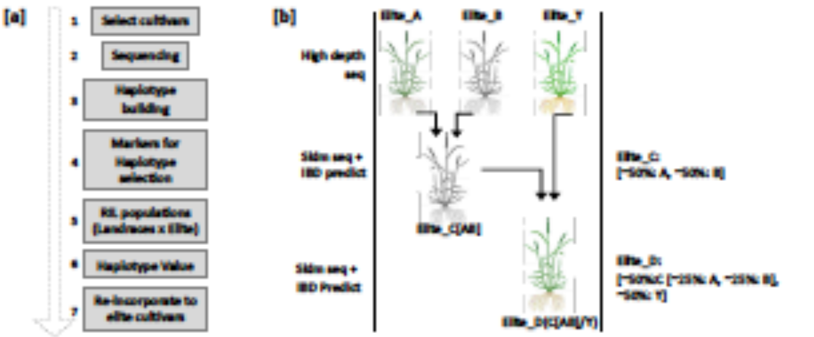


[a] Haplo-tig count (y-axis, log2 scale) across Chr6A (x-axis, in Mbp) of scaffold assembly (top), 10x (middle), and 1x (bottom) raw reads of Cadenza. Chromosome assembly of Jagger as a reference (Walkowiak *et al.*, 2020). With 10x raw reads the pipeline performed similar than chromosome assemblies to detect IBS regions (middle plot, yellow arrow). [b] Haplotype region shared between Jagger and Cadenza defined in Brinton *et al.*, 2020 is indicated by purple arrows and the green color of the bars which is the region matched in our analysis as IBS-like. [c] Haplo-tig size (y-axis) in bp across the Chr6A positions. Large haplo-tigs are located at chr telomeres as indicative of high variation. Note that there are a few large haplo-tigs in the IBS-like region (yellow arrow), which is indicative of a true shared haplotype block. The red triangle indicates 101 bp haplo-tigs size, which represent a single SNP.

Objective

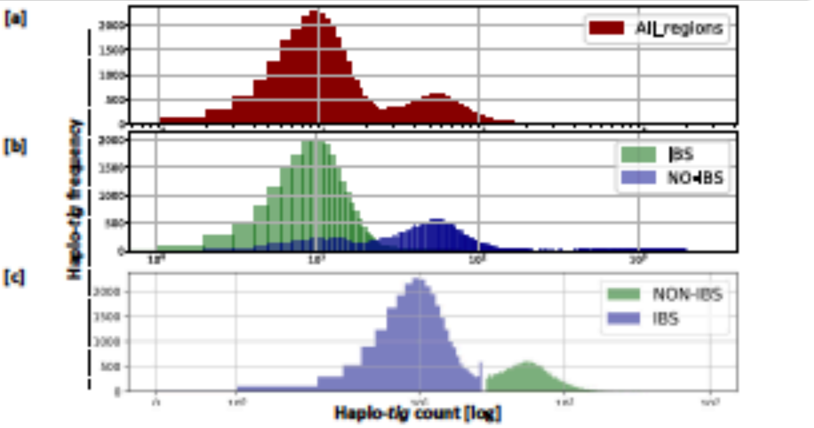
The aim of this study is to design a novel approach to capture genome variation by a haplotype-based method employing *k*-mers on promoter-exome capture, Whole Genome Sequences (WGS) and pedigree information. We are exploring UK Recommended List (RL) varieties and the Watkins collection. To test our model we are levering the pan-genome assemblies and Illumina short raw reads. At the end of the project we hope to build an UK haplotype database that will serve as a reference for anchoring skim sequences across the whole genome and will provide valuable information for the breeding community.

Fig. 2. Main steps of the pipeline for haplotype building and re-incorporation of allelic variation into commercial lines.



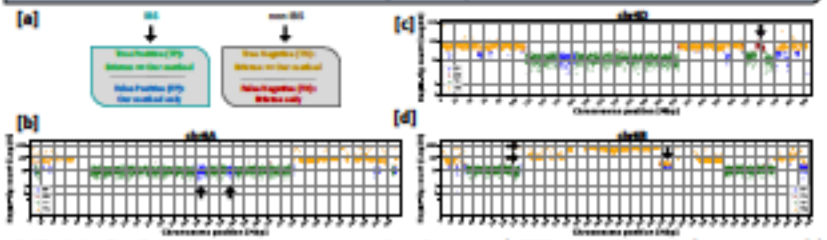
[a] Project general framework. (1) First, select cultivars (Elite and Landraces) and (2) sequencing method (promoter-exome capture and WGS. (3) Raw reads and variants for haplotype building. (4) Haplotype markers to select and test phenotype in (5) Landrace x Elite populations (F4:5) in the field and determine their genetic value (6). Further haplotype validation and re-introgression into commercial cultivars (7) can be achieved by other methods. [b] Strategy to define sequencing depth per line/cultivar in order to capture representative haplotypes from a specific population.

Fig. 5. Results. Whole genome haplo-*tig* distribution shape and categorization



[a] Whole genome haplo-tig frequency of Mattis vs Claire 10x raw reads. Frequency of the haplo-tig counts (y-axis) within 200 Kbp windows (x-axis). The single red mixture of two distributions was created with all haplo-tigs captured across the whole genome. We believe that low-count (left) and high-count (right) distributions belong to the IBS-like and NON-IBS regions, respectively. In plot [b], we utilized haplotypes defined by Brinton *et al.*, 2020 to determine in which of the two distributions the IBS and non-IBS are located. The green (IBS) and blue (non-IBS) are the regions that matched with our haplo-tig sequences and were categorized by Brinton. The overlapping points between the two colors is the data that was miss-categorized by either, our approach or Brinton. [c] *De novo* IBS and non-IBS classification. We used the Gaussian Mixture Model (GMM) of scikit-learn which implements the Expectation-Maximization (EM) algorithm to assign IBS regions to our data.

Fig. 6. Results. *De novo* haplotype blocks determination and validation by Precision and Recall across the whole genome by chromosome



[a] Precision-Recall outputs. Using haplotype regions from Brinton *et al.*, 2020 as a positive control to test our model we expected four outputs. IBS regions comparison generates TP and FP when the two methods agree and when only our method call an IBS region, respectively. Similarly, non-IBS generates TN or FN when the two methods and Brinton only call a non-IBS region, respectively. In [b], [c], and [d], the Mattis pan-genome reference was tested against 10x raw reads of Claire as a query. Haplo-tig count (y-axis) within 200 kb across chromosome physical positions (x-axis in Mbp). As described in [a], green (TP) blue (FP), yellow (TN), and red (FN). Roughly ~75% of the data agree between the two approaches with accuracy above ~80% (data not shown). Note in chr6A (purple arrows) IBS regions identified in our approach (bule dots) that was missing in Brinton et al., 2020. From ~100 to 440 Mbp these two samples share one large IBS block. In chr4D (arrow) we detected an FN region which was called IBS by Brinton which is clearly non-IBS in our approach since the level of variation is slightly higher than the threshold. These two samples share less IBS regions in chr6B. A problematic region is indicated by the arrow where our model is no yet able to categorize as IBS or non-IBS. The orange arrows point to two more levels of variation detected between these two samples which may be indicative of re-introgression, or conservation of ancestral variation.

Conclusions

1. Our approach can detect variations using *k*-mers in a similar way to alignment-based methods and have the potential to capture structural variations in haplotypes.
2. We detected IBS-like regions and identified high variation and low variation regions which are located in telomeric and centromeric regions, respectively.
3. *De novo* haplotypes blocks were identified which include novel regions not detected with alignment-based methods.
4. Finally, our method detects haplotypes using raw reads with low sequencing depth (~10x), overcoming the scaffold level assemblies.

Future work

Immediate analysis will be focused to define haplotypes from multiple samples from the UK RL varieties and the Watkin collection to create a haplotype database and test the imputation of genome region form skim sequences. Later in the project we hope to associate those haplotypes to phenotypes from the field which have been collected since 2017 to the present.

1. Bevan, M. W., C. Uauy, B. B. H. Wulff, J. Zhou, K. Krasileva, and M. D. Clark. 2017. 'Genomic innovation for crop improvement', Nature, 543: 346-54
2. Walkowiak *et al.*, 2020. Multiple wheat genomes reveal global variation in modern breeding.
3. Brinton *et al.*, 2020. A haplotype-led approach to increase the precision of wheat breeding.
4. Acknowledgements to Dr,. Brinton for providing with haplotype files and scripts from the alignment method.

Jesus.Quiroz-Chavez@jic.ac.uk