Comprehensive and cost-effective genomic regulatory element sequencing for wheat

Junli Zhang[1], German Burguener[1], Juan Debernardi[1], Frédéric Choulet[2], Etienne Paux[3], Jacob Enk[4], and Jorge Dubcovsky[1,5]

1. University of California, Davis, CA, USA
2. GDEC, Université Clermont Auvergne, INRAE, Clermont-Ferrand, France
3. VetAgro Sup, Lempdes, France; formerly INRAE
4. Daicel Arbor Biosciences, Ann Arbor, MI, USA
5. Howard Hughes Medical Institute, Chevy Chase, MD, USA

As genome resources for wheat expand at a rapid pace, it is important to update targeted sequencing tools like hybridization capture panels to incorporate improved sequence assemblies and regions of previously unknown significance. **Here we developed a regulatory region target enrichment panel for hexaploid and tetraploid wheat**. We used the upstream ~2 Kbp of each annotated gene in the most up-to-date Chinese Spring wheat genome assembly as the primary target source, but supplemented this using homologous sequences from a draft assembly of tetraploid Kronos wheat, and finally also included regions of observed open chromatin state identified with ATAC-seq. To improve specificity compared to similar legacy designs, we aggressively filtered candidate repetitive sequences using a combination of cross-alignment clustering, TREP19 affinity filtration, and kmer frequency capping. Finally, once converted to candidate probes, these were once again filtered for specificity using a standard design pipeline, resulting in a final target space of about 168 Mbp on RefSeq v1.0. Test captures using the probe set on hexaploid, tetraploid, and diploid wheat exhibit excellent coverage of the target with significantly improved specificity compared to captures performed using a legacy probe design. Captures of 24 lines from the Kronos TILLING population (EMS induced mutations) detected an average of ~3300 mutations per line (~5 million predicted mutations in the complete TILLING population of 1500 lines). The probe set is publicly available through Daicel Arbor Biosciences as either a **stand-alone capture kit or a full service option** from library prep through secondary bioinformatics analysis.

TARGET SELECTION AND PROBE DESIGN

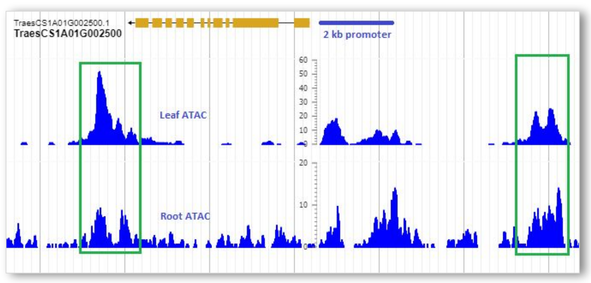
1 2 Kbp upstream of transcripts in RefSeq v1.0

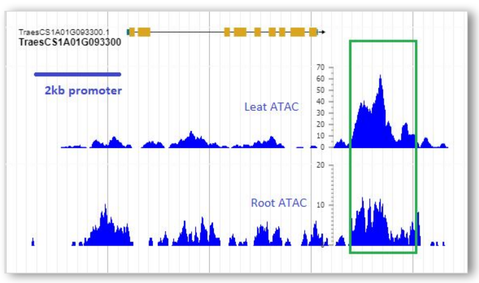
Example window on chr2A. **Green:** initial putative promoter target regions. **Blue lines:** annotation set v1.1 high confidence genes. **Blue blocks:** annotated exons. **Orange:** target regions in the Daicel Arbor Biosciences Expert Wheat Exome v1.0 myBaits kit. Note the partial overlap of Exome and starting promoters targets.

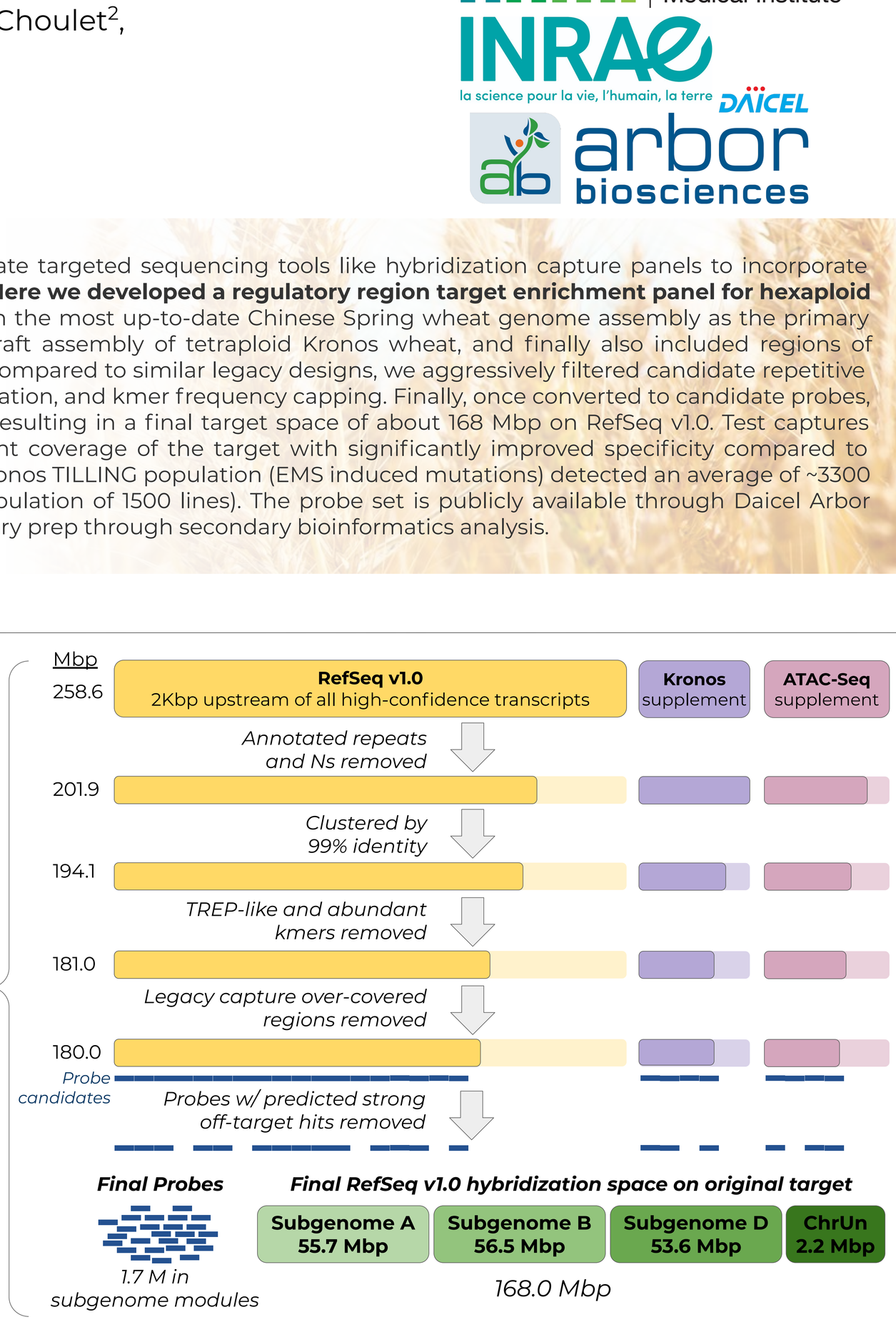
2 Kronos-specific candidate promoter sequences

3 Open chromatin regions identified with ATAC-Seq

Several genome regions showed consistent and significant high coverage when assayed with open-chromatin-specific sequencing (ATAC-Seq). Examples of such regions outside of putative annotation-derived promoter areas are in green boxes below. This comprised a final 23.49 Mbp consensus among several leaf- and root-derived ATAC-Seq datasets that were added to the starting target set to more comprehensively cover potential regulatory elements.



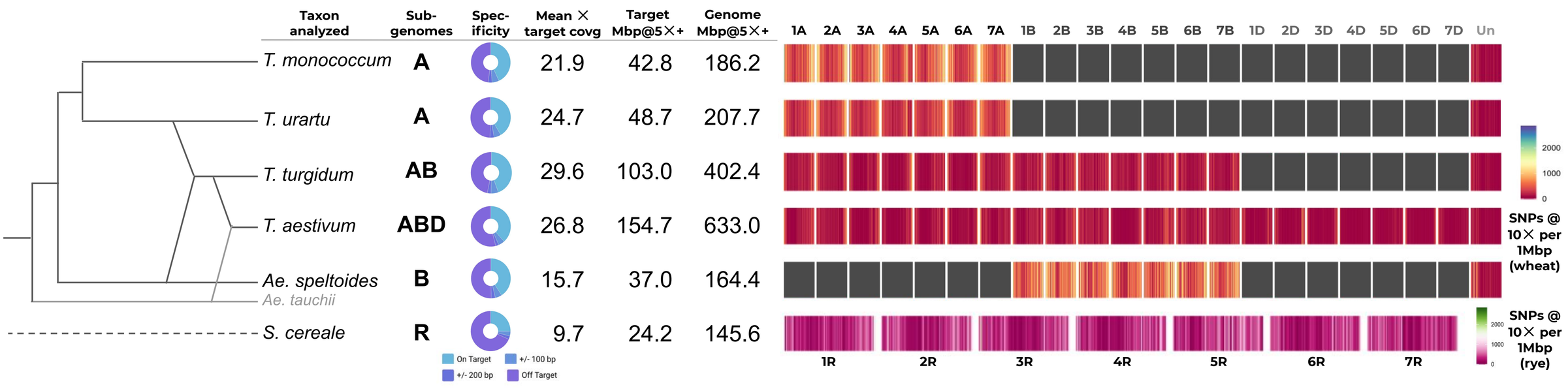




PROBE SET PERFORMANCE

We converted dozens of hexaploid, tetraploid, and diploid wheat and rye genomic DNAs to Illumina sequencing libraries and performed various enrichment reactions using different combinations of the probe modules. For hexaploid *T. aestivum*, Kronos-specific probes were excluded from the probe captures. For tetraploid *T. durum* captures, subgenome D-specific probes were excluded. For diploids and rye, however, we captured several taxa in the same pools, so included all probe modules (including subgenomes D and Kronos-specific). Hexaploids pooled 8 libraries, 1 µg each per capture reaction; tetraploids pooled 12 libraries, 750 ng each; diploids pooled 16 libraries, 500 ng each.

After target capture following the same protocol as the myBaits Expert Wheat Exome kit, we sequenced the resulting pools on a NovaSeq 6000 S4 flowcell using PE150 protocol. Before alignment and analysis, we downsampled the reads to 60M pairs for hexaploids, 40M pairs for tetraploids, and 20M pairs for diploids. Then the fastqs were aligned with *bwa mem* in the Curio Genomics platform to either taxon-appropriate subgenome sets of RefSeq v1.0, or to the *Secale cereale*/Weining Rye genome assembly. Variants were called for sites with a minimum of 10× unique read coverage.



IMPROVEMENT OVER PREVIOUS SETS

Two dozen samples included in the trial capture set described above were members of a Kronos TILLING population. These same samples were also previously captured using a wheat promoter probe design described by Gardiner et al. 2019 (https://doi.org/10.1093/gigascience/giz018) and manufactured by Roche NimbleGen. After aggregating all data and then downsampling to comparable total obtained sequencing reads, we observed an improvement in overall target coverage depth as well as in the number of EMS mutations detected using the new Arbor design.

	Arbor	NimbleGen
Total sequenced reads (PE150)	94,316,136	94,316,136
% read duplication	13.9%	35.5%
% reads mapped	93.9%	95.1%
% reads on target	37.7%	25.0%
Coverage in target region	19.4×	12.0×
Total EMS mutations detected	79,762	68,771

Special Thanks To:

Kellye Eversole
Isabelle Caugant
Jonathan Jones
Sebastian Fairhead
The Sainsbury Laboratory

web: www.arborbiosci.com
email: info@arbor.daicel.com
phone: 1-734-998-0751
twitter: @ArborBio