RELEASE NOTES FOR THE OAT cv. SANG REFERENCE GENOME RESOURCES
Date: March 1, 2022

This data release contains the genome sequence assembly (pseudomolecules) and structural annotation (both protein coding genes and TEs/repeats) of hexaploid (6x) *Avena sativa* (cv. Sang), tetraploid (4x) *Avena insularis* (BYU209) and diploid (2x) *Avena longiglumis* (CN58138).
All oat cv. Sang data is provided by ScanOats (PI Nick Sirijovski, https://scanoats.se) in collaboration with the Helmholtz Center Munich (PI Manuel Spannagl, https://www.helmholtz-munich.de/pgsb).
A. insularis and *A. longiglumis* data is provided by Jeff Maughan and Rick Jellen (https://lifesciences.byu.edu/directory/jeff-maughan and https://lifesciences.byu.edu/directory/rick-jellen; contact jeff_maughan@byu.edu) and Nick Tinker (Agriculture and Agri-Food Canada, https://profils-profiles.science.gc.ca/en/profile/nicholas-nick-tinker-phd; contact nick.tinker@agr.gc.ca) and Wubishet Bekele and Yong-Bi Fu (Agriculture and Agri-Food Canada).

TERMS OF USAGE & CITATION:
If you use the available resources, please cite: Kamal, N., Tsardakas Renhuldt, N., Bentzer, J., Gundlach, H., Haberer, G., Juhasj, A., Lux, T., Bose, U., Tye-Din, J., Lang, D., van Gessel, N., Reski, R., Fu, Y.-B., Spégel, P., Ceplitis, A., Himmelbach, A., Waters, A.J., Bekele, W.A., Colgrave, M., Hansson, M., Stein, N., Mayer, K., Jellen, E.N., Maughan, P.J., Tinker, N.A., Mascher, M., Olsson, O., Spannagl, M., and Sirijovski, N. 2022. The mosaic oat genome gives insights into a uniquely healthy cereal crop. Nature (2022) DOI:10.1038/s41586-022-04732-y
https://www.nature.com/articles/s41586-022-04732-y

TYPE OF DATA:
Researchers are able to access the annotation file as a track on the genome browser in GrainGenes. Additional files including a GFF, gene nucleotide fasta, peptide fasta are also available with this download.
List of files and resources available:
a) the genome assembly sequences (pseudomolecules) of *A. sativa* cv Sang (6x), *A. insularis* (4x) and *A. longiglumis* (2x).
b) the gene predictions (CDS, protein/peptide sequences and GFF) for *A. sativa* cv Sang (6x), *A. insularis* (4x) and *A. longiglumis* (2x). Representative/Primary gene models are included in a separate folder for A. sativa cv. Sang v1.1. Please note that LC genes with a CDS ending in an incomplete codon are excluded from the peptide sequence file!
c) the functional annotations, the Trait (TO) and Plant (PO) Ontology annotations, and the manually curated gene models for *A. sativa* cv Sang (6x).
d) the TE/repeat annotations for *A. sativa* cv Sang (6x), *A. insularis* (4x) and *A. longiglumis* (2x).

IMPORTANT NOTES:
- the gene predictions are classified into high- and low-confidence (HC and LC) gene models. For classification criteria please consult the corresponding methods section in the publication.

- representative transcripts (Sang v1.1 only): this folder contains files with one selected, representative transcript per gene locus, in cases where multiple splice variants/isoforms are predicted. This should assist analyses that require a non-redundant set of genes.

- ontologies: we provide a set of different ontology terms for the predicted Sang gene models. This includes gene ontology (GO), plant ontology (PO) and trait ontology (TO) terms. Please note that all ontology terms are electronically inferred.

- Sang version 1.0 and 1.1 gene annotation differences: both annotation versions refer to the same genome assembly sequence (included with this release) and only differ in the addition of manually curated gene models (e.g. for the prolamin gene family) in release version 1.1. Gene identifiers have been kept stable between the two annotation versions wherever possible (that means, some genes may have been removed, merged or added in v1.1).

- additional resources: we provide normalized gene expression counts (TPM) for Sang genes in the "expression_data.csv" file as well as differentially expressed genes (null hypothesis |LFC| <= 1, adjusted $P < 0.01$) between early, mid, and late stages of seed development in the DESeq2 folder. The "orthogroups.tsv" file contains gene groups derived by OrthoFinder for the following species and gene sets:
Z. mays (B73v5), *O. sativa* (IRGSP-1.0), *B. distachyon* (v3.2), *H. vulgare* (Morex v3), *S. cereale* (IRGSC Lo7 v3), hexaploid bread wheat (*T. aestivum*; IWGSC v1.1), hexaploid oat (*A. sativa*, this study), tetraploid *A. insularis*, and the three diploid oat species, *A. longiglumis* and *A. atlantica* as A-lineage and *A. eriantha* as C-lineage representatives.

- correspondence files between oat cv Sang (this study) and cv OT3098 (PepsiCo oat assembly; available from GrainGenes) genome resources: to facilitate the transfer of knowledge and analyses results, as well as to compensate for some of the limitations of short read assemblies (Sang), we generated the following correspondence files:
a) *lowGeneDens.tsv:* regions of reduced gene density between the Sang short read assembly and the OT3098 genome assemblies. For this, we used overlapping sliding windows for each segment in the syntenic framework described below, to identify regions of reduced gene density in the oat cv. Sang assembly.
b) *invSangOT3098.tsv*: locations of small-scale inversions (0.4 to 23.4 Mbp) between the oat cv Sang and OT3098 genome assemblies: A total of 38 inversions was identified and indexed by their (approximate) coordinates. A more detailed explanation of the approaches and thresholds for these analyses are provided in the respective supplementary methods sections included in the genome publication.
c) *geneAssocTableSangOT3098.tsv:* gene association table between oat cv Sang and OT3098 syntenic genes.
Syntenic pairs between the genes predicted in the Sang (v1.1) and OT3098 (v2) genome assemblies were determined. The analysis was complemented by a reciprocal/bidirectional best blast hit (BBH) approach that included the low-confidence (LC) Sang genes in order to establish links between Sang genes on "chromosomeUnknown" and OT3098 genes. This framework established links between Sang genes and the corresponding, syntenic gene(s) in OT3098 (that means not all Sang genes are associated with a corresponding/orthologous OT3098 gene model in this table, only those within syntenic relationships!!). This analysis serves as the backbone to link (Sang <-> OT3098) and contextualise the genes in:
1) 'Chromosome unknown'/unplaced genes -> column 5
2) (syntenic) regions of unbalanced/reduced gene density between Sang and OT3098 -> column 4
3) regions of identified inversions between Sang and OT3098 -> column 3
The identifiers of the blockIDs refer to the regions identified and labeled in a) and b).

In addition, we added information on whether a gene is located within one of the seven major translocations (T1 to T7) identified in our study of hexaploid oat cv Sang (column 6).

CONTACT: In case of questions please contact Nick Sirijovski (nick.sirijovski@tbiokem.lth.se) and Manuel Spannagl (manuel.spannagl@helmholtz-muenchen.de).