Building pangene sets from plant genome alignments confirms presence-absence variation

Bruno Contreras-Moreira[1,2], Shradha Saraf [1], Guy Naamati[1], Sandeep S. Amberkar[3], Paul Flicek[1], Andrew R. Jones[3], Sarah Dyer[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

[2]Laboratorio de Biología Computacional y Estructural, Estación Experimental Aula Dei-CSIC, Zaragoza, 50059, Spain

[3]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, UK

* Corresponding authors: bcontreras@eead.csic.es, sdyer@ebi.ac.uk

Core ideas

1: Whole Genome Alignments capture overlapping gene models and genome segments.

2: A pangene represents homologous collinear gene models from different gene sets.

3: Lift-over can be used to refine gene models and to confirm gene Presence-Absence Variation.

Abbreviations: ACK, Ancestral Crucifer Karyotype; ANI, Average Nucleotide Identity; PAV, Presence Absence Variation; WGA, Whole Genome Alignment.

ABSTRACT

Consistent gene annotation in crops is becoming harder as genomes for new cultivars are frequently published. Gene sets from recently sequenced accessions have different gene identifiers to those on the reference accession, and might be of higher quality due to technical advances. For these reasons there is a need to define pangenes, which represent all known syntenic orthologues for a gene model and can be linked back to the original sources. A pangene set effectively summarizes our current understanding of the coding potential of a crop and can be used to inform gene model annotation in new cultivars. Here we present an approach (get_pangenes) to identify and analyze pangenes that is not biased towards the reference annotation. The method involves computing Whole Genome Alignments (WGA), which are used to estimate gene model overlaps. After a benchmark on *Arabidopsis*, rice, wheat and barley datasets, we find that two different WGA algorithms (minimap2 and GSAlign) produce similar pangene sets. Our results show that pangenes recapitulate known phylogeny-based orthologies while adding extra core gene models in rice. More importantly, get_pangenes can also produce clusters of genome segments (gDNA) that overlap with gene models annotated in other cultivars. By lifting-over CDS sequences, gDNA clusters can help refine gene models across individuals and confirm or reject observed gene Presence-Absence Variation. Documentation and source code are available at https://github.com/Ensembl/plant-scripts/tree/master/pangenes.

INTRODUCTION

For a growing number of crops and plants there are now multiple genome assemblies available in public repositories. These data are driving the analysis of the pangenome, the union of all known genomes of a species. For instance, recently published pangenome reports include staple crops wheat and barley (Jayakodi et al., 2020; Walkowiak et al., 2020). While these efforts have greatly advanced our understanding of the variability of genomes within species, they have also prompted a new class of problems, those related to annotating and naming genes across cultivars. Different strategies are possible. For instance, the barley pangenome consortium lifted-over gene models from three genotypes (Morex, Barke, HOR10350) to all other assemblies. This procedure biases the gene space to that of the reference cultivars. In contrast, in other species, fresh gene annotations have been produced for different individuals or populations (Gordon et al., 2017). In this case, care should be taken to follow the same annotation protocols throughout to avoid inflating the number of clade-specific genes (Weisman et al., 2022) or to conserve gene identifiers.

In this context it is useful to define a pangene, a gene model found in some or all individuals of a species in a similar genomic location. A pangene should integrate additional naming schemes, e.g. so that a cluster of gene models can share a common identifier that links back to their original gene identifiers. A pangene set defines our current understanding of the total coding potential of a species and can assist in gene model curation, by providing a pool of possible gene models for assessment.

Pangenes can be produced by a variety of approaches, such as iterative mapping and assembly (Golicz et al., 2015), local alignments of nucleotide sequences (Contreras-Moreira et al., 2017), molecular phylogenies of chromosome-sorted proteins (Lovell et al., 2022) or as a secondary product of genome graphs (Sheikhizadeh et al., 2016; Li et al., 2020; Guarracino et al., 2022). Whatever the approach, a common use case for pangenes is to capture Presence

Absence Variation (PAV) at the gene level. However, previous work has observed that absent gene models are rarely caused by complete sequence deletions; instead, they might not be expressed in certain conditions or the underlying genomic regions might contain genetic variants such that the criteria for calling a gene model are not satisfied. For instance, sequence variants in introns or splice sites can reduce evidence for a gene model (Lovell et al., 2021).

Here we present an approach to identify and analyze pangenes in sets of plant genomes which can explicitly confirm or reject PAVs by lifting-over gene models on candidate genomic segments. This approach requires computing pairwise Whole Genome Alignments (WGAs), which are then used to estimate gene model overlaps across individuals. Finally, pairs of overlapping genes are iteratively merged to produce pangene clusters. The algorithm produces pangene clusters that are not biased towards the reference annotation and that can optionally be used to refine individual gene model annotation with information from all cultivars. We benchmark this approach on diverse datasets that cover monocots and dicots, as well as small and large genomes.

MATERIALS AND METHODS

Genome sequences and genesets

A total of four datasets were used in this study (Arabidopsis ACK2, rice3, chr1wheat10 and barley20) of increasing size and complexity. They are listed in **Table 1**, where ACK stands for the Ancestral Crucifer Karyotype (Lysak et al., 2016) and chr1wheat10 for chromosome 1 in ten different hexaploid wheats. Note that ACK2 and rice3 include different species (*Arabidopsis lyrata* and *Oryza nivara*); the others include cultivars of the same species. The ranges of Average Nucleotide Identity (ANI) among genomes in each dataset are indicated in **Table 3**.

Protocol for calling pangene clusters based on WGA evidence

The repository https://github.com/Ensembl/plant-scripts/tree/master/pangenes contains the documentation and source code for calling pangenes. The main script (*get_pangenes.pl*), illustrated in **Figure 1**, sequentially runs the scripts *_cut_sequences.pl*, *_collinear_genes.pl* and *_cluster_analysis.pl*. These tasks can be performed serially on a computer (default) but can also run in batches over a high performance computer cluster. cDNA and CDS sequences are cut with GffRead with arguments -w, -y and -x (Pertea & Pertea, 2020). Genomic sequences (gDNA) are cut with *bedtools getfasta* (Quinlan & Hall, 2010). WGAs in PAF format are computed with minimap2 with parameters --cs -x asm20 --secondary=no -r1k,5k (Li, 2018, 2) or GSAlign with parameters -sen -no_vcf -fmt 1 (Lin & Hsu, 2020). Unlike minimap2, GSAlign provides ANI estimates. Feature overlap is computed with *bedtools intersect* with parameters -f 0.5 -F 0.5 -e (Quinlan & Hall, 2010) after converting WGAs to BED files, which requires parsing the CIGAR strings contained in PAF files. When features are actual gene models, strandedness is also required (-s). Results can be downloaded at

https://github.com/Ensembl/plant-scripts/releases/download/Dec2022/pangenes_bench.tgz (FASTA files not included).

Pangenome terminology

We will often use pangenome-related terms to describe the pangene clusters output by our protocol. We define occupancy as the number of genomes represented in a cluster. Core clusters contain sequences from all analyzed genomes. Soft-core clusters contain sequences from 95% of the input genomes. Finally, in this paper shell clusters are those with less occupancy than soft-core clusters after excluding singletons (occupancy=1).

Dotplots

The *_dotplot.pl* script can be used to make a genome-wide dotplot of collinear gene models resulting from a pairwise WGA stored in TSV format. This is done in two steps: i) the TSV file is converted to a PAF file, ii) the dotplot is produced with R package *pafr*, available at https://github.com/dwinter/pafr.

BUSCO analysis

In order to evaluate the completeness of the core and soft-core collections of pangenes produced by *get_pangenes.pl*, the corresponding protein FASTA files containing all known isoforms of genes were analyzed with BUSCO v5.4.3 (Manni et al., 2021). The poales_odb10 lineage was selected for all datasets except ACK2, where brassicales_odb10 was used instead.

Venn diagrams

Comparisons of CDS pangene sets produced with both WGA algorithms and the resulting Venn diagrams were carried out with script *compare_clusters.pl* from the GET_HOMOLOGUES-EST software (Contreras-Moreira et al., 2017). As explained in the documentation at https://github.com/Ensembl/plant-scripts/tree/master/pangenes, other scripts from this package can be used to simulate and plot the pangene set growth.

Ensembl orthologues and InterPro annotations

High-confidence orthogroups produced with Ensembl Compara (Herrero et al., 2016) were retrieved with script *ens_syntelogs.pl* (Contreras-Moreira et al., 2022) from Ensembl Plants (Yates et al., 2022). These are derived from phylogenetic trees of aligned protein sequences from most plant species in Ensembl and have extra supporting collinearity evidence. Only pairs of orthologues with Whole Genome Alignment score ≥ 50% and Gene Order Conservation ≥ 75% were taken. This ensures that only collinear genes are included.

For ACK2 and rice3 genomes, annotated InterPro protein domains were retrieved from Biomart (Kinsella et al., 2011). These were taken to curate pangene clusters where all their sequences shared at least one InterPro domain.

Lifting-over gene models on genomic segments

The script *check_evidence.pl* uses precomputed collinearity evidence, stored in a TSV file, and lift-over alignments to project cDNA/CDS sequences on a reference genomic sequence.

Briefly, collinear genome sequences are extracted with *bedtools getfasta* and then pre-clustered cDNA/CDS sequences are mapped to them with GMAP, a software tool that efficiently connects exons while jumping intervening introns (Wu & Watanabe, 2005). GMAP is run with parameters -t 1 -2 -z sense_force -n 1 -F.

RESULTS AND DISCUSSION

A protocol for calling pangenes based on Whole Genome Alignments

The first result from this work is the design of a protocol for calling pangenes in a series of related genomes. Note that a pangene is defined as a gene model found with homologous coordinates in a set of genomes. In order to find pangenes, WGAs are computed, which in turn produce pairs of collinear genomic segments. Collinear evidence is stored in TSV files and can be produced by two WGA algorithms: minimap2 and GSAlign. The protocol is represented as a flowchart in **Figure 1A**. As illustrated schematically in **Figure 1B**, WGAs are used to project the coordinates of gene models across assemblies. By default, a pair of genes are said to be collinear when at least half the length of one matches the other in genomic space (**Figure 1C**). Finally, clusters of genes (pangenes) emerge by merging pairs of collinear genes from different input taxa (**Figure 1D**).

An example collinear region of *Oryza sativa* Japonica group (bottom) and *Oryza nivara* (dataset rice3) as displayed in the Ensembl Plants genome browser is shown in **Figure 2**, together with a summary of the supporting WGA evidence. Besides five 1-to-1 collinear gene pairs, it can be seen that the gene ONIVA01G00100 was mapped to two consecutive *O. sativa* models (Os01g0100100, Os01g0100200) and that two *O.sativa* genes map to unannotated genome segments in *O. nivara*. In a nutshell, this figure shows that WGA evidence allows matching of long genes to split genes if they are collinear, as well as matching annotated gene models to homologous regions in other genomes, even if the genes in question failed to be annotated.

Benchmark on several plant datasets

A systematic benchmark of *get_pangenes.pl* was performed with the four datasets in **Table 1**. In order to describe the performance several variables were collected. As computing WGAs is costly, particularly for large genomes, the maximum amount of RAM consumed by pairwise genome alignments was recorded. The size of the collinear fragments produced by WGA is captured in two variables, N50 and the percentage of fragments that contain blocks of 3 or more genes. Several further variables were also calculated to describe the clusters of collinear genes produced by the protocol: the total number of clusters, the number of pangenes present in all genotypes (core clusters), the pangenes present in 95% of genotypes (soft-core clusters) and the percentage of complete BUSCOs, which are universal single-copy orthologs. The results are summarized on **Table 2** (minimap2) and **Table 3** (GSAlign). Here the outcomes of both algorithms are compared. Although GSAlign consumes more RAM than minimap2 as genome size grows, this is not a fair comparison; in fact, while GSAlign was fed raw genome sequences, minimap2 only completed the WGAs for barley and wheat after masking long geneless genomic regions. The collinear segments aligned by minimap2 are longer and contain more genes than those produced by GSAlign. The dotplots in **Supplemental Figure S1** confirm that the paired genomic segments are generally collinear. **Supplemental Table S1** shows that the number of hits per gene is similar for both algorithms, with minimap2 failing to map more genes than GSAlign. Despite these differences, the numbers of pangenes clustered using WGA evidence from both algorithms are comparable: 91% of all clusters are identical (95% for core clusters, see **Supplemental Figure S2**).

As an extra test, clusters of cDNA and CDS sequences resulting from the rice3 analysis were aligned locally to compute their sequence identity both at the nucleotide and protein level. Note that these clusters contain all isoforms annotated, so often there might be several sequences for the same gene. The results, plotted in **Supplemental Figure S3**, yielded

median sequence identities of 99.6% for nucleotides (cDNA and CDS) and 98.3% for proteins for both minimap2 and GSAlign. This means that annotated sequences clustered together are nearly identical, although, as seen in **Supplemental Figure S4**, that does not guarantee that the same protein sequence is always encoded by clustered genes, as a result of divergent gene model annotation. Note that we also found some cases, 201/29922 for minimap2 and 327/30087 for GSAlign, where cDNA sequences of the same cluster could not be aligned. These occur when overlapping genes do not share exons.

Tables 2 and **Table 3** also contain a summary of BUSCO analysis. BUSCOs are sets of universal single-copy orthologs tailored to different taxa and are typically used to estimate the completeness of genome assemblies. In this context, BUSCOs provide a biologically meaningful metric based on expected gene content (Manni et al., 2021). The percentage of complete BUSCOs found in the core and soft-core pangene sets are similar for both WGA algorithms. The ACK2 dataset contains the two most divergent genomes, with ~85% nucleotide identity (see **Table 3**). Despite this, the core pangenes found across *A. thaliana* and *A. lyrata* contain a similar percentage of complete BUSCOs with both algorithms (~74%). In the rice3 dataset the nucleotide identity rises to 96% and the core pangene set contains ~84% complete BUSCOs. As BUSCO analysis does not make sense for a single chromosome, the wheat dataset was left out. As for the barley20 dataset, the nucleotide identity is generally higher than that of rice3, as expected for a set of cultivars, and the benchmark produced core sets with close to 60% complete BUSCOs. This number increased to 76-82% when the soft-core pangenes are considered, and in the case of minimap2, to 95.7% when pangenes present in 15+ cultivars are analyzed. To put all these BUSCO scores in perspective please see the scores of individual input genome annotations in **Supplemental Table S2**.

Finally, **Table 3** also shows the Average Nucleotide Identity values computed by the GSAlign algorithm for pairwise genome alignments. These values are useful to measure the divergence of the genomes being compared.

Comparison to ancestral karyotype and Ensembl orthogroups

Additional analyses were carried out to gain insights into the performance of our protocol by comparing our results to data produced independently.

First, we estimated its recall on the ACK2 dataset, which represents the most difficult scenario tested due to having the lowest nucleotide identity (~85%). This experiment counted the number of collinear genes identified by minimap2 and GSAlign within 23 blocks of the Ancestral Crucifer Karyotype. The results, summarized in **Supplemental Table S3**, indicate that in these conditions half of the genes making up the blocks are called collinear.

Second, taking advantage of the fact that the genomes in datasets ACK2 and rice3 are included in Ensembl Plants, it was possible to compare the pangenes to precomputed Ensembl Compara orthogroups. In this comparison, clusters are said to match orthogroups when they include all the orthologues annotated in Ensembl; the results are shown in **Table 4**. The analysis of the challenging ACK2 dataset shows that *get_pangenes.pl* fails to recover a quarter of core clusters found among Compara orthogroups, while producing more clusters with multiple copies. This means that ANI does limit the recall of our protocol. Nevertheless, over 86% of pangene clusters match orthogroups and contain sequences that encode common InterPro protein domains. This indicates that pangene clusters are generally biologically relevant.

In contrast, the rice3 data show that *get_pangenes.pl* captures over seven thousand more core pangenes (20K vs 13K). Overall, 85% of pangene clusters match Compara orthogroups and three quarters contain sequences that encode common InterPro protein domains. Inspection of

the data, as the examples on **Supplemental Figure S5**, demonstrate that our protocol is able to cluster together overlapping gene models which might be split or incomplete. This would illustrate why pangene clusters are more likely to group together multiple sequences from the same species. Note that incomplete gene models, or clusters with sequences of contrasting length, would in turn explain why some clusters contain sequences that do not encode a common protein domain (see **Supplemental Figure S6**). The table also shows that up to six thousand shell pangene clusters are produced that pair an annotated gene model with an overlapping genomic segment (gDNA) from a different species. **Supplemental Table S4** shows how core and shell clusters are represented as a BED-like pangene matrix produced by *get_pangenes.pl* .

Confirming gene Presence-Absence Variation

A use case of pangenome analysis of plants and other organisms is to find genes which might be present and functional only in some individuals or populations. These would be annotated by our pipeline as shell genes. As discussed in the previous section, the *get_pangenes.pl* protocol can produce gDNA sequence clusters that contain the genomic intervals of annotated gene models plus overlapping unannotated segments from other species. This feature takes advantage of pre-computed WGAs, which match genomic segments whether they harbor gene models or not. Such clusters can effectively be used to lift-over or project gene models from the species where they are annotated to other individuals. In particular, CDS or cDNA isoform sequences from annotated genes can be mapped to the matching genomic segments and the resulting alignment will directly confirm whether exon/intron boundaries and the embedded coding sequence are conserved. When conserved, the segment likely contains an overlooked gene model; when not, it probably contains a gene fragment or a pseudogene. Next we illustrate how this test works with an example pangene,

Horvu_MOREX_1H01G020700 from dataset barley20. First, the flowchart in **Figure 3A** summarizes how the script *check_evidence.pl* retrieves the WGA evidence supporting the cluster (**Figure 3B**), defines consensus and outlier isoform sequences, and then projects consensus CDS or cDNA sequences on candidate genomic segments with GMAP (See Materials and Methods). If the projection succeeds, a GFF file is created that captures the genomic coordinates of the projected gene (**Figure 3C**). In addition to missing gene models, if there is enough WGA evidence lifting-over can also merge split gene models or, conversely, divide gene models that might have been merged during gene annotation. Finally, the multiple alignment in **Figure 3D** shows that the projected gene model encodes a protein sequence nearly identical to proteins annotated in other cultivars. More generally, we find that the combination of WGA evidence and gene model projection could be a powerful way to refine gene annotation across individuals of the same species. Moreover, as shown in the example, lift-over alignments can be used to confirm or reject observed PAV. The script *check_evidence.pl* facilitates these tasks, but note that additional data beyond sequence evidence would be required to fully characterize such genes, such as expression data and, ultimately, proteomics evidence.

ACKNOWLEDGEMENTS

We are grateful to all members of the PanOryza consortium for feedback during the development of this work. We thank the Gramene team for continuous support and cooperation, as well as members of the Ensembl team for developing and maintaining the front-end and back-end software and infrastructure that underpins Ensembl Plants [Wellcome Trust WT222155/Z/20/Z]. We would like to thank BBSRC/NSF for funding [BB/T015691/1, BB/T015608/1] and the European Molecular Biology Laboratory.

CONFLICT OF INTEREST

Paul Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd.

ORCID

0000-0002-5462-907X Bruno Contreras Moreira

0000-0002-6433-8356 Shradha Saraf

0000-0002-0523-4071 Guy Naamati

0000-0002-8965-1648 Sandeep S Amberkar

0000-0002-3897-7955 Paul Flicek

0000-0001-6118-9327 Andrew R Jones

0000-0001-5690-9633 Sarah Dyer

SUPPLEMENTAL MATERIAL

Supplementary tables and figures are included in file Supplementary_data.pdf.

REFERENCES

Contreras-Moreira, B., Cantalapiedra, C.P., García-Pereira, M.J., Gordon, S.P., Vogel, J.P., Igartua, E., Casas, A.M., & Vinuesa, P. (2017). Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species.. *Frontiers in plant science*, *8*, 184. https://doi.org/10.3389/fpls.2017.00184

Contreras-Moreira, B., Naamati, G., Rosello, M., Allen, J.E., Hunt, S.E., Muffato, M., Gall, A., & Flicek, P. (2022). Scripting Analyses of Genomes in Ensembl Plants. In D. Edwards (Ed.), *Plant Bioinformatics* (pp. 27–55). Springer US, New York, NY.

Golicz, A.A., Batley, J., & Edwards, D. (2015). Towards plant pangenomics. *Plant Biotechnol J*, . https://doi.org/10.1111/pbi.12499

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T.E., Amasino, R., Caicedo, A.L., Goodstein, D., Davidson, P., Mur, L.A.J., Figueroa, M., Freeling, M., Catalan, P., & Vogel, J.P. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure.. *Nature Communications*, *8*, 2184. https://doi.org/10.1038/s41467-017-02292-8

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics*, *38*, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*, bav096. https://doi.org/10.1093/database/bav096

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., Ens, J., Zhang, X.-Q., Angessa, T.T., Zhou, G., Tan, C., Hill, C., Wang, P., Schreiber, M., Boston, L.B., Plott, C., Jenkins, J., Guo, Y., Fiebig, A., Budak, H., Xu, D., Zhang, J., Wang, C., Grimwood, J., Schmutz, J., Guo, G., Zhang, G., Mochida, K., Hirayama, T., Sato, K., Chalmers, K.J., Langridge, P., Waugh, R., Pozniak, C.J., Scholz, U., Mayer, K.F.X., Spannagl, M., Li, C., Mascher, M., & Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, *588*, 284–289. https://doi.org/10.1038/s41586-020-2947-8

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space.. *Database: the Journal of Biological Databases and Curation*, *2011*, bar030. https://doi.org/10.1093/database/bar030

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.. *Bioinformatics*, *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*, 265. https://doi.org/10.1186/s13059-020-02168-z

Lin, H.-N., & Hsu, W.-L. (2020). GSAlign: an efficient sequence alignment tool for intra-species genomes. *BMC Genomics*, *21*, 182. https://doi.org/10.1186/s12864-020-6569-1

Lovell, J.T., Bentley, N.B., Bhattarai, G., Jenkins, J.W., Sreedasyam, A., Alarcon, Y., Bock, C., Boston, L.B., Carlson, J., Cervantes, K., Clermont, K., Duke, S., Krom, N., Kubenka, K., Mamidi, S., Mattison, C.P., Monteros, M.J., Pisani, C., Plott, C., Rajasekar, S., Rhein, H.S., Rohla, C., Song, M., Hilaire, R.St., Shu, S., Wells, L., Webber, J., Heerema, R.J., Klein, P.E., Conner, P., Wang, X., Grauke, L.J., Grimwood, J., Schmutz, J., & Randall, J.J. (2021). Four chromosome scale genomes

and a pan-genome annotation to accelerate pecan tree breeding. *Nature Communications*, *12*, 4125. https://doi.org/10.1038/s41467-021-24328-w
Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D.M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, e78526. https://doi.org/10.7554/eLife.78526
Lysak, M.A., Mandáková, T., & Schranz, M.E. (2016). Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, *30*, 108–115. https://doi.org/10.1016/j.pbi.2016.02.001
Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., & Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199
Mascher, M. (2020). Assembly, annotation and analysis of the barley (Hordeum vulgare L.) pan-genome. https://doi.org/10.5447/IPK/2020/24
Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, 304. https://doi.org/10.12688/f1000research.23297.1
Quinlan, A.R., & Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features.. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033
Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D., & Smit, S. (2016). PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, *32*, i487–i493. https://doi.org/10.1093/bioinformatics/btw455
Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J.N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., Ban, T., Venturini, L., Bevan, M., Clavijo, B., Koo, D.-H., Ens, J., Wiebe, K., N'Diaye, A., Fritz, A.K., Gutwin, C., Fiebig, A., Fosker, C., Fu, B.X., Accinelli, G.G., Gardner, K.A., Fradgley, N., Gutierrez-Gonzalez, J., Halstead-Nussloch, G., Hatakeyama, M., Koh, C.S., Deek, J., Costamagna, A.C., Fobert, P., Heavens, D., Kanamori, H., Kawaura, K., Kobayashi, F., Krasileva, K., Kuo, T., McKenzie, N., Murata, K., Nabeka, Y., Paape, T., Padmarasu, S., Percival-Alwyn, L., Kagale, S., Scholz, U., Sese, J., Juliana, P., Singh, R., Shimizu-Inatsugi, R., Swarbreck, D., Cockram, J., Budak, H., Tameshige, T., Tanaka, T., Tsuji, H., Wright, J., Wu, J., Steuernagel, B., Small, I., Cloutier, S., Keeble-Gagnère, G., Muehlbauer, G., Tibbets, J., Nasuda, S., Melonek, J., Hucl, P.J., Sharpe, A.G., Clark, M., Legg, E., Bharti, A., Langridge, P., Hall, A., Uauy, C., Mascher, M., Krattinger, S.G., Handa, H., Shimizu, K.K., Distelfeld, A., Chalmers, K., Keller, B., Mayer, K.F.X., Poland, J., Stein, N., McCartney, C.A., Spannagl, M., Wicker, T., & Pozniak, C.J. (2020). Multiple wheat genomes reveal global variation in modern breeding.. *Nature*, *588*, 277–283. https://doi.org/10.1038/s41586-020-2961-x
Weisman, C.M., Murray, A.W., & Eddy, S.R. (2022). Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology*, *32*, 2632-2639.e2. https://doi.org/10.1016/j.cub.2022.04.085
Wu, T.D., & Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310
Yates, A.D., Allen, J., Amode, R.M., Azov, A.G., Barba, M., Becerra, A., Bhai, J., Campbell, L.I., Carbajo Martinez, M., Chakiachvili, M., Chougule, K., Christensen, M., Contreras-Moreira, B., Cuzick, A., Da Rin Fioretto, L., Davis, P., De Silva, N.H., Diamantakis, S., Dyer, S., Elser, J., Filippi, C.V., Gall, A., Grigoriadis, D., Guijarro-Clarke, C., Gupta, P., Hammond-Kosack, K.E., Howe, K.L., Jaiswal, P., Kaikala, V., Kumar, V., Kumari, S., Langridge, N., Le, T., Luypaert, M., Maslen, G.L., Maurel, T., Moore, B., Muffato, M., Mushtaq, A., Naamati, G., Naithani, S., Olson, A.,

Parker, A., Paulini, M., Pedro, H., Perry, E., Preece, J., Quinton-Tulloch, M., Rodgers, F., Rosello, M., Ruffier, M., Seager, J., Sitnik, V., Szpak, M., Tate, J., Tello-Ruiz, M.K., Trevanion, S.J., Urban, M., Ware, D., Wei, S., Williams, G., Winterbottom, A., Zarowiecki, M., Finn, R.D., & Flicek, P. (2022). Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Research*, *50*, D996–D1003. https://doi.org/10.1093/nar/gkab1007

FIGURES AND TABLES

Table 1. Datasets used in this study. Geneset sources correspond to Ensembl Plants releases (EP) and the barley gene annotation from IPK (Mascher, M, 2020).

dataset	species / genotype	INSDC accession	geneset	size (MB)	#genes
ACK2	*Arabidopsis thaliana*	GCA_000001735.1	EP52	116	27,655
	Arabidopsis lyrata	GCA_000004255.1	EP52	200	32,667
rice3	*Oryza sativa* Nipponbare	GCA_001433935.1	EP54	363	37,960
	Oryza sativa indica 93-11	GCA_000004655.2	EP54	414	40,745
	Oryza nivara IRGC:100I897	GCA_000576065.1	EP54	327	36,313
chr1wheat10	Chinese Spring	GCA_900519105.1	EP52	1,736	18,017
	Arinalrfor	GCA_903993985.1	EP52	1,725	13,542
	Jagger	GCA_903993795.1	EP52	1,740	17,810
	Julius	GCA_903994195.1	EP52	1,729	17,669
	Lancer	GCA_903993975.1	EP52	1,721	17,951
	Landmark	GCA_903995565.1	EP52	1,738	17,792
	Mace	GCA_903994175.1	EP52	1,713	17,724
	Sy Mattis	GCA_903994185.1	EP52	1,718	17,941
	Norin61	GCA_904066035.1	EP52	1,726	18,091
	Stanley	GCA_903994155.1	EP52	1,737	17,767

barley20	MorexV2	LR722616-LR722623	IPK2020	4,210	46,294
	Akashinriki	ERS4201448	IPK2020	4,401	44,446
	B1K-04-12	ERS4201449	IPK2020	4,142	44,566
	Barke	ERS4201450	IPK2020	4,073	45,999
	Golden Promise	GCA_902500625.1	IPK2020	3,946	42,464
	HOR10350	ERS4201451	IPK2020	4,086	45,810
	HOR13821	ERS4201452	IPK2020	4,324	44,714
	HOR13942	ERS4201453	IPK2020	4,249	44,718
	HOR21599	ERS4201454	IPK2020	4,344	44,456
	HOR3081	ERS4201455	IPK2020	4,201	45,146
	HOR3365	ERS4201456	IPK2020	4,722	47,588
	HOR7552	ERS4201457	IPK2020	4,228	44,641
	HOR8148	ERS4201458	IPK2020	4,212	45,026
	HOR9043	ERS4201459	IPK2020	4,270	45,028
	Hockett	ERS4201460	IPK2020	4,201	46,450
	Igri	ERS4201461	IPK2020	4,202	45,213
	OUN333	ERS4201462	IPK2020	4,392	44,699
	RGT Planet	ERS4201463	IPK2020	4,213	45,413

	ZDM01467	ERS4201447	IPK2020	4,540	44,746
	ZDM02064	ERS4201446	IPK2020	4,153	45,050

Table 2. Summary of pangene analyses based on minimap2 Whole Genome Alignments (WGA). WGA N50 is shown as a range of observed values. The percentage of genes in blocks of 3+ contiguous genes is also shown as a range. Note that barley and wheat datasets require optional argument -H, which masks geneless regions longer than 1Mbp, where repeated sequences accumulate. In all cases only chromosomes of the same name were aligned (optional argument -s). Maximum RAM use was measured for pairwise WGA batch jobs. In the barley20 row the numbers in curly brackets correspond to clusters with occupancy >= 15.

	max RAM (GB)	WGA N50 (Kbp)	% genes blocks3+	total clusters	(soft) core clusters	% BUSCO complete
ACK2	4.5	6.1	34.3	43,951	15,768	74.3
rice3	1.4	[27.4, 29]	[74.4, 78.4]	62,915	18,681	83.9
chr1wheat10 (-H)	64.5	[80.8, 142.4]	[38.8, 54.2]	30,173	7,872	
barley20 (-H)	46.3	[43.6, 75.7]	[24.9, 35.4]	180,667	15,674 (23,888) {34,074}	61.3 (82.1) {95.7}

Table 3. Summary of pangene analyses based on GSAlign Whole Genome Alignments (WGA). WGA N50 is shown as a range of observed values. The percentage of genes in blocks of 3+ contiguous genes and the Average Nucleotide Identities (ANI) are also shown as ranges. Note that optional argument -H was also tested on barley. In all cases only chromosomes of the same name were aligned (optional argument -s). Maximum RAM use was measured for pairwise WGA batch jobs. The bottom row is included as a control to measure the effect of using optional argument -H.

	max RAM (GB)	WGA N50 (Kbp)	% genes blocks3 +	total clusters	(soft) core clusters	% BUSCO complete	% ANI
ACK2	4.5	4.3	23.4	43,340	16,432	74.7	84.7
rice3	3.3	[15.2, 16.9]	[51.6, 56.6]	62,844	18,626	84.2	[96.4, 97.6]
chr1wheat10	83.4	[40.9, 72.1]	[20.2, 34.2]	30,135	7,723		[98.9, 99.4]
barley20	113.1	[17.1, 34.3]	[10.5, 15.9]	173,984	13,934 (21,171)	56.8 (76.2)	[96.9, 99.3]
barley20 -H	110.1	[16.6, 32.9]	[10.5, 15.9]	188,289	13,957		[96.8, 99.3]

Table 4. Summary of pangene clusters obtained for datasets ACK2 and rice3 and the corresponding orthogroups in Ensembl Plants. Core clusters contain genes from all analyzed genomes; in rice, shell clusters contain genes from two species. Clusters with multiple copies have several genes from the same species. gDNA segments are shell clusters that bring together a gene model and a matching genomic segment from the underlying WGA. Column 'match Compara' shows the number of pangene clusters that contain the same genes as the corresponding Compara orthogroups. The last column shows the number of pangene clusters that contain sequences that share an InterPro domain (the number in square brackets is for core clusters only).

	dataset	core clusters	multiple copies	shell clusters	gDNA segments	match Compara	share InterPro domains
Compara orthogroups	ACK2	20,192	161				[18,259]
minimap2 clusters	ACK2	15,768	490			14,044	[14,371]
GSAlign clusters	ACK2	16,432	446			14,145	[14,790]
Compara orthogroups	rice3	13,020	219	6,386			16,766 [11,571]
minimap2 clusters	rice3	20,419	3,022	9,503	5,593	17,317	22,796 [17,232]
GSAlign clusters	rice3	20,224	2,831	9.863	6,173	17,103	22,818 [16,957]

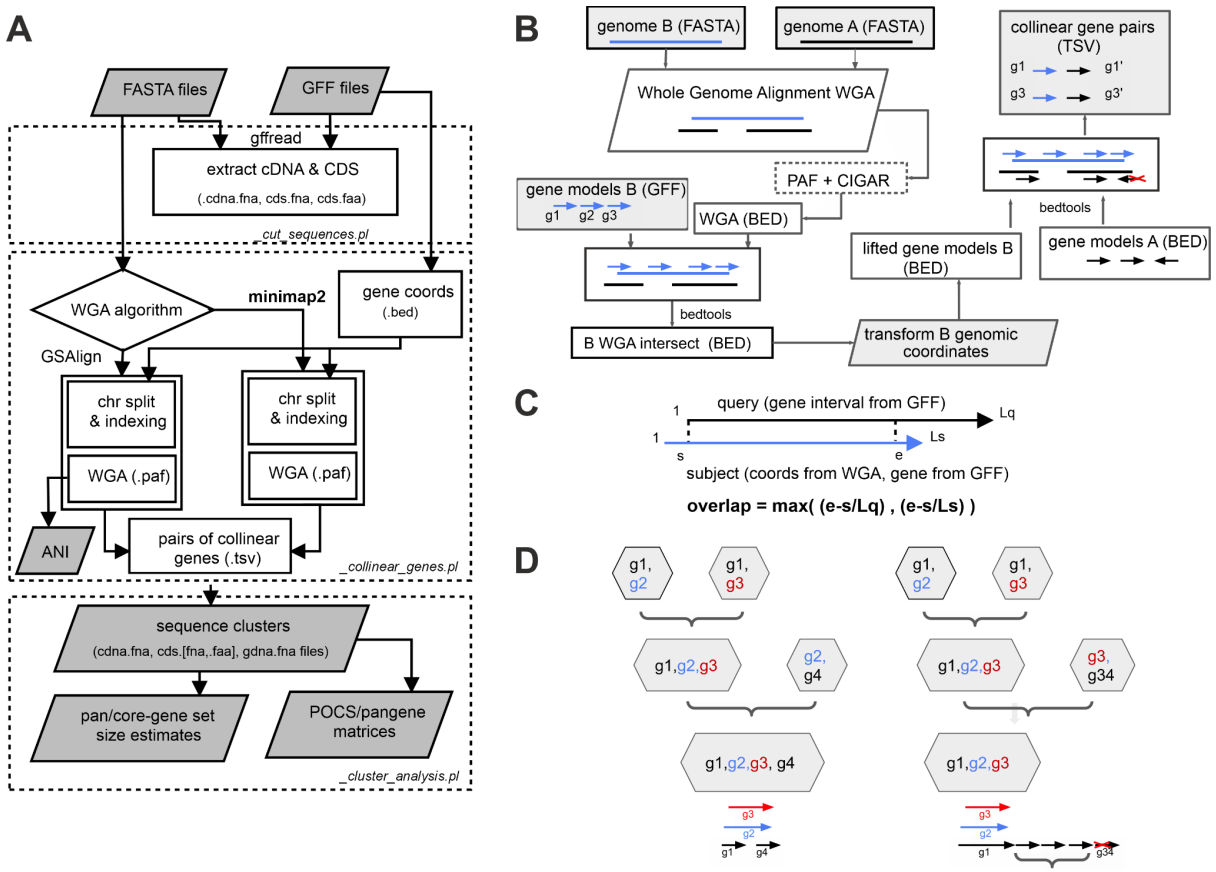


Figure 1. Features of get_pangenes.pl.

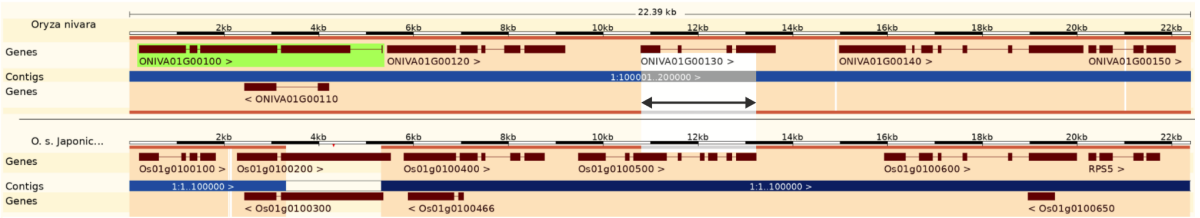
A) Flowchart of the main tasks and deliverables of script *get_pangenes.pl*: cutting cDNA and CDS sequences (top), calling collinear genes (middle, panels B and C) and clustering (bottom, panel D). By default, only cDNA and CDS sequences longer than 100bp are considered. Whole Genome Alignments (WGA) can be computed with minimap2 (default) or GSAlign, and the input genomes can optionally be split in chromosomes or have their long geneless regions (>1Mbp) masked. Resulting gene clusters contain all isoforms and are post-processed to produce pangene and Percentage of Conserved Sequences (POCS) matrices, as well as to estimate pan-, soft-core-, and core-genomes. GSAlign also produces Average Nucleotide Identity (ANI) matrices. Several tasks can be fine-tuned by customizing an array of parameters, of which alignment coverage is perhaps the most important.

B) WGA of genomes A and B produces BED-like files that are intersected with gene models from B. Intersected coordinates are then used to transform B gene models to the genomic

space of A. Finally, overlapping A gene models on the same strand are defined as collinear genes.

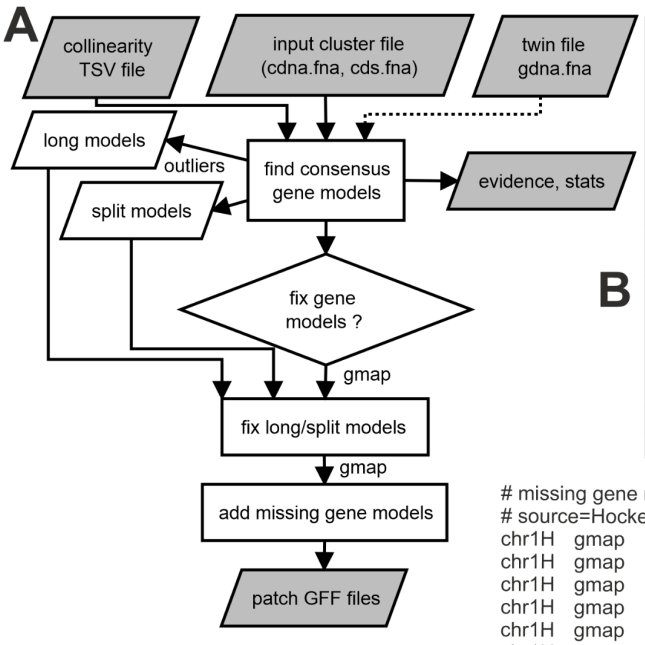
C) Feature overlap is computed from WGAs and gene coordinates from source GFF files. When checking the overlap of A and B gene models, strandedness is required. Overlaps can also be estimated between gene models annotated in one assembly and matched genomic segments from others.

D) Making greedy clusters by merging pairs of collinear genes. This algorithm has a key parameter, the maximum distance (in genes) among sequences of the same species that go in a cluster (default=5). Its effect is illustrated on the right side, where gene g34 is left unclustered for having too many intervening genes.



Oryza *nivara*	ONIVA01G00100	ONIVA01G00110	ONIVA01G00120	[gDNA segment]	ONIVA01G00130	ONIVA01G00140	[gDNA segment]	ONIVA01G00150
minimap2 overlap (bp)	7839, 3065	1787	2963	1169	2428	4063	564	1471
GSAlign overlap (bp)	7827, 6546		2956	1173	2415	4052	562	1501
Oryza sativa Japonica	Os01g0100100 Os01g0100200	Os01g0100300	Os01g0100400	Os01g0100466	Os01g0100500	Os01g0100600	Os01g0100650	RPS5

Figure 2. Aligned genomic region in chr1 of *Oryza nivara* (top) and *Oryza sativa* Japonica group cv. Nipponbare (bottom) as displayed in the Ensembl Plants browser. Genes on the forward strand (>) are above contigs, whilst those in the negative strand (<) are underneath. As a result of the genomic alignment, genes of *O. nivara* overlap with gene models from *O. sativa*. This evidence can be used to identify collinear genes that take equivalent positions in different genomes, as illustrated with gene models ONIVA01G00130 and Os01g0100500, which overlap over 2.4kb (white rectangle). The example shows that overlapping gene models might share only some exons. The table below shows the collinear gene models identified based on minimap2 ang GSAlign alignments, together with the corresponding overlapped based pairs.



B

length	pairs	overlap	genename	taxon
892	14	12005	Horvu_AKASHIN_1H01G016200	Akashinriki
898	13	11167	Horvu_21599_1H01G015900	HOR21599
904	13	11258	Horvu_10350_1H01G020000	HOR10350
892	11	9519	Horvu_PLANET_1H01G014600	RGT_Planet
733	11	8040	Horvu_BARKE_1H01G018700	Barke
904	11	9434	Horvu_9043_1H01G017700	HOR9043
1380	10	8989	Horvu_HHOR_1H01G019100	HOR3365
898	10	8152	Horvu_8148_1H01G016700	HOR8148
733	9	6558	Horvu_HOCKETT_1H01G013500	Hockett
771	8	6166	Horvu_MOREX_1H01G020700	Morex
904	7	5961	Horvu_IGRI_1H01G016300	Igri
897	7	5962	Horvu_7552_1H01G018700	HOR7552
1444	6	5769	Horvu_HUANG_1H01G010400	ZDM01467
1374	6	5344	Horvu_7552_1H01G018300	HOR7552
1380	2	2247	Horvu_FT11_1H01G019600	B1K-04-12
1380	2	1661	Horvu_FT11_1H01G019300	B1K-04-12

C

```
# missing gene model: corrected chr1H:3798001-3798892(-) [HOR3081]
# source=Hockett matches=639 mismatches=9 indels=0
chr1H   gmap   gene   3798173   3798892   .     -   .
chr1H   gmap   mRNA   3798173   3798892   .     -   .
chr1H   gmap   exon   3798591   3798892   97    -   .
chr1H   gmap   exon   3798173   3798518   100   -   .
chr1H   gmap   CDS    3798591   3798892   97    -   0
chr1H   gmap   CDS    3798173   3798518   100   -   2
```

D

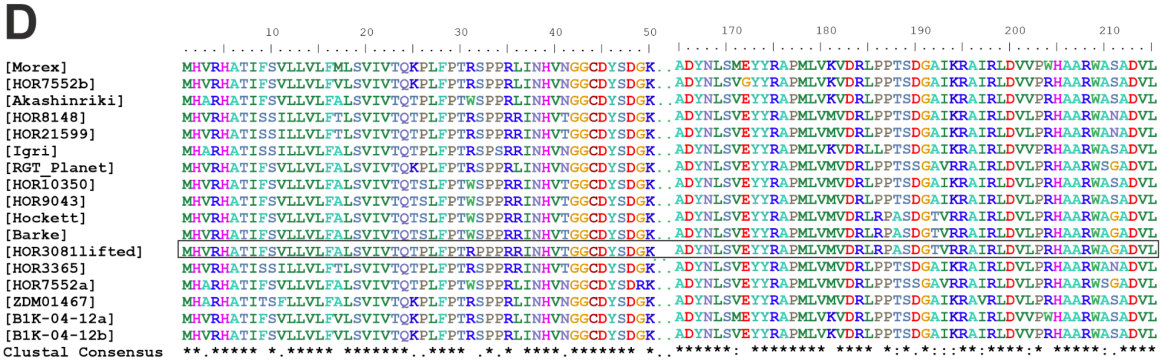


Figure 3. Evidence-based analysis of shell pangene cluster.

A) Flowchart of script check_evidence.pl , which uses as input a cluster in FASTA format and precomputed collinearity evidence in TSV format. If available, a twin FASTA file with the corresponding collinear genomic segments is also taken, which would be needed for confirming presence/absence of gene models.

B) Summary of Whole Genome Alignment evidence for the gene models in CDS cluster Horvu_MOREX_1H01G020700 resulting from the analysis of dataset barley20. This cluster contains isoforms from 16 gene models and 14 barley cultivars. Note that there are no genes from cultivar HOR3081.

C) Patch GFF file produced by lifting-over a CDS isoform from cultivar Hockett on the overlapping genomic segment from cultivar HOR3081. The actual CDS vs genomic DNA

alignments are computed with GMAP. In this case the CDS sequence was aligned with no indels, 639 matches and 9 mismatches.

D) Alignment of amino acid sequences of CDS isoforms from cluster Horvu_MOREX_1H01G020700. Note that the lifted-over gene model encoded in the patch GFF file is also included, producing a protein sequence of the same length of cultivars Barke and Hockett. Only the first and last 50 residues of that protein are shown for brevity.