Gene models for wheat

A good genome assembly (as discussed in the Genome assemblies section) is an essential prerequisite for obtaining high quality gene models; the models can only be as good as the assembly they are based on. As wheat is a polyploid species its genes are usually present in multiple copies, one from each homoeologous chromosome. Homoeologous genes are at least 95% similar across their coding region, which complicates the process of creating accurate gene models. Combining transcriptome data and gene models from related species has proven to be a useful "formula" to define correct gene models in wheat.

Although many genome assemblies were released over the last few years, only a few of these also contained annotated gene models. In this section we will compare the various sets of gene models in wheat and also explain the nomenclature behind them.

Please note that the wheat genome sequencing efforts are rapidly updating the "state of the art" resources so gene models and genome sequences may change.

a) Gene models in wheat (ordered by release date)

CSS gene models (PGSB v2.2, IWGSC2.26)

This set of gene models was released together with the Chinese Spring Survey Sequence (CSS) of wheat in 2014. It was the first time that a set of gene models could be assigned to each of the wheat chromosomes individually. But many of the predicted 104,934 high confidence gene models were wrong and many genes were missing from this set. This was due to the fragmented nature of the assembly, consisting of more than 10 million small scaffolds. The gene models were often truncated or split erroneously into two separate gene models.

The CSS gene models are available from the Ensembl Plants archive pages (http://archive.plants.ensembl.org/Triticum_aestivum/Info/Index). They are still being used in the WheatEXP expression browser (https://wheat.pw.usda.gov/WheatExp/) as well as the wheat *in silico* TILLING database (http://www.wheat-tilling.com/). The expVIP expression browser (http://www.wheat-expression.com/) also contains these gene models for legacy reasons.

CSS reference: IWGSC 2014, DOI: 10.1126/science.1251788

TGACv1 gene models

The TGACv1 gene models represent an improvement on the CSS gene models. This was mainly due to the improved scaffold sizes, which allowed accurate prediction of 104,390 whole high confidence (HC) gene models. This set of gene models was the reference standard until the release of the RefSeq gene models (see below). The gene models are available from the Ensembl Plants archive pages (http://oct2017-lants.ensembl.org/index.html); the expVIP expression browser (http://www.wheat-expression.com/) allows for the use of the TGACv1 gene models.

TGACv1 reference: Clavijo *et al.*, 2017 DOI: 10.1101/gr.217117.116

WEWSeq v.1.0 and v.2.0 gene models

The WEWSeq v.1.0 gene models are based on the wild emmer genome assembly (WEWSeq v.1.0). They represent an accurate prediction of the genes present in wild emmer, as the underlying reference is of high quality. This set of gene models (65,012 high confidence gene models) allows for comparative evolutionary studies between wild emmer and domesticated wheat. The gene models represent the current gene annotation of wild emmer on the Ensembl Plants website.
As part of the annotation of the 'Svevo' genome (see below), the gene models of the WEWSeq genome were annotated using the same pipeline as the Svevo gene models. The WEWSeq version 2.0 annotation contains 67,182 high confidence gene models but is not incorporated into the Ensembl Plants website. They can be downloaded at https://doi.org/10.5447/ipk/2019/0. Please note that the version 2 gene models use different gene identifiers than version 1 (see **Figure 1**).

WEWSeq v.1.0 reference: Avni *et al.*, 2017 DOI: 10.1126/science.aan0032
WEWSeq v.2.0 reference: Maccaferri *et al.,* 2019 DOI: 10.1038/s41588-019-0381-3
S. Twardziok (2018) DOI:10.5447/IPK/2019/0

RefSeqv1.0 gene models

The RefSeqv1.0 gene models build upon the TGACv1 gene models. Using an even more contiguous assembly and two different prediction pipelines allowed for the accurate prediction of homoeologous sets of gene models. In total 110,790 high confidence genes (homology to genes in other species) and 158,793 low confidence genes (e.g. truncated genes, genes lacking transcriptional evidence or lacking homology to other species) were annotated.

RefSeqv1.0 reference: IWGSC 2018, DOI: 10.1126/science.aar7191

RefSeqv1.1 gene models

This set of gene models is almost identical to the RefSeqv1.0 gene models. The difference is within ~2,000 gene models, which were re-annotated manually. The RefSeqv1.1 gene models are the most complete and accurate set of gene models currently available in hexaploid wheat and represents the gene annotation used on the Ensembl Plants website. The expVIP expression browser (http://www.wheat-expression.com/) includes both the RefSeqv1.0 and v1.1 gene models for expression analyses.

Svevo v.1.0 gene models

The Svevo v.1.0 gene models are based on the genome assembly of tetraploid durum wheat (cultivar 'Svevo'). Protein reference sequences and expression data from 21 RNA-Seq studies were used as evidence to call open reading frames, which were further filtered and refined. This resulted in a set of 66,559 high confidence gene models, which represent the current gene annotation of *Triticum turgidum* on the Ensembl Plants website.

Svevo v.1.0 reference: Maccaferri *et al.,* 2019 DOI: 10.1038/s41588-019-0381-3

b) Overview of gene model nomenclature

The gene models belonging to the various released assemblies all use a different nomenclature. To help users distinguish between the different sets of gene models, we have listed and decoded the nomenclature (see **Figure 1**).

In **Figure 1**, we exemplify the differences in nomenclature using one gene. Up to six elements, called fields, make up the various gene model names. These fields are shown at the top of **Figure 1** with matching colours for the corresponding features in the gene names.

Lastly, the nomenclature used in the RefSeqv1.0 and v1.1 gene models is highlighted with differently shaded blue backgrounds. The RefSeqv1.0 (release annotation) and RefSeqv1.1 (improved annotation) gene models are solely differentiated by the number in front of the biotype; "01" for RefSeqv1.0 and "02" for RefSeqv1.1.

The CSS gene nomenclature simply consists of the species name, the chromosomal location of the gene, and nine alphanumeric characters, i.e. a sequence of nine numbers and/or letters (yellow background in **Figure 1**).

The TGACv1 gene nomenclature consists of the species name, the sequenced accession, the chromosomal location of the gene, the annotation version, and a unique identifier (green background in **Figure 1**). This identifier is composed of two sections; the first number refers to the scaffold the gene is located on (here: TGACv1_scaffold_404669_5BL), while the second number denotes the gene order along a single scaffold in steps of 10. TRIAE_CS42_5BL_TGACv1_404669_AA1307950 is located on the same scaffold as our example gene and precedes it. Please note that gene order in the TGACv1 assembly is only coherent within a single scaffold, but not between scaffolds!

The RefSeqv1.0 and v1.1 nomenclature consists of the species name, the sequenced accession, the chromosomal location of the gene, the annotation version, the biotype, and a unique identifier (blue shaded backgrounds in **Figure 1**). The unique identifiers within the RefSeqv1.0 and v1.1 annotation are progressive numbers in steps of 100s, which reflect the relative position of gene models on the RefSeqv1.0 assembly. This means that our example gene *TraesCS5B02G236400* lies between genes *TraesCS5B02G236300* and *TraesCS5B02G236500*.

For the 'Svevo' annotation, the name of the species is TRITD (*TRITicum Durum*), while for the 'Zavitan' annotation it is TRIDC (*TRIticum DiCcocoides*); for both sets of gene models the gene identifiers increase in steps of 10s.

The RefSeqv1.0 and v1.1 annotations both contain High Confidence (HC) and Low Confidence (LC) gene models. Low confidence gene models follow the same nomenclature as described above but are flagged by an "LC" at the end (not shown here). It is vital to understand that HC and LC genes with sharing the same unique identifier are in fact **not** the same locus and they are not located in sequential order. As such, while both *TraesCS5B02G236400* and *TraesCS5B02G236400**LC*** are

located on chromosome 5B they are not physically adjacent. Instead, *TraesCS5B02G236400**LC*** will be flanked by *TraesCS5B02G236300**LC*** and *TraesCS5B02G236500**LC***.

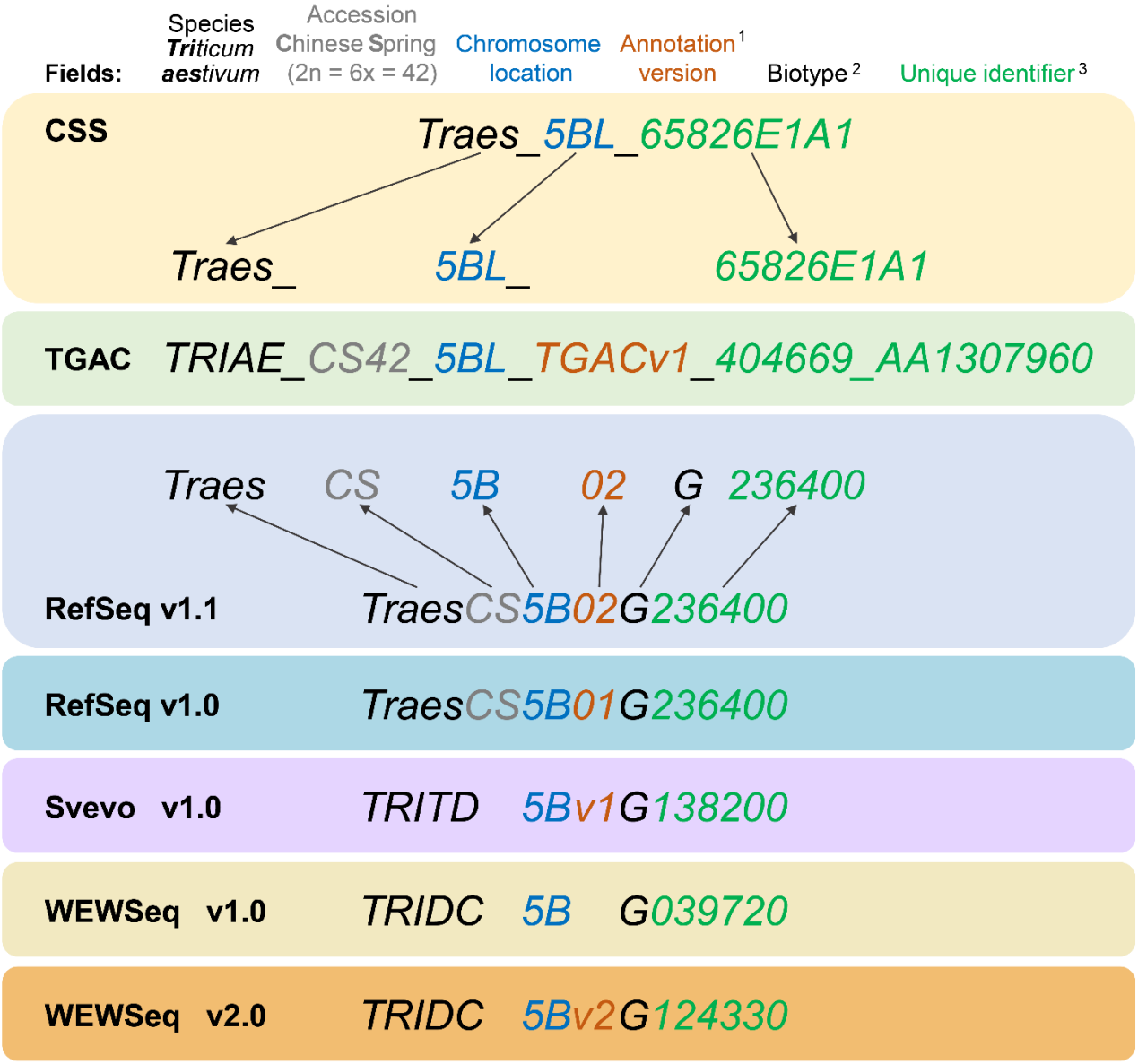

Figure 1. Gene nomenclature of the seven gene annotations available for wheat.