Variation data

This document explains the variation data present on the Ensembl Plants website.

Here the term variation data is used synonymously for Single Nucleotide Polymorphism (SNP) data and small Insertion/Deletions (InDels). SNP data in particular has seen a steep rise in abundance from ~1,536 publicly available SNPs in 2010 to over 14 million publicly available SNPs in 2019. Other variants like large InDels or even large structural variations like inversions or translocations are not considered here; currently there is not enough reliable data for this type of variation in wheat.

Variation data is essential for a number of analyses including genetic markers and haplotype analyses. The following document will explain what data is available on Ensembl Plants, how to access the data, and how to use it.

All SNPs in the Ensembl Plants database have been assessed using a SNP-effect prediction tool. This program predicts the effect of a SNP on its corresponding protein, i.e. whether a SNP is synonymous or whether it leads to a missense mutation or a premature STOP codon. The predicted missense mutations have been analysed using another program called “Sorting Intolerant From Tolerant” (SIFT). This tool predicts the likelihood of an amino acid substitution to affect protein function, i.e. whether the mutation is tolerated or whether it has a deleterious effect on protein function. While the SIFT score can be useful for distinguishing the severity of several missense mutations, it is only a prediction and might not be accurate.

a) Natural Variation Data

Natural variation data describes variation between different species, accessions or cultivars of wheat. It is mostly this type of variation that breeders have used to create improved varieties. Here, we specifically mean SNP variation and small InDels, i.e. single nucleotide (or small oligonucleotide) changes in or near genes that lead to differences in phenotypic traits. Below we will explore how the data was generated.

What data is on Ensembl Plants?

The natural variation data on Ensembl Plants has currently one major source:

1. HapMap variants (currently unavailable on Ensembl Plants, but can be downloaded from http://129.130.90.211/hapmap/#dataset)
2. CerealsDB variants, which are also available at the CerealsDB website

HapMap variants

The Haplotype Map (HapMap) data is a set of 1.57 million SNPs and 161,719 small InDels that were generated by re-sequencing a diverse wheat panel (see **Figure 1**) using whole exome capture (WEC) as well as using a genotyping-by-sequencing (GBS) approach. The wheat panel consists of a collection of 62 landraces, cultivars and breeding lines from the Americas, Europe, Africa, Australia and Asia. Only about 10 % of the SNPs occur within the coding sequence (CDS), of which 76,361 are synonymous SNPs (SNPs that do not cause a change in amino acid of the translated protein) and 83,622 are non-synonymous SNPs (SNPs that do cause a change of amino acid in the translated protein). Only 1,600 of the non-synonymous SNPs cause premature STOP codons while another 1,583 SNPs cause mutations at splice-site junctions. For more information on the HapMap data go to the official website (http://129.130.90.211/hapmap/#intro) and/or read the associated publication (Jordan et al. 2015; http://www.ncbi.nlm.nih.gov/pubmed/25886949).

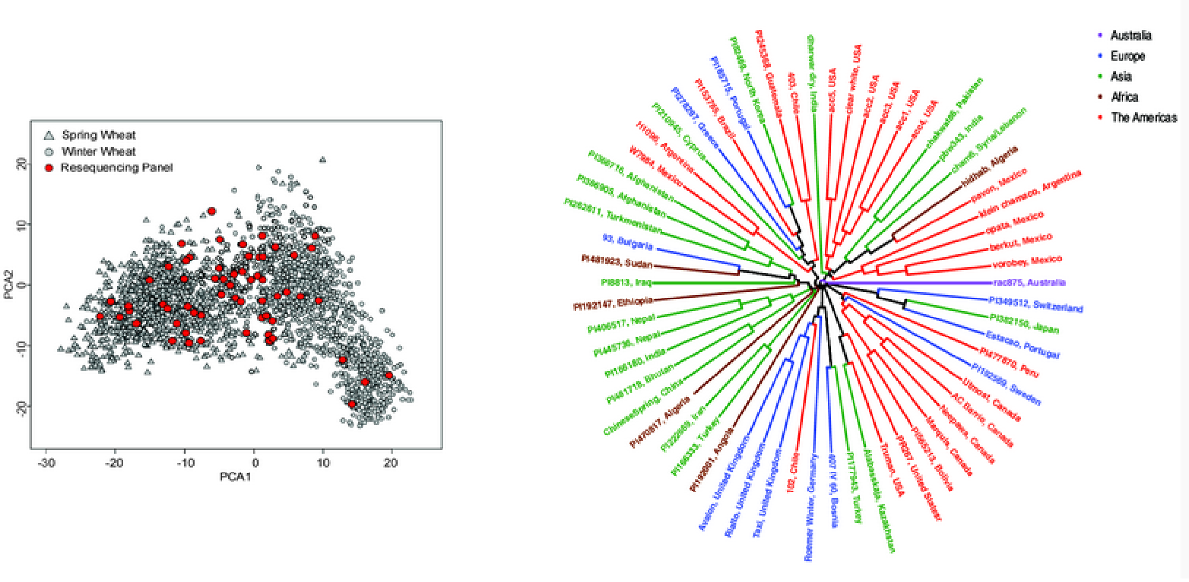


Figure 1: HapMap wheat diversity
This screenshot from the HapMap website shows a principal component analysis (left) and neighbour-joining tree (right), both illustrating the diversity of the accessions used.

CerealsDB variants

The CerealsDB variants in the Ensembl Plants database represent a subset of the 820k Axiom high-density SNP array (http://plants.ensembl.org/Triticum_aestivum/Info/Annotation#variation). Out of the total 820,000 SNPs it was possible to accurately map 768,864 SNPs using the stringent criteria required by the Ensembl Plants database. The SNPs themselves were discovered in a diverse panel of more than 50 hexa-, tetra- and diploid winter and spring accessions (including wild progenitors). For full information about the creation of the data and a complete list of all SNPs go to the CerealsDB website (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php).
The CerealsDB data is split into two subsets; one is the 820K Axiom array and the second the 35k Axiom array. The latter is a subset of the most diverse SNPs from the 820k Axiom SNP array.

b) Induced Variation data

Induced variation data describes variation that was artificially induced using chemicals (e.g. EMS) or irradiation (e.g. X-rays). The resulting plant populations are not classified as transgenic and thus can and have been used by breeders to create improved varieties. The effects on the genome depend on the mutagen used. Ethyl methanesulfoante (EMS) preferentially leads to G-to-A transitions whereas irradiation mostly leads to deletions, the size of which depends on the type of electromagnetic radiation used.

What data is on Ensembl Plants?

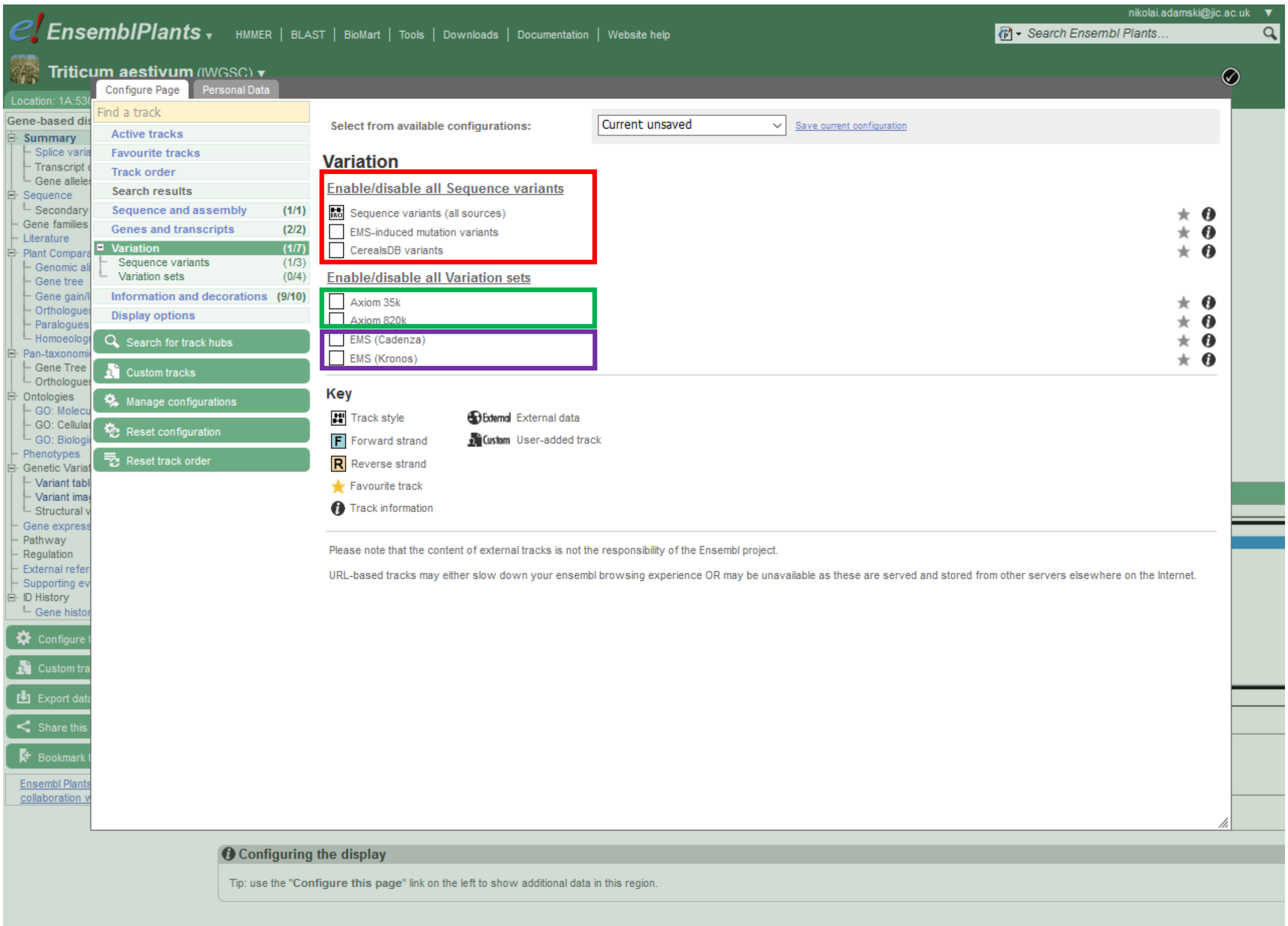
Two TILLING populations, one in hexaploid background Cadenza (UK variety) and one in tetraploid background Kronos (US variety), have been exome-sequenced (at M2 generation). The reads from both mutant populations were mapped against the RefSeqv1.0 assembly and the induced EMS SNPs have been predicted.
There is also a legacy version of the TILLING SNPs at www.wheat-tilling.com. The SNP data there is based on the CSS genome assembly (see section "Genome Assemblies for more details"). See the TILLING mutant resource section for more details.

c) How to access variation data?

Visualizing the variation data

We have talked about the type of variation data available on Ensembl Plants and now we will take a look at how to visualize and access it.
Variation data tracks are available both in the Location and Gene tabs of the Ensembl Plants summary page for any gene. There are currently seven tracks that users can choose from allowing the available data described above to be displayed in different subsets. For instance, users can have all CerealsDB SNPs displayed or only the SNP subsets belonging to either the Axiom 820K or the Axiom 35k SNP array (see **Figure 2**). This allows for a great deal of flexibility whether users are interested in a specific SNP set or all available variation data.

Hovering your mouse cursor above a SNP name opens up a toolbox with additional information, e.g. the nucleotide change, a predicted effect based on the underlying gene model as well as a link to a detailed Variant tab (red box in **Figure 3**).



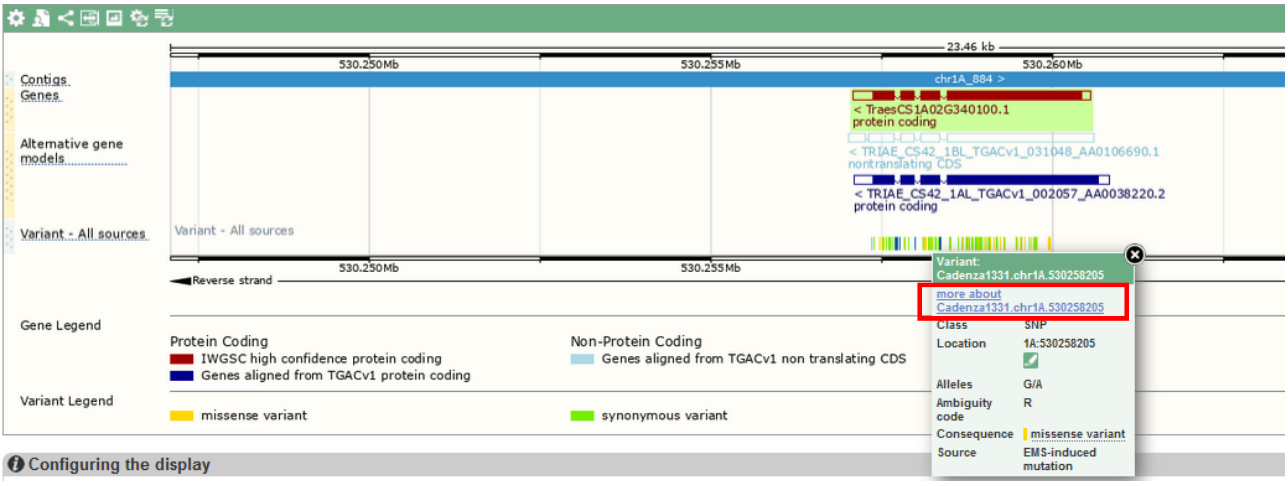
Please note that the content of external tracks is not the responsibility of the Ensembl project.

URL-based tracks may either slow down your ensembl browsing experience OR may be unavailable as these are served and stored from other servers elsewhere on the Internet.

Tip: use the "Configure this page" link on the left to show additional data in this region.

Figure 2: Tracks for Variation data

This screenshot shows the different tracks for variation data on the Ensembl Plants website. Different subsets of the variation data can be displayed; either according to the source of the SNPs (red box), CerealsDB SNPs (green box) or TILLING SNPs (purple box).



Variant: Cadenza1331.chr1A.530258205	
more about Cadenza1331.chr1A.530258205	
Class	SNP
Location	1A:530258205
Alleles	G/A
Ambiguity code	R
Consequence	missense variant
Source	EMS-induced mutation

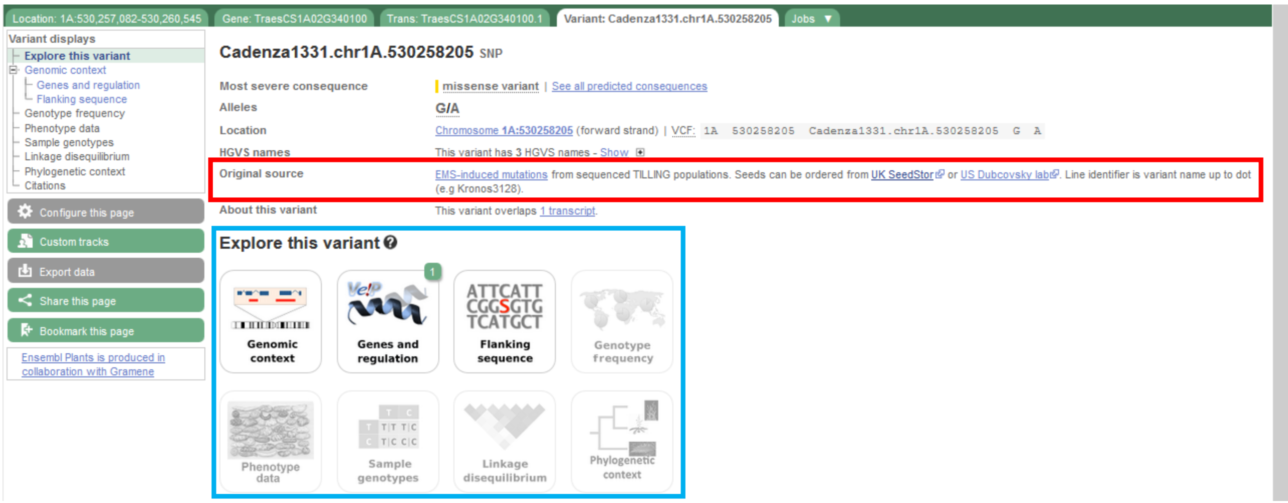
Figure 3: Variation tracks in detail

This screenshot shows an active variation track. The data is ordered by position and named using a standardized nomenclature system. Hovering your mouse cursor over an entry opens a toolbox with further information, including a link to a detailed variant tab (red box).

The Variant tab contains a lot of additional information, like the variants original source (see red box in **Figure 4**). In addition, this tab contains links to other information related to the variant:

- Genomic context:
  This opens up a 5kb window displaying the genomic location with all variants therein; useful to jump to nearby variants.
- Genes and Regulation:
  This opens a window displaying the predicted effect of a variant on its associated gene in addition to other useful information.
- Flanking Sequence:
- This opens a window displaying the nucleotide sequence, 400 bp either side of the variant. This also includes other variants, which allows for a quick assessment of variants. Furthermore, the variants are colour-coded based on their predicted effect on the underlying gene model.
- Genotype Frequency (For CerealsDB SNPs only):
  This opens a window with information on the number of accessions the variant was detected in; from this a frequency of the genotype is calculated. This is in essence the summary of the Sample Genotypes view (see next point).
- Sample Genotypes:
  This opens a window listing all accessions that the variant was detected in, as well as the allele calls for each accession. From this data the genotype frequency was calculated (see above).

There are a few more categories (Linkage disequilibrium, phenotype data, etc) that are currently not available (greyed out). In future updates, when more data becomes available, this will hopefully change.



Cadenza1331.chr1A.530258205 SNP

Most severe consequence	missense variant | See all predicted consequences
Alleles	G/A
Location	Chromosome 1A:530258205 (forward strand) | VCF: 1A 530258205 Cadenza1331.chr1A.530258205 G A
HGVS names	This variant has 3 HGVS names - Show
Original source	EMS-induced mutations from sequenced TILLING populations. Seeds can be ordered from UK SeedStor or US Dubcovsky lab. Line identifier is variant name up to dot (e.g Kronos3128).
About this variant	This variant overlaps 1 transcript.

Figure 4: Variant tab

This is a screenshot of the Variant tab. It displays additional information about a variant, most importantly the variant's original source (red box). This tab also contains links to more detailed effects of the variant (blue box).

In addition, all variants associated with one gene can be visualized either as a "Variant image" or a "Variant table" (**Figure 5** and **Figure 6**, respectively). Both of these visualizations can be accessed from the Gene or the Transcript tab.

The Variant Image displays the selected gene with tracks for predicted protein domains, tracks for the predicted SNP effects and all variants shown as boxes at the bottom of the window. The latter makes selection of individual specific SNPs easier.

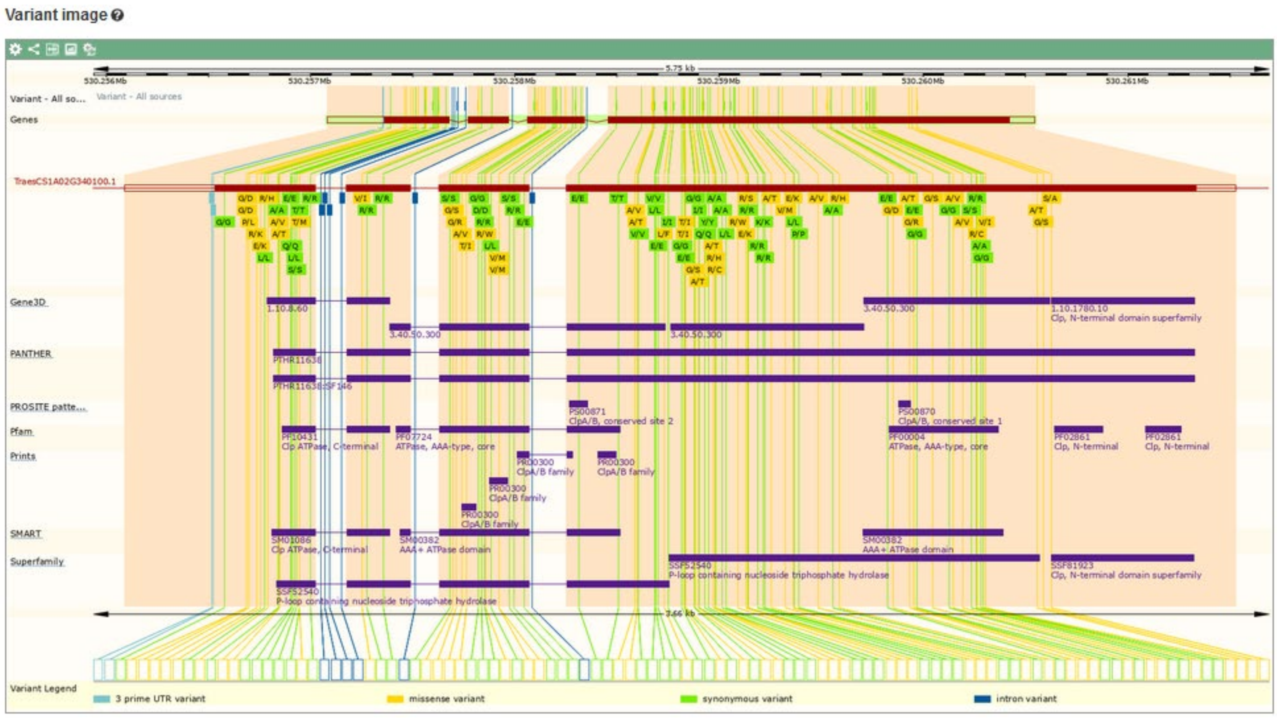
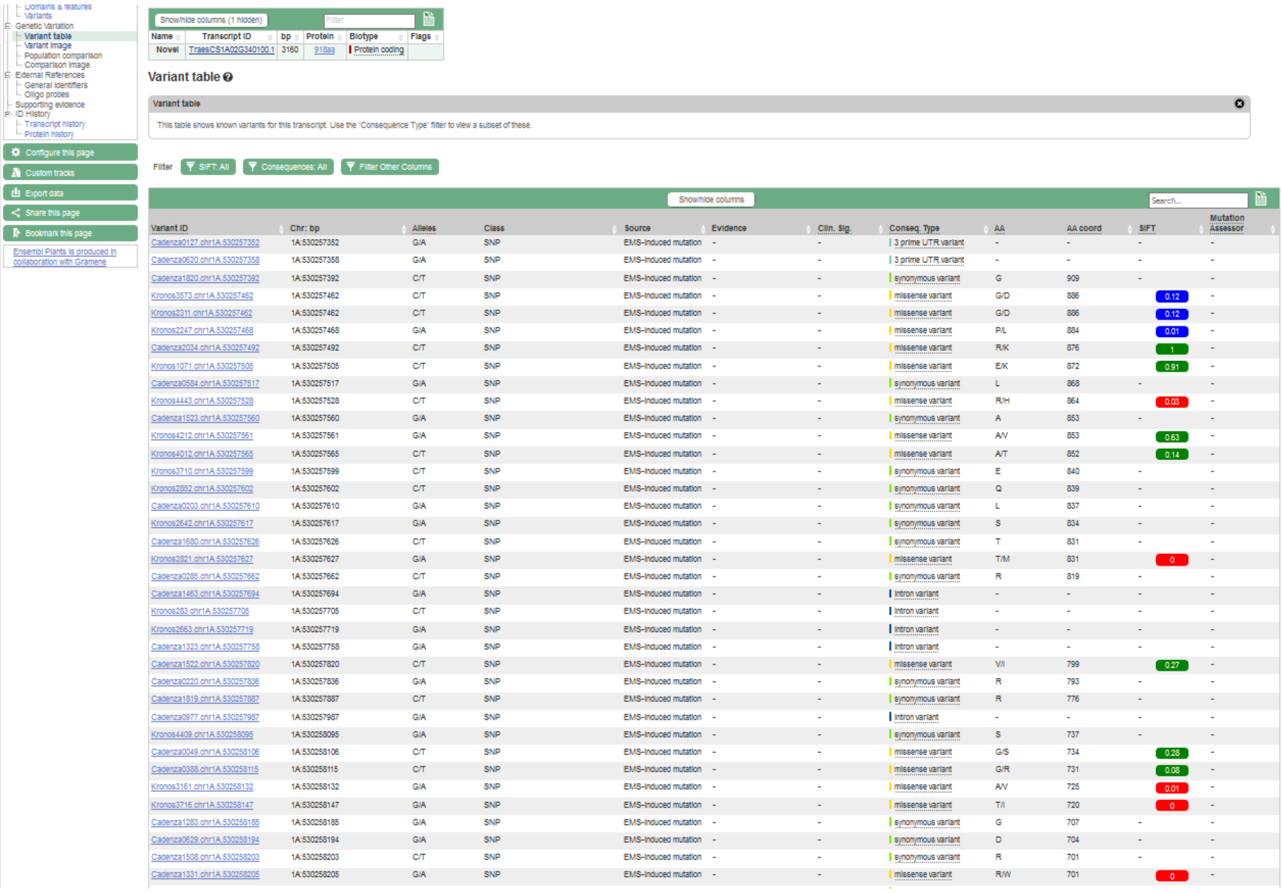


Figure 5: Variant Image

The Variant Image provides a great overview of the SNPs across a selected gene. It allows to select SNPs underlying specific protein domains to target mutations to specific parts of a gene. A colour-coded track also displays the predicted SNP effects on the associated protein sequence.

The Variant Table, as the name suggests, displays all variants associated with a gene in a tabular format. This makes it easy to scan a gene for variation of a specific type, e.g. EMS-induced premature STOP codons (**Figure 6**).

These two new visualization tools enable users to quickly browse the variation associated with any given gene.



Name	Transcript ID	bp	Protein	Biotype	Flags
Novel	TraesCS1A02G340100.1	3160	918aa	Protein coding	

Variant table

Variant table

This table shows known variants for this transcript. Use the 'Consequence Type' filter to view a subset of these.

Variant ID	Chr: bp	Alleles	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA coord	SIFT	Mutation Assessor
Cadenza0127.chr1A.530257352	1A:530257352	G/A	SNP	EMS-induced mutation	-	-	3 prime UTR variant	-	-	-	-
Cadenza0620.chr1A.530257358	1A:530257358	G/A	SNP	EMS-induced mutation	-	-	3 prime UTR variant	-	-	-	-
Cadenza1820.chr1A.530257392	1A:530257392	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	G	909	-	-
Kronos3573.chr1A.530257462	1A:530257462	C/T	SNP	EMS-induced mutation	-	-	missense variant	G/D	886	0.12	-
Kronos2311.chr1A.530257462	1A:530257462	C/T	SNP	EMS-induced mutation	-	-	missense variant	G/D	886	0.12	-
Kronos2247.chr1A.530257468	1A:530257468	G/A	SNP	EMS-induced mutation	-	-	missense variant	P/L	884	0.01	-
Cadenza2034.chr1A.530257492	1A:530257492	C/T	SNP	EMS-induced mutation	-	-	missense variant	R/K	876	1	-
Kronos1071.chr1A.530257505	1A:530257505	C/T	SNP	EMS-induced mutation	-	-	missense variant	E/K	872	0.91	-
Cadenza0584.chr1A.530257517	1A:530257517	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	L	868	-	-
Kronos4443.chr1A.530257528	1A:530257528	C/T	SNP	EMS-induced mutation	-	-	missense variant	R/H	864	0.03	-
Cadenza1523.chr1A.530257560	1A:530257560	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	A	853	-	-
Kronos4212.chr1A.530257561	1A:530257561	G/A	SNP	EMS-induced mutation	-	-	missense variant	A/V	853	0.63	-
Kronos4012.chr1A.530257565	1A:530257565	C/T	SNP	EMS-induced mutation	-	-	missense variant	A/T	852	0.14	-
Kronos3710.chr1A.530257599	1A:530257599	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	E	840	-	-
Kronos2852.chr1A.530257602	1A:530257602	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	Q	839	-	-
Cadenza0203.chr1A.530257610	1A:530257610	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	L	837	-	-
Kronos2642.chr1A.530257617	1A:530257617	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	S	834	-	-
Cadenza1680.chr1A.530257626	1A:530257626	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	T	831	-	-
Kronos2821.chr1A.530257627	1A:530257627	G/A	SNP	EMS-induced mutation	-	-	missense variant	T/M	831	0	-
Cadenza0285.chr1A.530257662	1A:530257662	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	R	819	-	-
Cadenza1463.chr1A.530257694	1A:530257694	G/A	SNP	EMS-induced mutation	-	-	intron variant	-	-	-	-
Kronos283.chr1A.530257705	1A:530257705	C/T	SNP	EMS-induced mutation	-	-	intron variant	-	-	-	-
Kronos2663.chr1A.530257719	1A:530257719	G/A	SNP	EMS-induced mutation	-	-	intron variant	-	-	-	-
Cadenza1323.chr1A.530257758	1A:530257758	G/A	SNP	EMS-induced mutation	-	-	intron variant	-	-	-	-
Cadenza1522.chr1A.530257820	1A:530257820	C/T	SNP	EMS-induced mutation	-	-	missense variant	V/I	799	0.27	-
Cadenza0220.chr1A.530257836	1A:530257836	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	R	793	-	-
Cadenza1819.chr1A.530257887	1A:530257887	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	R	776	-	-
Cadenza0977.chr1A.530257987	1A:530257987	G/A	SNP	EMS-induced mutation	-	-	intron variant	-	-	-	-
Kronos4409.chr1A.530258095	1A:530258095	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	S	737	-	-
Cadenza0049.chr1A.530258106	1A:530258106	C/T	SNP	EMS-induced mutation	-	-	missense variant	G/S	734	0.28	-
Cadenza0388.chr1A.530258115	1A:530258115	C/T	SNP	EMS-induced mutation	-	-	missense variant	G/R	731	0.08	-
Kronos3161.chr1A.530258132	1A:530258132	G/A	SNP	EMS-induced mutation	-	-	missense variant	A/V	725	0.01	-
Kronos3716.chr1A.530258147	1A:530258147	G/A	SNP	EMS-induced mutation	-	-	missense variant	T/I	720	0	-
Cadenza1283.chr1A.530258185	1A:530258185	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	G	707	-	-
Cadenza0629.chr1A.530258194	1A:530258194	G/A	SNP	EMS-induced mutation	-	-	synonymous variant	D	704	-	-
Cadenza1508.chr1A.530258203	1A:530258203	C/T	SNP	EMS-induced mutation	-	-	synonymous variant	R	701	-	-
Cadenza1331.chr1A.530258205	1A:530258205	G/A	SNP	EMS-induced mutation	-	-	missense variant	R/W	701	0	-

Figure 6: Variant Table

The Variant Table provides an easy-to-use and quick access to the SNPs across a selected gene. It allows users to filter the display for specific SNP types or other criteria. The data can be downloaded as a spreadsheet, like all other data from Ensembl Plants.

Downloading the variation data

While possible, it would be quite tedious and time-consuming for users to look up every gene model and download the associated variant information one by one. To speed up this process users can make use of the BioMart tool (see How to use BioMart for a step by step guide). This allows downloading variants on different levels by either specifying a chromosome location or a list of gene models. It is a considerable time saver!

Using the variation data

At the beginning of the document we mentioned that the chief usage of variants is the development of genetic markers. There are many tools available to perform this task, e.g. Primer3web, PrimerBlast or BiSearch to name a few. However, these tools were designed for diploid organisms and as such do not take into account the presence of homoeologous sequences; In wheat, homoeologous sequences are >95% similar to each other, which in most cases would result in unspecific primer binding.

The PolyMarker tool was designed specifically with polyploid organisms in mind; it allows users to create genome-specific markers. Researchers only have to provide a list of bi-allelic SNP variants with flanking sequence (100 bp either side is sufficient), which can be obtained using the BioMart tool (see above). To learn more about the PolyMarker tool see the step by step guide (How to use Polymarker).