Improving gene models

As discussed in "Selecting TILLING Mutants", the consequence predictions of mutations on www.wheat-tilling.com are based on the IWGSC gene model. It is therefore very important that the gene model you are using is correct as otherwise mutations may not have the predicted effect. As such, it is sometime necessary to improve and reannotate you IWGSC CSS gene model. This document describes some tips for improving an incomplete/incorrect gene model and how to use this information to predict the consequences of mutations.

For more information on gene models and how to assess their quality see "Selecting TILLING Mutants"

a) Common problems with gene models

There are a number of problems that can arise with gene models and these are usually due to IWGSC scaffolds not spanning the entire length of the gene. As such, the most common problem will be either the 5' or 3' end of the gene has been missed due to the gene being right at the end of a scaffold. In this case the gene model will require extension at the 5' and/or 3' end (Figure 1).

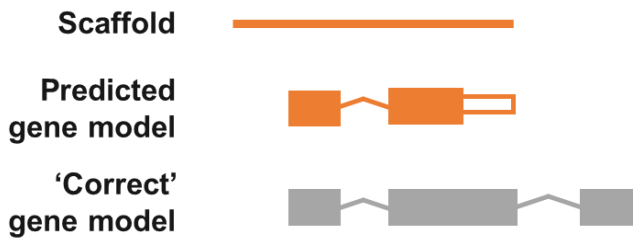

Figure 1: Gene model requiring extension

Another common problem are split gene models. This occurs when the sequence of a gene is split over several short scaffolds. This causes a number of individual incomplete gene models to be predicted when there should be a single full length gene model comprising all gene models (Figure 2).

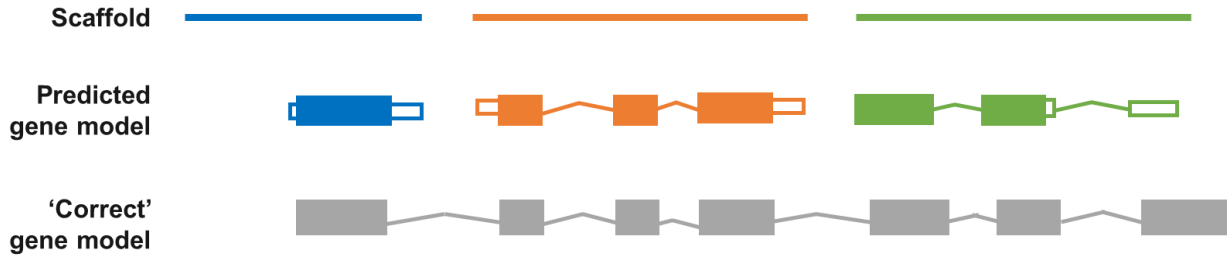

Figure 2: Split gene model

b) Improving the gene model

There are many things that you can do to improve a gene model, below we describe a non-exhaustive list of approaches.

TGACv1 assembly and gene models

The first step should be to BLAST the original IWGSC gene sequence against the new TGACv1 assembly (http://plants.ensembl.org/Triticum_aestivum/Tools/Blast?db=core). This assembly consists of much longer scaffolds than the IWGSC CSS assembly and so often the TGACv1 gene models are more complete than the IWGSC gene models. See "Genome assemblies"[JB(1)] for more information. The TGAC assembly and gene models are now the default on Ensembl Plants. This improvement will often improve the gene model sufficiently to proceed. It is important to remember that the TGAC gene model will still need to be put into the context of the IWGSC assembly in order to proceed with finding TILLING mutants.

If, for whatever reason, the TGAC gene models do not help with improving your gene model (e.g. the same problem with the gene model exists or the gene model is not present) there are other steps that can be taken. For these it is still useful to obtain the TGAC scaffold that your gene is on as this provides valuable extra sequence to work with.

Comparison with other homoeologues

Whilst the gene model for your particular gene may be incorrect or incomplete, this may not be the case for the other two homoeologues of your gene (see "Introduction to Wheat Growth" for more information on homoeologues). Looking at the gene models for other homoeologues and extending your gene model based on these can be a good approach (see "Ensembl plants primer" for how to find the homoeologues of your gene). The TGAC scaffold can be used to provide the additional sequence information necessary for this.

Comparison with other species

You can also extend your protein based on information from other closely related species that have better annotated genomes, (e.g. Rice, Barley, Maize). To do this, carry out a protein BLAST (e.g. on http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). If, when looking at the alignments with other species, you see that in most cases the protein sequences from other species all extend further at the 5' or 3' end, or have some extra sequence in the middle, this could be an indication that the real sequence of your gene also has these. Of course, it is important to note that there can be interspecies variation so this should be used in combination with other evidence.

Ab initio prediction

Another option is to use an *ab initio* gene prediction tool, such as FGENESH (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind). For this, you could use part of the TGAC scaffold that your gene is in as the input sequence for this. It is important to note that these are only predictions and are not always biologically correct. It is therefore important to combine this with other evidence including from other species and the other homoeologues.

c) Re-annotating the gene model to find TILLING mutants

Once you have your improved gene model you will need to take some extra steps to identify and predict the consequence of mutations in the TILLING mutant resource (see "Selecting TILLING Mutants").

Use of parseSNP to identify mutations

We will use CODDLE and parseSNP to predict the consequence of mutations according to our improved gene model.

1. Obtain the scaffold name of the IWGSC CSS scaffold that your gene is on.
2. Use the scaffold name as the search query on www.wheat-tilling.com to download an excel file with all the mutations in this scaffold
3. Use this to create a variants text file: this should consist of two columns. The first column should contain all the mutation positions following the format REF base, scaffold position, MUT base. The second column should contain the name of the mutant line containing this mutation. See Figure 3 for an example. This should be saved as a .txt file.
    a. *Hint: to generate the information for column 1 you can concatenate the "ref", "pos" and "mt" columns from the exported csv file. Paste the formula "=concatenate(H2,E2,I2)" into the excel output to produce this format for the first mutation in row 2.*

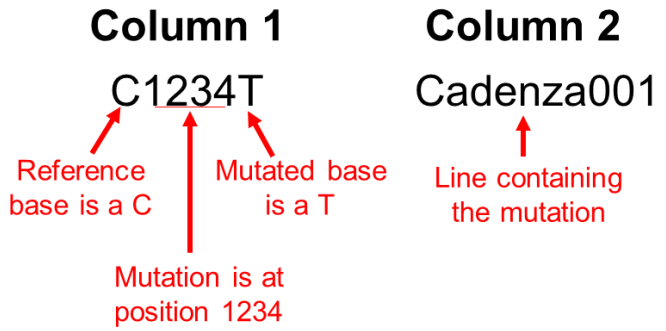

Figure 3: Variants text file format
Column 1 of the variants text file should contain the position of the mutation and column 2 should contain the name of the mutant line.

4. Download the genomic sequence of the *whole* IWGSC CSS scaffold by clicking on the name of the scaffold in the output table
5. Make a coding sequence for your gene using fragments of the genomic sequence NB. It is very important that there are no SNPs between the genomic sequence and coding sequence
6. Go to the CODDLE website: http://blocks.fhcrc.org/~proweb/input/
7. Copy and paste your genomic sequence into the 'Submit genomic sequence' box (Figure 4a)
8. Copy and paste your CDS into the 'Submit cDNA sequence' box (Figure 4b)
9. Click 'Begin Processing' (Figure 4c)
10. You will first be taken to a page with predicted 'blocks' in your protein, click the 'Proceed with PARSESNP' button
11. On the PARSESNP page upload your variants text file (Figure 5a), set the 'No. of variants to enter by hand' (Figure 5b) to 0 and click 'PARSE-SNPs in Your Gene' (Figure 5c)
12. On the next page (your variants file displayed in a table on the webpage) click 'submit'.

13. The result will be a table containing a lot of information, including the predicted effect of all the mutations in your variants file on your gene of interest.
14. From here you can proceed with selecting your mutant line as normal ("Selecting TILLING Mutants").

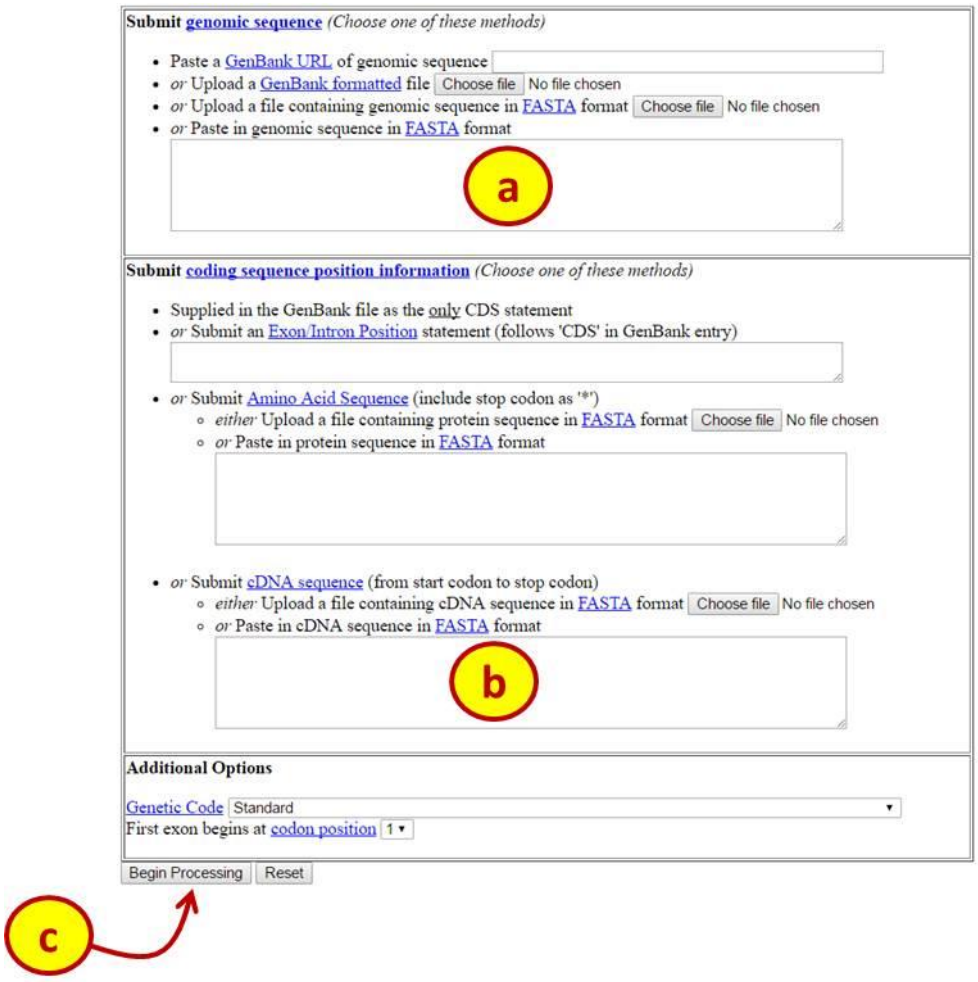

Figure 4: CODDLE homepage

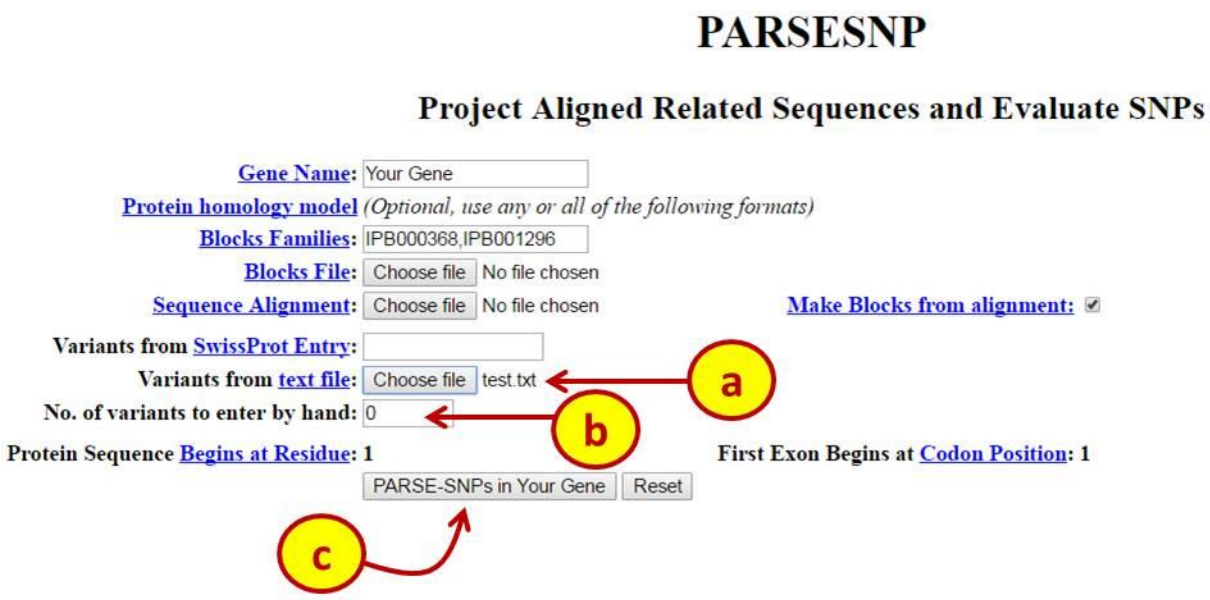

Figure 5: PARSESNP page