Introgression Browser tutorial

EU-TransPLANT Training course: exploring plant variation data

Jan-Peter Nap, Sven Warris, Saulo Alves Aflitos

Access at EBI:

Family name A-I, please use: http://10.7.243.39:10000
Family name J-R: please use: http://10.7.243.40:10000
Family name S-Z: please use: http://10.7.243.41:10000

Note: Access via Firefox or Chrome, NOT via Internet Explorer

1. Background

In this part of the workshop, you will use a novel way to visualize whole genome variation data, in a tool called Introgression Browser (iBrowser). Using extensive sets of genome data, the tool is suitable for SNP mining, analysis of genome structure as well as pedigree analysis. Based on SNP distributions in genomic windows, the tool allows identification of introgressions from -for example- wild relatives to rationalize and speed up of plant breeding and marker assessment. Notably in recurrent backcross strategies, detection of introgressed regions is important to be able to follow desired and/or unwanted sections (genes) of a genome. The latter is known as linkage drag. Prior to be able to visualize and inspect for introgression, several computationally intensive calculations have to be done; you are referred to your local bioinformatics expert to set up these calculations with your own data. Details can be found in the accompanying Plant Journal paper (Aflitos *et al.* (2015) Plant Journal **82**, 174-182) and its supplementary material. Also the analysis of whole genomes (e.g. 600 Arabidopsis accessions or 60 tomato RILs) is resource intensive. Therefore, the data you get now is limited to a relatively short section of 4 Mb of chromosome 6 of 84 different accessions of tomato, including wild species such as *Solanum pimpinellifolium* (*S. pimp*) and others, in addition to various tomato (*S. lycopersicum*) accessions, with the tomato Heinz genome as reference. Note that the approach and power of iBrowser easily extends to all chromosomes and many more accessions, allowing fast surveys of genome structures and events. For the purpose of this tutorial, however, a limited data set is presented.

Paper: http://onlinelibrary.wiley.com/doi/10.1111/tpj.12800/abstract
Manual: https://github.com/sauloal/introgressionbrowser/wiki

2. Functionality

You will first have to familiarize yourself with the basic data and the basic functionality of iBrowser.

1. To get started, open Firefox or Chrome (IE is more recalcitrant, unfortunately) and start iBrowser as indicated by your tutor. The start-up screen is very basic (Fig. 1), not to waste resources on frivolities.

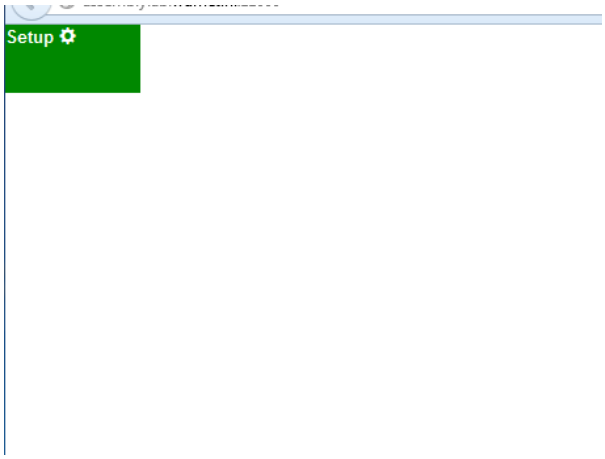

Fig 1. Basal opening screen iBrowser. Mouse over the green square to get the next screen.

2. Move your mouse (mouse over) to the green area and see a menu pop up (Fig 2.). See that if you move your mouse outside the green box, you go back to Fig. 1.

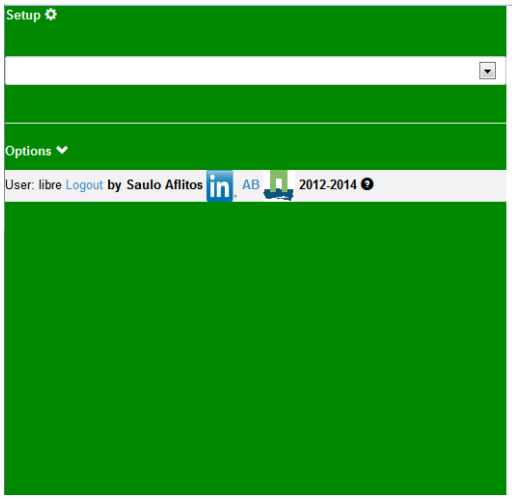

Fig 2. First interactive screen of iBrowser for data selection.

3. Go with the mouse to the first white bar (mouse click in the middle of the bar will do) and see that you have only one database to choose from: Tomato 84 – 50 Kb – Introgression (Fig 3.) Select this one database with a mouse click. Realize that many more databases are available in the full blown version of the iBrowser (Fig 4).

Fig 3. Selection of data in iBrowser: selection of chromosomes and/or chomosome sections.

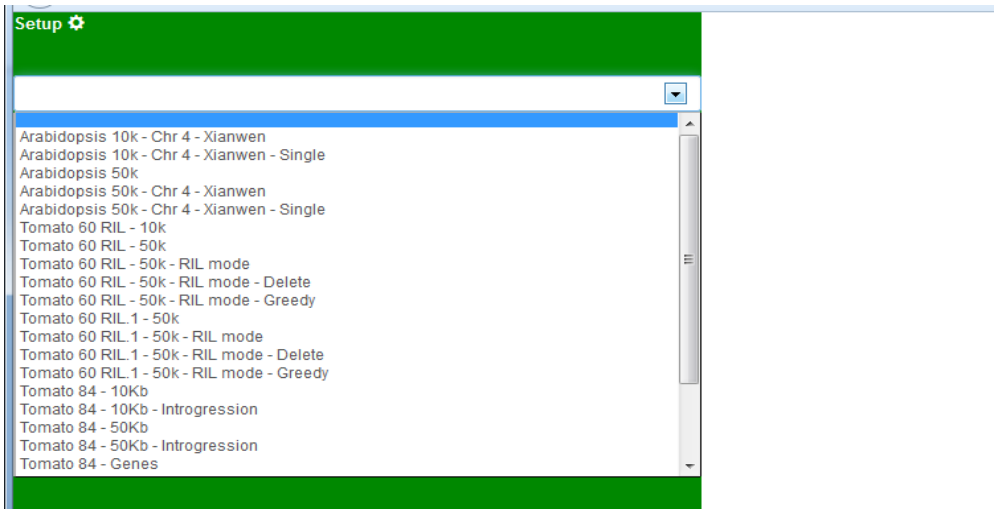

Fig 4. Partial list of the data available in the current full version of iBrowser.

4. Go with the mouse to the next bar and see that you have only one chromosome to choose from: SL2.40ch06. This is chromosome 6 of *Solanum lycopersicum* CV Heinz 1706, assembly version 2.40. Select this chromosome with a mouse click (Fig 5.)

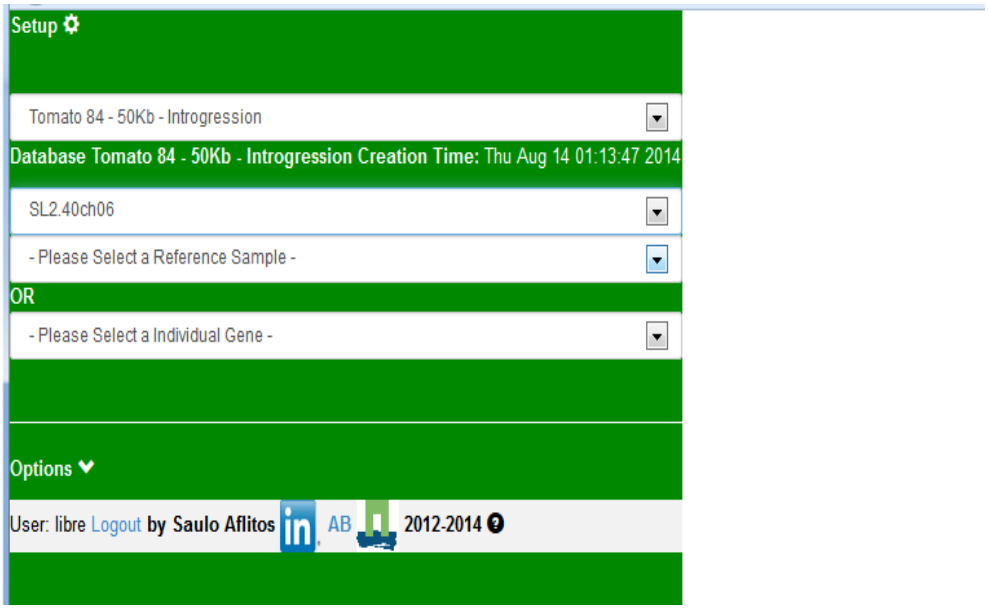

Fig 5. Continued data selection in iBrowser: selection of a reference genome to browse the sequence distance against.

5. Now select a reference sample. Look for the reference 'ref' (is the last in the list) and select this one. This reference is the Heinz tomato genome, the best assembled and annotated tomato genome to date.

Q1. *In addition to Heinz, you can select any accession as (relative) reference. How many different species are included in this particular dataset?*

6. Make three selections: (a) select colour to your taste (default= red, please select red for the purpose of this tutorial), (b) select 'show row' to display the species names and (c) select 'cluster rows alphabetical'. Click on send. The system shows you it is busy for you.

7. Now click on 'send' and blow the figure up so that it fills the screen using the magnifier plus button. These actions should result in Fig 6. Such a figure is referred to as 'heat map'.

Fig. 6. Output of iBrowser: heat map

8. The information condensed in Fig 6. may seem a bit overwhelming at first. It will need some time to get accustomed to. At the top, below the green bar labelled ‘set up’, the most relevant information is displayed. Assuming you have selected ‘red’ as colour for display, in red to yellow, the distance of the accessions is plotted. In this case, this distance runs from 0.0 to 0.5. This distance is based on the number of SNPs (actually, the number of polymorphic positions) in the whole dataset relative to Heinz.

9. In green, the number of SNPs is given; this is the number of SNPs per 50 kb region. In this case, the number runs from 1785 to 2884.

Q2. *In what colour the number of SNPs is given when you choose the option ‘rainbow’ for visualization?*

Q3. *Why does the plot start at 36.0 Mbp and not at 1?*

10. When you click on any green square at the top information about that region appears on the right hand site of the top bar.

Q4. *What is the end position of the 50 Kb fragment with the least amount of SNPs?*

Q5. *What happens if you deselect ‘show row’? When is that handy?*

11. The colour scheme gives a fast visual impression of the differences between accessions and allows the visual assessment of genomic regions. If you click on any square, you get the detailed information of that particular region of that particular accession.

Q6. *What does it mean that most lyc accessions are largely yellow?*

Q7. *Which S. lycopersicum is most close to the Heinz reference, yet differs visibly?*

12. When you double click on any square, or go to the set-up section and select a section, you get the distance tree showing the relationships of all accessions for this

particular section of the chromosome. In addition, the full sequence alignment and alignment matric are given (for the number crunchers in your lab). Go in the set-up panel to the fragment with start 36850001 and end 36900000 and generate the tree. The upper part of this tree should look like Fig 7. Is it probably easiest to generate and inspect the PNG file format. Try to do so.

Q8. *What does this tree tell you about the three S pimp accessions ?*

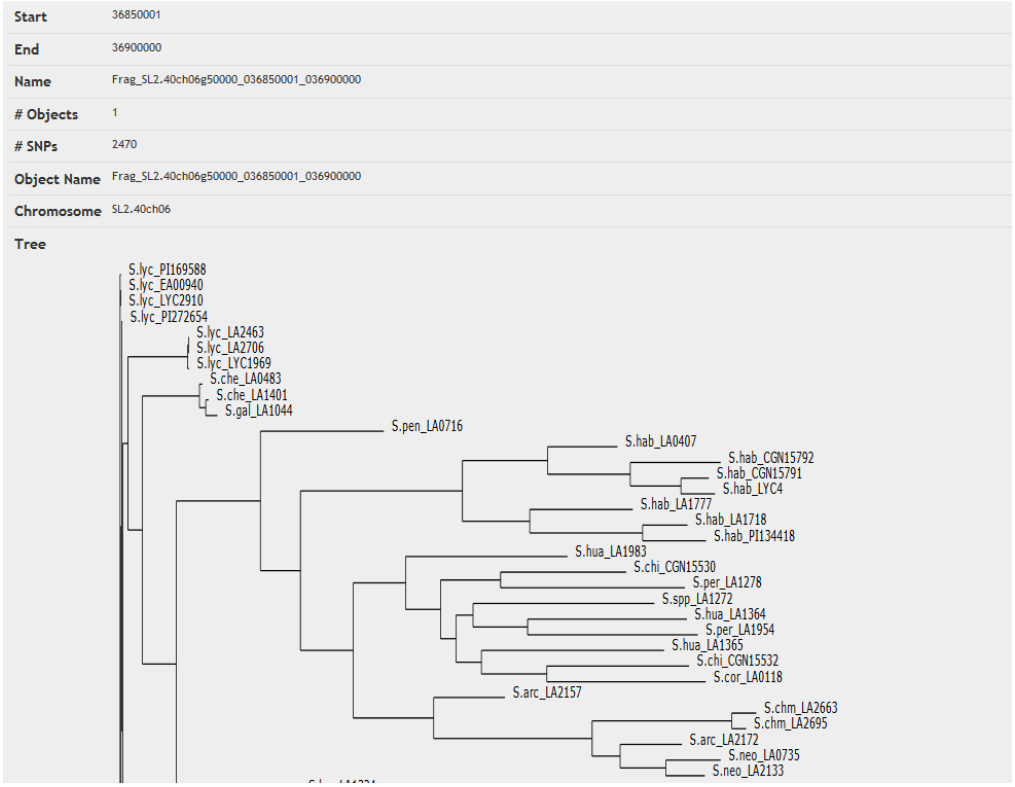

Fig. 7. Part of the distance tree generated by i-browser.

13. The above has covered the basics of the tool. Go back to the heat map. There are two other ways to cluster the data: either ‘per chromosome’, to analyse a single chromosome, or the ‘tomato 84 tree’, which is useful to compare chromosomes. The first was created by concatenating all SNPs of the individual chromosome in a multi-alignment-like manner and cluster; the second involved all SNPs of the whole genome. Play with these settings and get a feel for the different plots.

14. Note that when you select another reference than Heinz, the data plotted reflect the distance to the reference selected. It allows evaluating the distances between accessions. Assume the genotype of Heinz is AAAA, *S. arc* is ACGG and *S. pimp* is ACTG (mapped relative to Heinz). With Heinz as reference, the distances of *S. arc* and *S. pimp* are 3 and 3, whereas relative to *S. pimp*, the distance of *S. arc* is 1 (and Heinz is obviously again 3).

3. Comparative genomics

15. iBrowser is used to investigate different aspects of genomes and genome structures. In this part of this tutorial, we give few examples.

Q9. *Now focus on accession tomato* S. spp. *LA1272. In what country was this accession identified and what species was it identified as? (Hint: use Google and TGRC)*

Q10. *Analyse the sequencing data with the various options of iBrowser. What is your conclusion?*

Q11. *Explain the result. What could have happened here?*

Q12. *What is your opinion on S. gal. LA1044?*

Q13. *Zoom in on LA2706. What happened here?*

Q14. *Issues in comparative genomics are to be able to distinguish between introgressions (introduction of material from other accessions) and recombinations (reshuffling of material within a given genome, such as an inversion). How would you use iBrowser to make this distinction?*

Q15. *In case you really have too much time: what genes are located at the 5' border of the region identified in Q13? [Hint: go to solgenomics jbrowse]*