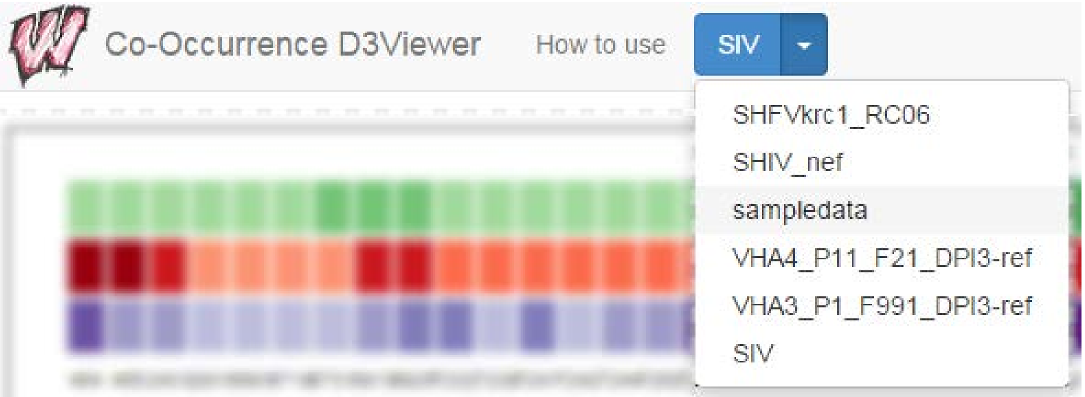
Alper Sarikaya (sarikaya@cs.wisc.edu); 2/17/15
http://pages.cs.wisc.edu/~sarikaya/static/cooccur/d3viewer.html#SIV

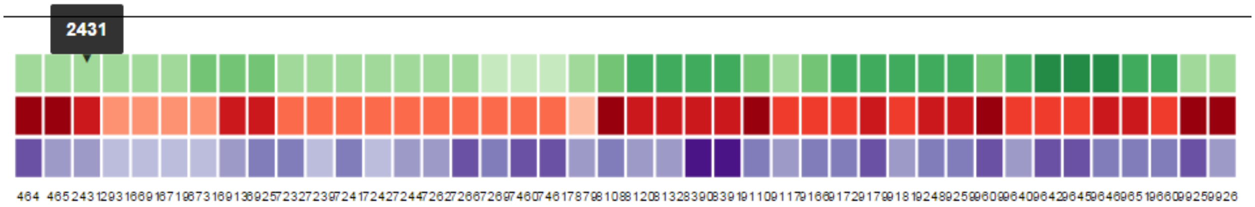
This is a little introduction to finding 'interesting' co-occurrences in viral populations where short read sequences (commonly found in NGS, *.sam/.bam files) can be used to establish co-occurring variants.

The visualization can present any arbitrary NGS data. In order to prepare the data for the visualization, the provided *.sam* file is passed through a program that generates metrics and counts in a format easily processed by the visualization.

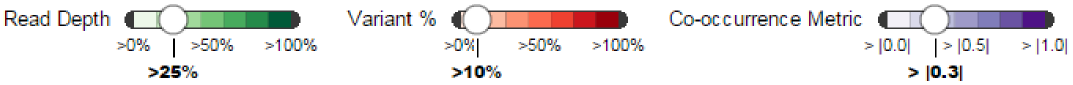
There is a 'dataset' switcher at the top. VHA3/4 designate H5N1 datasets (ferrets) at particular timepoints (three days post-infection), while SIV is the one *bam* file we had (*CY0334_54wk_Contig.bam*).



Once the data loads (usually around 100MB in abstracted binary format), the browser parses it and dumps a summary at the top. Click any of the columns to show the co-occurrences that exist with that position.



The colors correspond to the three sliders at the bottom, which trim down all possible co-occurring pairs.



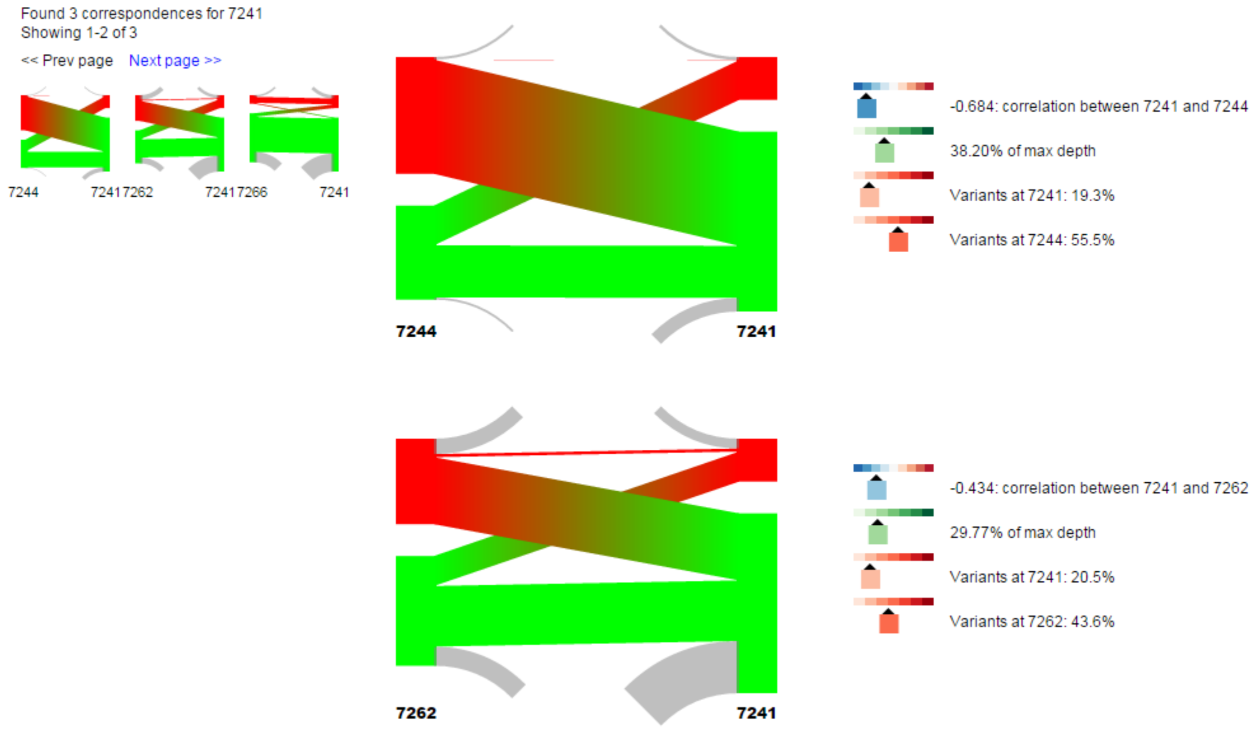
- Read depth is % of the maximum read depth seen at any pair of positions,
- Variant % is a minimum threshold for which each position in every pair must match, and
- The co-occurrence metric is an 'interestingness' metric that measures meaningful co-occurrence of variants (or reverse co-occurrence):

$$\Pr(j_{\text{var}}|i_{\text{var}}) - \Pr(j_{\text{var}}|\neg i_{\text{var}}),$$

where the above equation will go to 1 when variants at *i* influences variants at *j*, the above equation will go to -1 when variants at *i* influences non-variants at *j*, and go to 0 when no interaction is seen at *j* that depends on whether or not *i* is variant.

Dragging any of the sliders will force a refiltering operation of the dataset and reset the summary up top.

Click any one of the boxes up top (a particular position) to see what co-occurrences that position has to other positions. You'll see the view fill up with some information:



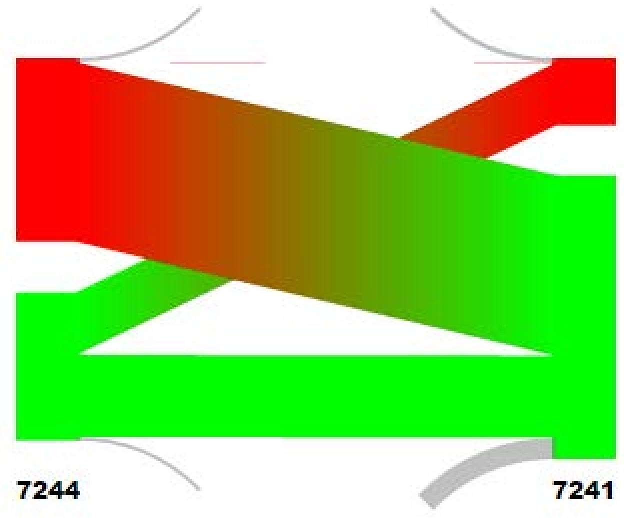
Found 3 correspondences for 7241
Showing 1-2 of 3

-0.684: correlation between 7241 and 7244
38.20% of max depth
Variants at 7241: 19.3%
Variants at 7244: 55.5%

-0.434: correlation between 7241 and 7262
29.77% of max depth
Variants at 7241: 20.5%
Variants at 7262: 43.6%

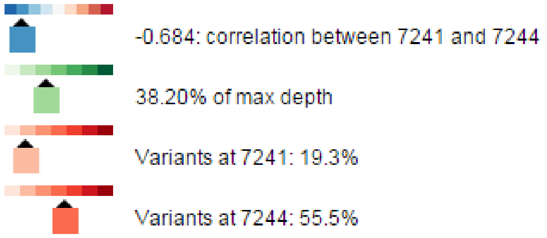
At most two pairs are shown in the middle of the screen. If there are more than two filtered correspondences, a "small-multiples" display is shown on the left; click on any one to bring it into the main view. You can also 'page' forward and backward through the found correspondences.

In the above view, let’s look at the first co-occurrence displayed.



The height of the bars corresponds to the number of reads at each location (7244 on the left, 7241 on the right). Red are variant reads (they didn’t match the given reference given in the FASTA file), green are non-variant. Gray drop-off chords are those reads that appear at one position but do not overlap with the paired position.

The visualization shows that the large majority of the variant reads at 7244 are non-variant at 7241, and likewise, variant reads at 7241 are non-variant at 7244. This is an example of ‘reverse’ co-occurrence, which the co-occurrence score shows:



-0.684: correlation between 7241 and 7244

38.20% of max depth

Variants at 7241: 19.3%

Variants at 7244: 55.5%

I’d love to hear any feedback. Is this sort of view useful? What would make it more useful, or what isn’t working?

-Alper