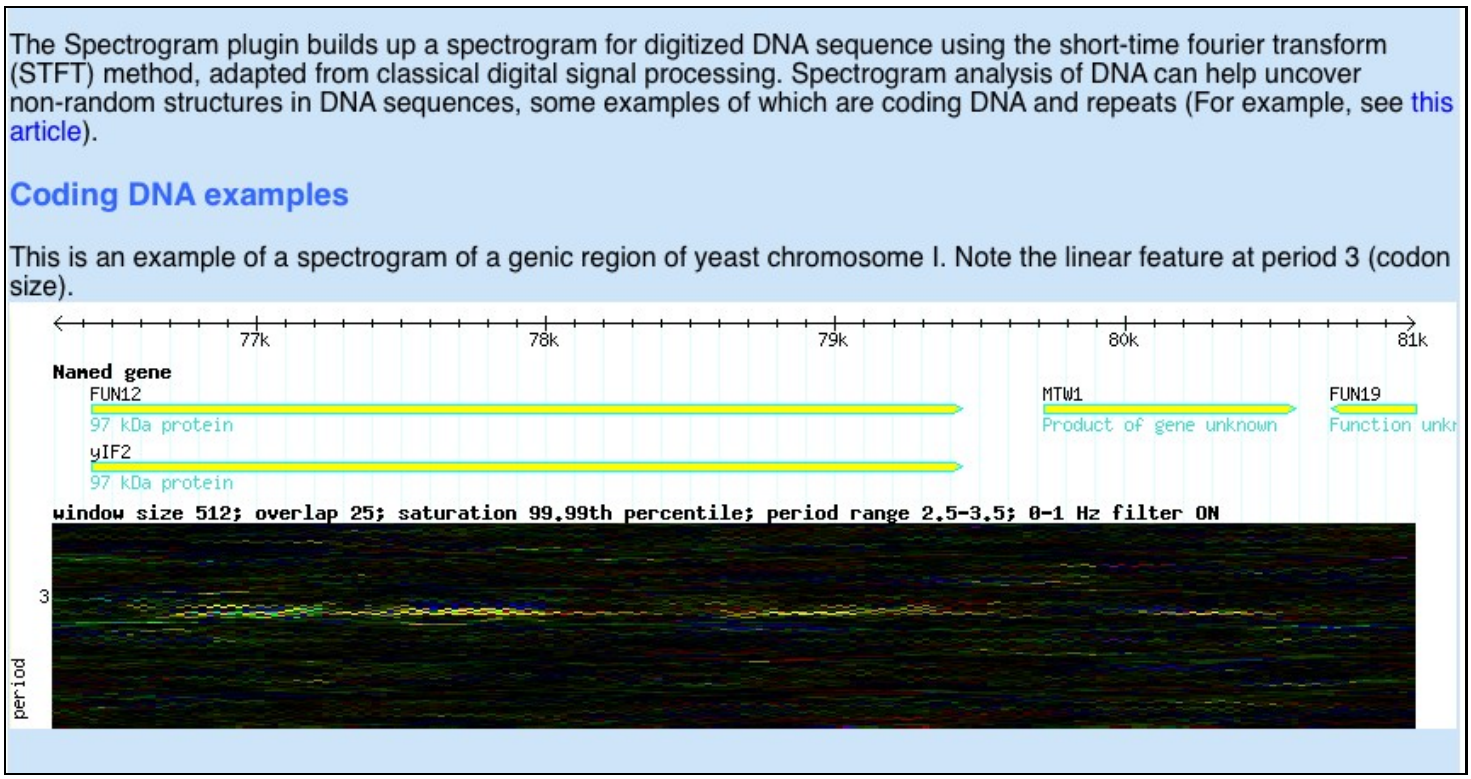
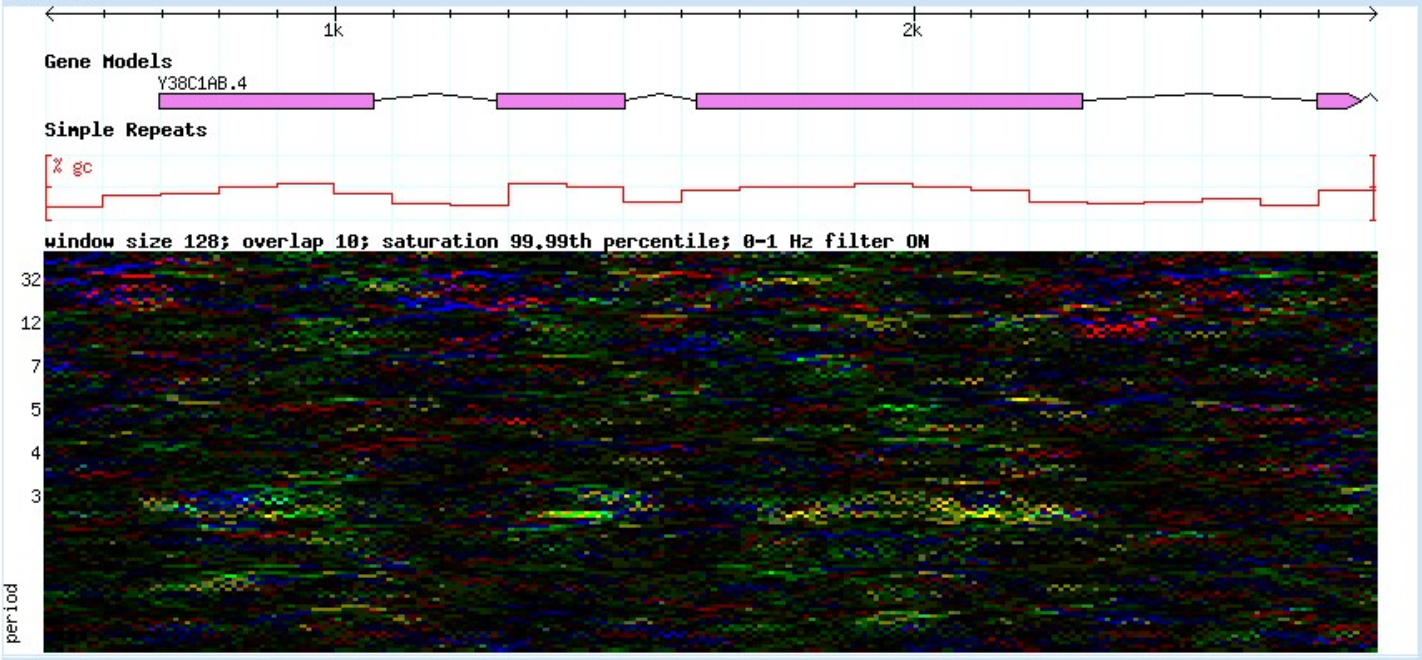
Spectrogram.pm is a GBrowse plugin written by Sheldon McKay. It draws DNA spectrograms of DNA of digitized DNA sequences using techniques borrowed from the digital signal processing world. Graphical rendering is accomplished using the HSV color space. The color of the spot on the spectrogram corresponds to the dominant nucleotide at that "frequency" and position and the intensity corresponds to the strength of the signal. DNA spectrograms reveal non-random sequence composition, the two most common examples of which are coding DNA and repeat sequences.

- Coding DNA has a signal due the the non-random occurence of nulceotides in codons and appear as a line in the spectrogram with a period of 3.
- Repeats have a characteristic ladder like appearance in the spectrogram.
- Follow this link for background reading of digital signal processing of DNA



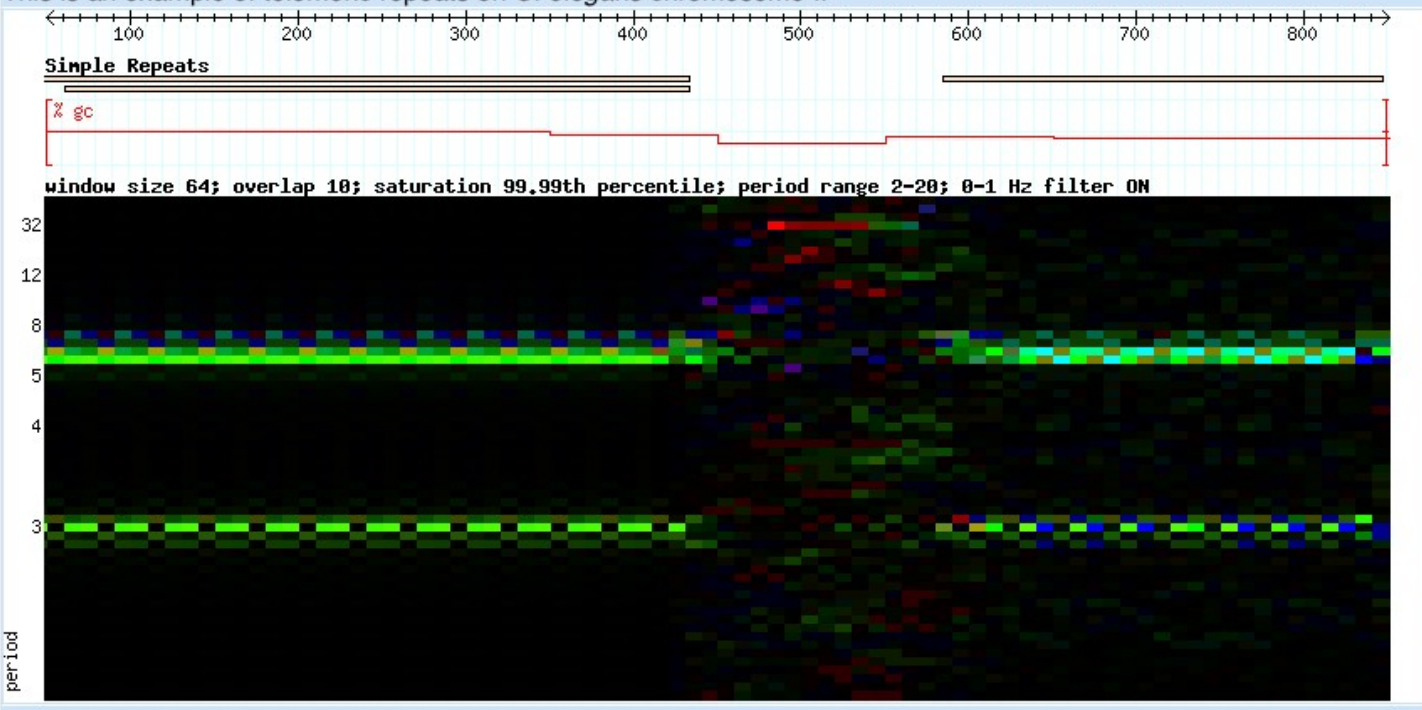
This is an example of a portion of *C. elegans* predicted gene Y38C1AB.4. Note the differences between exons and introns.



Repeats

Repeats cause a ladder-like series of horizontal lines. Short repeats, such as telomeric repeats, are most visible with small window sizes. Longer repeats, such as minisatellites, are best seen with larger window sizes.

This is an example of telomeric repeats on *C. elegans* chromosome I.



How is the DNA spectrogram calculated?

A sliding window of variable size and overlap is used to calculate the spectrogram, which is displayed graphically as a track in the genome browser. Each window is a subsegment of DNA and corresponds to a 'column' in the graphical display of the spectrogram. The window slides along the sequence, from left to right, at a set increment, which corresponds to the column width.

The spectrogram refers collectively to all of the rows and columns seen in the graphical display.

The spectrogram has n rows, where n is the number of bases in the window. Each row corresponds to a discrete 'frequency' from 0 -> n-1.

An arguably more intuitive way to relate this to DNA sequence to calculate the 'period' (n/frequency*2). If we see a feature in the spectrogram at period x, there is a non-random structure with a periodicity of x nucleotides. The chief example of this would be coding DNA at period 3.

The DNA sequence is converted from analog to digital by creating four binary indicator sequences:

```
  G A T C C T C T G A T T C C A A
G 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
A 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1
T 0 0 1 0 0 1 0 1 0 0 1 1 0 0 0 0
C 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0
```

The magnitude of the discrete fourier transform (DFT) is calculated seperately for each of the four indicator sequences. The algorithm used is the fast fourier transfrom (FFT; via Math::FFT), which is much faster than the original DFT algorithm but is limited in that only base2 numbers (128, 256, 512, etc) can be used for window sizes. This is necessary to make the spectrogram calculation fast enough for real-time use.

For graphical rendering, each transformed sequence is assigned a color (A=blue; T=red; C=green; G=yellow). The colors for each base are superimposed on the image. In a given spot on the spectrogram, the brightness corresponds to the magnitide (signal intensity) and the color corresponds to the dominant base at that frequency/period. If no single base predominates, an intermediate color is calculated based on the relative magnitudes.

The spectrogram is visible as a track in the generic genome browser. Please note that the calculations and graphical rendering are computationally intensive, so the image will take a while to load, especially with larger sequence regions and/or small increments for the sliding window.

After you have launched this plugin, the spectrogram will continue to be calculated in the main gbrowse display until you turn off the 'Spectrogram' track.

The plugin was written by Sheldon McKay (mckays@cshl.edu)