Michael Bradshaw

Mile Stone Results For The Variance Visualizer

Figure 1.

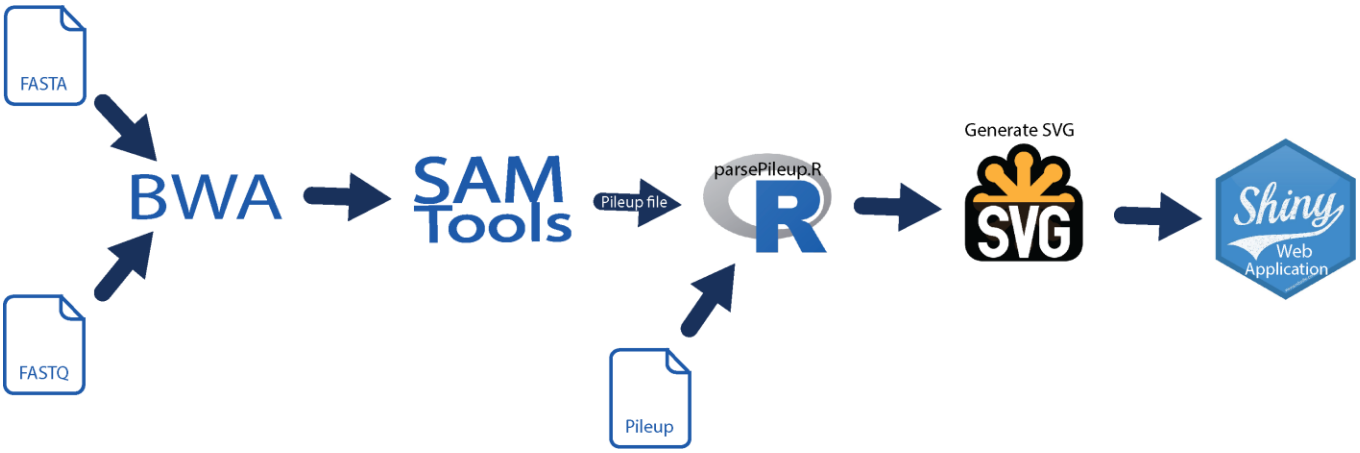

Figure 1 outlines the flow of information through the Shiny application. There are two difference places of input for the process to start at. The program can receive either fastq read files and a fasta reference sequence files which will be piped through BWA and SamTools to generate a pileup file. Or the user can simply upload a readymade pileup file. The pileup file is then parsed using an R script for relevant information. Once parsed, the data can then be used to generate SVG images of the given sequences. The SVGs are then passes off to Shiny and rendered as crystal clear image in the web application.

Figure 2.

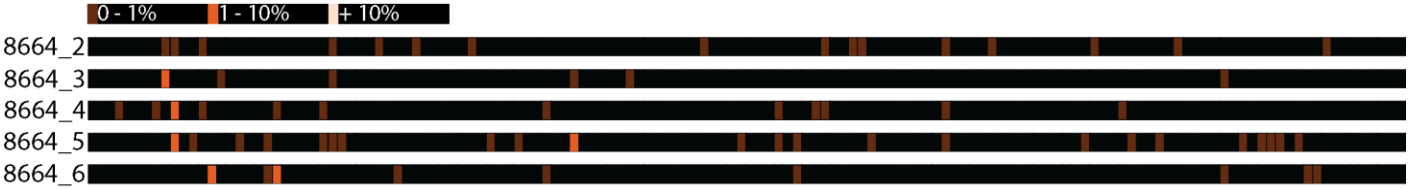

Figure 2 shows the variance of a small segment of the ITS region for 5 difference lichen samples from the species Rhizoplaca suhasini. Unlike usual sequence visualizations that focus on the nucleic acid at each position, this image reveals the amount of base pair variance at each position. Regions with absolutely no variance are left black; otherwise the brightness of the marker is representative of the amount of intragenomic variation found at that location.

Time Log

Task	Over All Time
Pileup Parser	6 Hours
SVG Creator	3 Hours
Flow Diagram	2 Hours
Total	11 Hours

The flow chart provides a map of what the application is actually accomplishing under the hood. It also provides a visual representation of the types of input it can handle. Pileup files can directly be loaded into the application and result in quick rendering of sequence images. However pileup files do not have perfectly consistent file formats. The software used to generate the pileup is curcial, thus an alternative input format is available. As an alternate input type, the user may also provide raw reads in the fastq format and a reference sequence in a fasta file. This option is much slower due to several expensive steps required to create the pileup file. However fasta and fastq are much more common files types to have on hand and do not suffer from major inconsistencies as pileups do.

Pileups are then parsed for information relating to base pair sequence and variance. This provides sufficient information to create a scalable vector graphic (SVG) of the sequence. SVG was chosen as the image file type for 3 reasons. First this type of file can easily be produced; SVGs are really just text XML documents. Secondly because they are XML in a web application SVGs can also have functionality linked to make them interactive to hovering, scrolling and clicking. Third their scalability, it does not matter how big or small you make an SVG it will always be crystal clear.

In Figure 2 you will notice a good deal of the brown markers speckled seemingly randomly across the sequences. To call its dispersal random is accurate, with variation occurring in 1% or less of the reads mapping to this sequence it most likely that these variances are the results of Illumina sequencing error rather than true nucleotide variance. Of more interest are the positions marked in orange to more easily catch the eye on a black background. These markers represent sequence positions with percent variance above the Illumina error threshold. It is important to note that several of these variances line up across samples. The higher level of variance and location consistency across samples suggests that these suhasini have none identical rDNA repeats in their genome.