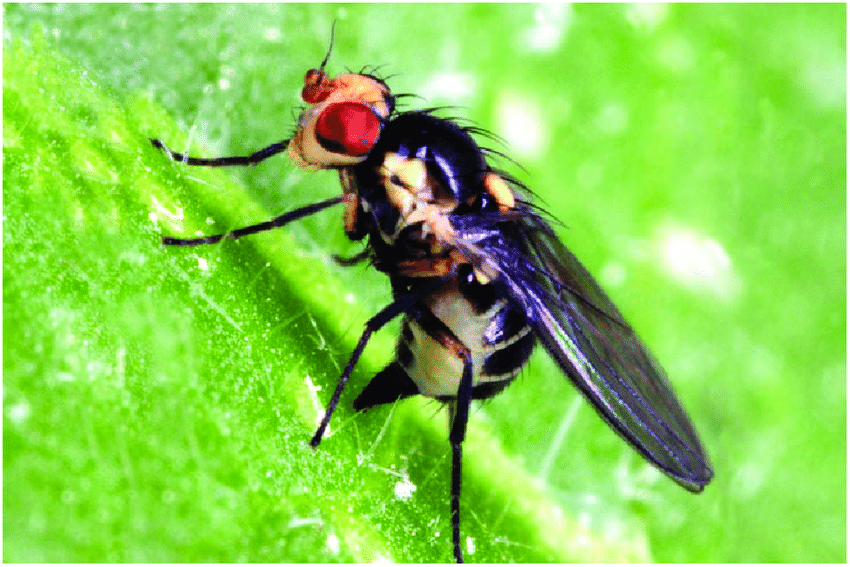
The Variance Visualizer

By Michael Bradshaw

**Introduction**

The influx of Next-Generation Sequencing (NGS) in the past decade and a half has revolutionized the use of DNA in the study of all forms of life. One technique that has become made possible by NGS is DNA barcoding, also referred to as a DNA fingerprint 1. Barcoding “uses standardized 500- to 800-bp sequences to identify species of all eukaryotic kingdoms”2. In the past, species have been distinguished from one another using key anatomical structures but distinguishing small insects, plants and fungi based solely on physical appearance can be difficult and ambiguous. For example see the three different species of vegetable leaf miners below:



Liriomyza huidobrensis - Liriomyza trifolii - Liriomyza sativae

Images courtesy of Research Gate, EPPO Global Database and GrandenDrum.org

Morphologically extremely similar, these three species are often mistaken for each other on quick examination3. But why is it important to tell these 3 insects apart? Well, one of these insects has spread from the Americas to Asia and has begun to destroy tomato plants in India, the second largest producer of tomatoes in the world3. When visual identification failed, researchers were able to determine which species of leaf miner it was terrorizing these tomatoes using DNA barcoding3.

Identifying insects visually is difficult enough, even for trained professionals. Now imagine trying to identify lichen forming fungi visually.

Images courtesy of lichenportal.org, dr-ralf-wagner.org and waysoflichenment.org

The first two images are the same species, Rhizoplaca melanophthalma the second is Rhizoplaca haydenii. As forms of life get smaller our ability to accurately classify them becomes more and more difficult. This mandates the use of more precise methods than visual comparison.

Barcoding has proven itself to be a useful tool in numerous fields. But not all species have universally accepted barcoding. Experts studying the Fungal kingdom still have some ongoing debate about what the standardized barcoding region ought to be. One of the problems of choosing a barcoding sequence for all bacteria is finding a region that is present in all species that is unique enough to identify closely related species from each other. For fungus the ITS region of the rDNA has been proposed as the barcoding region2.

Prior to proposing the ITS region as the fungal barcode, several other sequences were investigated. Because of its ease of PCR amplification and initial success in species identification, the ITS region was identified as a strong candidate for the fungal barcode2.

There are potential problems with using the ITS region as the barcode. The rDNA cistron (which contains the ITS region) is an abundantly repeated region of the genome and there is concern that the cistron may be subject to intragenomic variation4. Too much intragenomic variation could invalidate the ITS region as a viable barcoding sequence. In order to confirm if the ITS region can be useful for barcoding an investigation of its intragenomic variation is needed; this is the question this project aims to address.

Most of the analytical tools for gathering and analyzing the information about intragenomic variation already exist. First NGS reads need to be aligned to a reference sequence. A number of software tools already exist for accomplishing this. The most effective algorithms out there currently use implementations of the Burrows Wheeler transformation for both fast and accurate read alignments5,6. Two such implementations widely used today are BowTie7 and BWA6. After alignment, reads need to be indexed and put into a more usable format to reveal coverage and base pair consistency. This can largely be accomplished with Samtools8. The last step needed for evaluation is to place this information in a human readable format. Many software tools have been created for visualizing alignments. However most of these tools have major dependency issues when installing them and are not compatible with modern operating system versions. One large exception to these generalizations is Integrative Genomics Viewer (IGV) by the Broad Institue9. IGV is easy to install, simple to use and works on the current Mac and Windows operating systems. But IGV lacks the functionality to visualize intragenomic variation. In IGV you can hover over a single base pair you can see the percentage of each nucleic acid were mapped to that exact position, this is the core information needed to establish what kind of variation is occurring in the data. In order to gather information about intra genomic variation from IGV, a researcher would be required to physically mouse over each base pair and calculate the percent variance at each location to check for values above the sequencing error rate. For just one sequence this is a long and tedious process. Now imagine having to do it for multiple different samples, it sounds like an unpleasant and error prone process. Add cross referencing those positions to cross species diagnostic markers and you have a problem that is not well suited for IGV and its users. There is the possibility of simply extending the functionality of IGV to automate the calculations and cross referencing, is has the potential to be a great solution for the variation problem. But the IGV source code does not appear to have been built in a way that supports easy extension by external developers. Because of this, visualizing the information necessary to answer the ITS variation question requires a new, previously non-existent tool.

The purpose of this project is to create a tool, The Variance Visualizer (VV) for visualizing a sequence of DNA and its intragenomic variation. VV will be useful in answering the question of whether or not the ITS region is consistent, unique, and not overly variant within an individual. VV will also allow users to visualize multiple samples at the same time. The accuracy of this tool will be validated in two ways. The first way will compare the sequence image it creates to a handmade image displaying the intragenomic variation of the exact same data sets. The second way will compare the relative abundance of variation/sequencing error of Illumina and PacBio reads. Illumina reads are known to have a relatively low sequencing error, 1%10,11, while PacBio tends to be much higher, experiencing 22% sequencing error11,12.

**Proposed Materials and Methods**

*NGS Reads:*

Illumia reads from 31 individuals representing 7 different fungal species have been provided by Dr. Leavitt from the Biology Department at Brigham Young University. These are the samples that will be used to validate the software in comparison to PacBio reads (also provided by Dr. Leavitt). Additionally, these 31 samples will be used in the final evaluation of the ITS region as a barcoding region.

*Alignment:*

Alignment of NGS reads will be done using the BWA6 software. Due to its speed, accuracy, ease of installation and use, BWA appears to be an excellent candidate for this task.

*Data Formatting:*

The SAM files generated from BWA will be indexed and converted into a more useful file type using Samtools8. Specifically, Samtools will be used to generate pileup files, text files representing the alignment containing the coverage at each position and the base pair contained by each read at that position. Pileup files can then easily be parsed using an R script to gather and generate the relevant information.

*Visualization:*

One of the major problems I have encountered while trying to use software for visualizing sequences is simply getting it to install. The variance visualizer will avoid this problem entirely since it is a web based Shiny application. The sequence images themselves will be created as scalable vector graphic (SVG) images. Because SVGs are text based vector images, they can easily be generated and remain crystal clear regardless of how large or small the user chooses to make them. Additionally, if the user should desire to customize the output in unforeseen ways, SVGs provide a format in which similar elements can all be simultaneously changed relatively easily or converted into any other picture format such as .png, .jpeg or a .pdf.

I will be working on all parts of this alone. If pitfalls should occur I will tackle them with the help of Stack Overflow.

**Results:**

*Figure 1* will be the schematics diagram outlining the flow from user input to the final output.

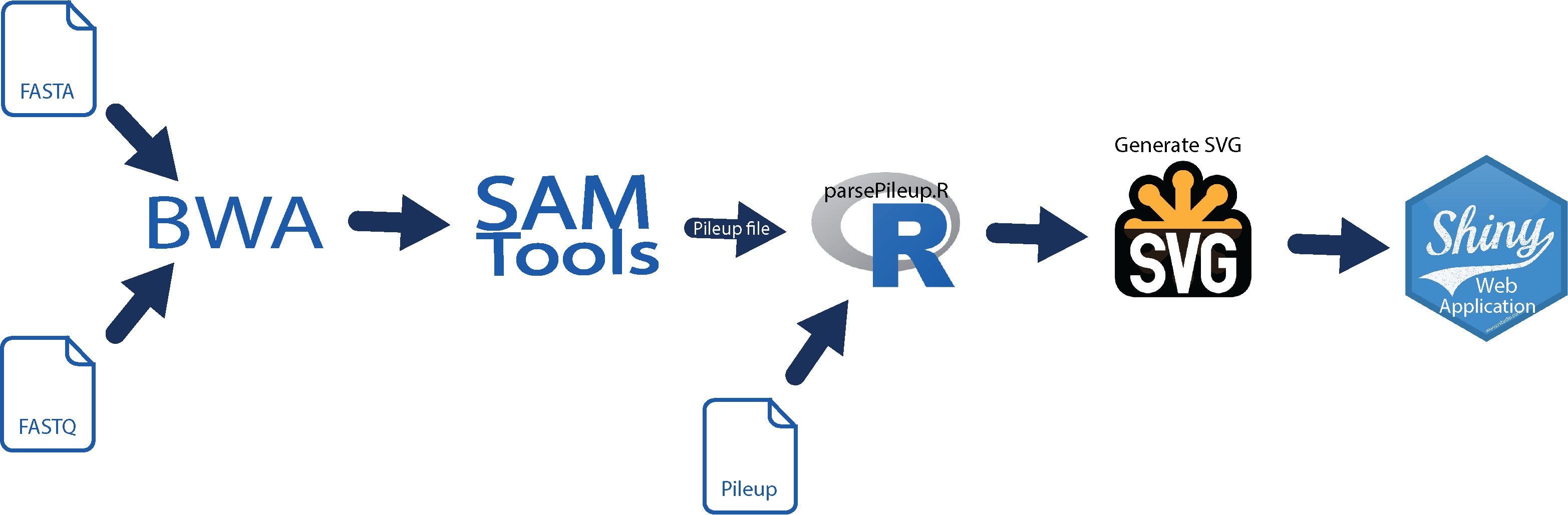
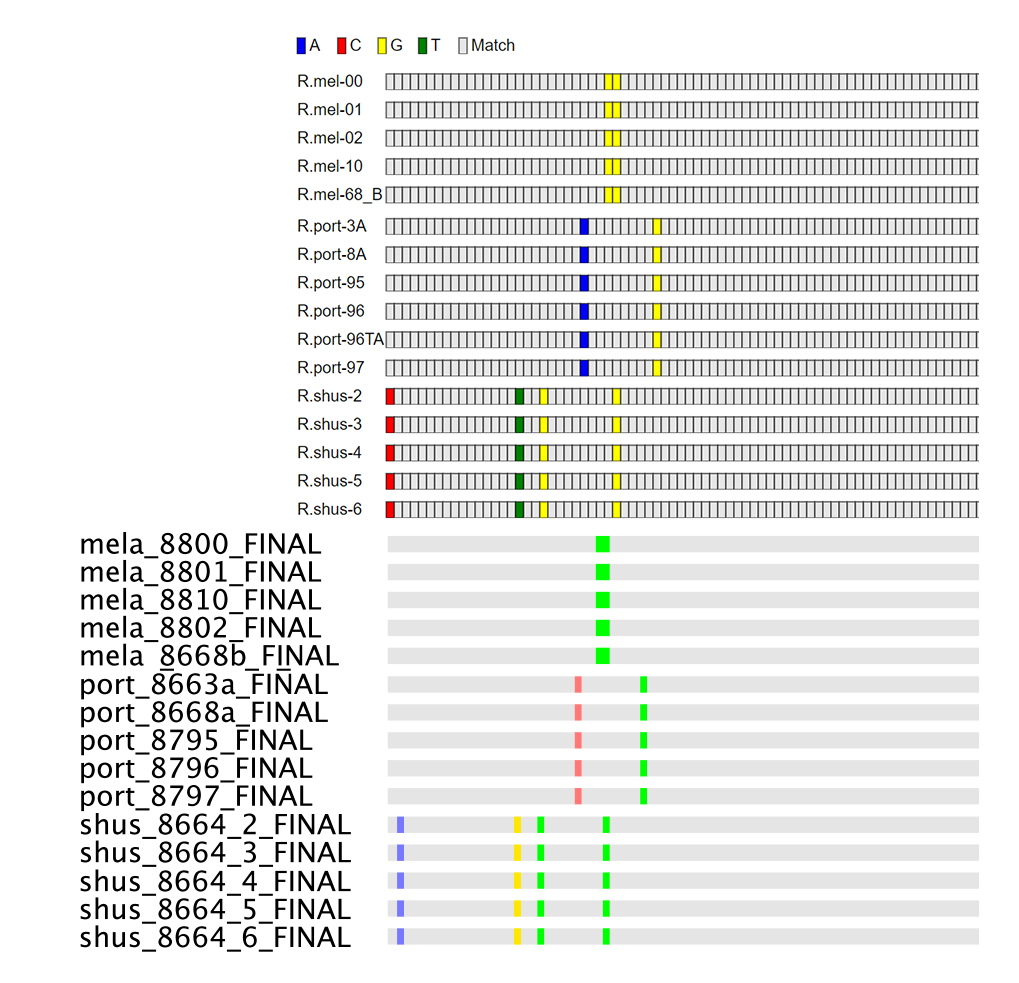
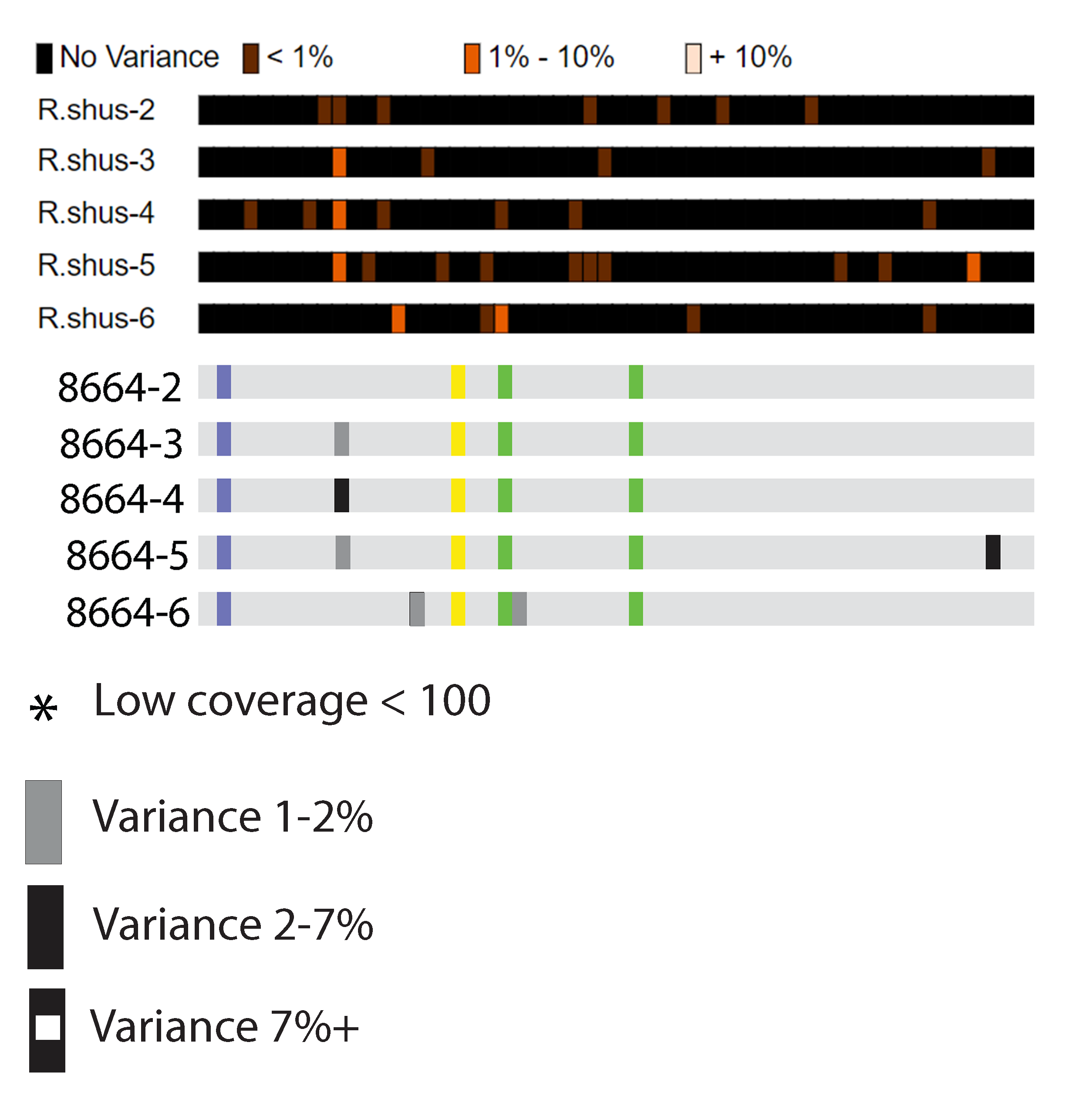


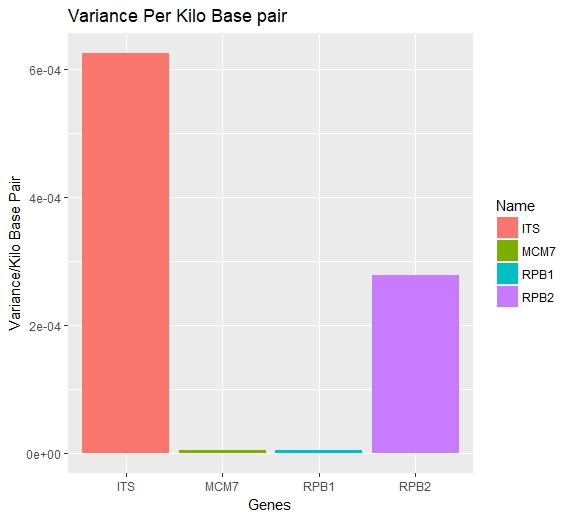
Figure 1 outlines the flow of information through the Shiny application. There are two difference places of input for the process to start at. The program can receive either fastq read files and a fasta reference sequence files which will be piped through BWA and SamTools to generate a pileup file. Or 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.



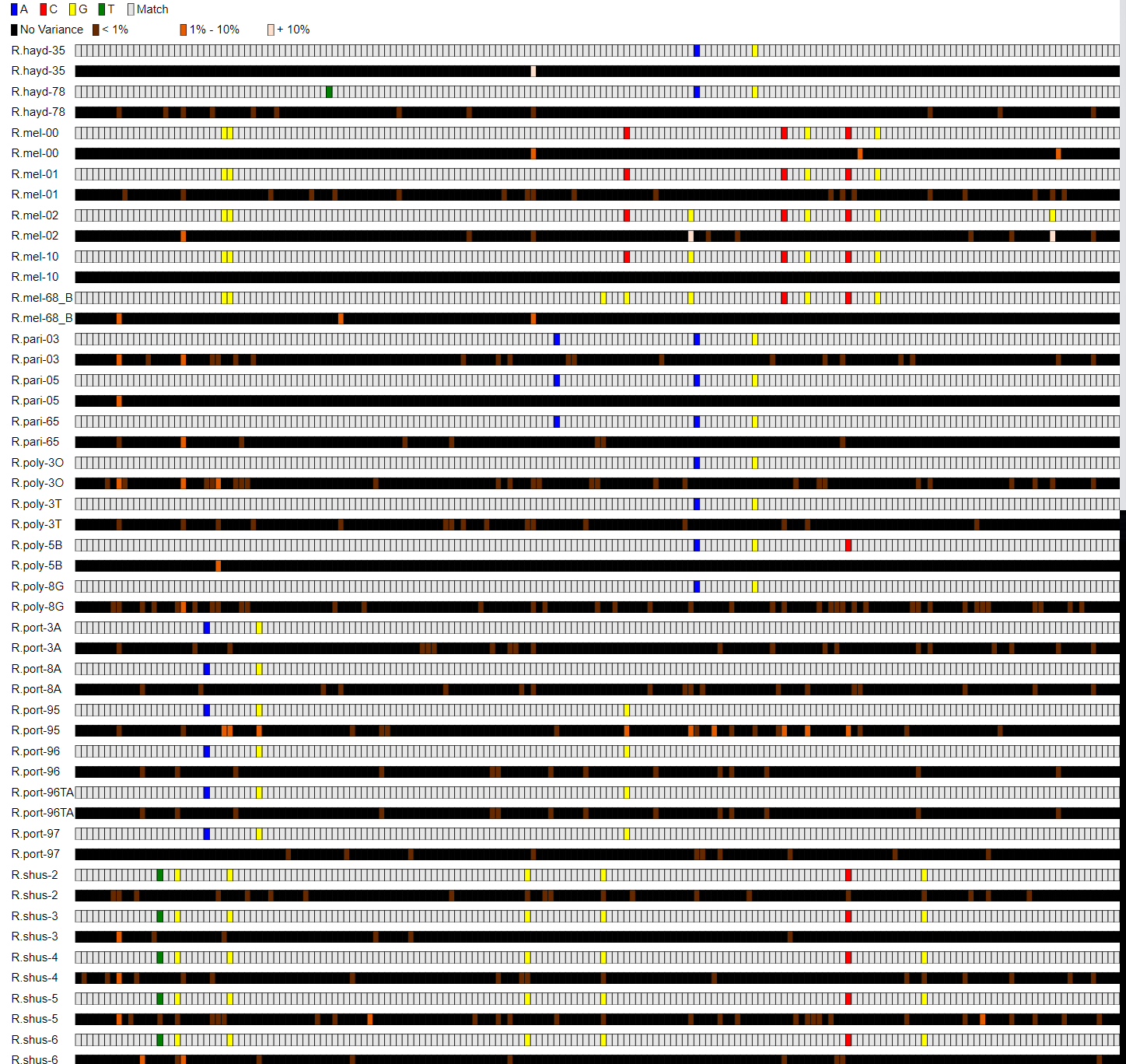
*Figure 2* shows multiple sequence alignments from VV (the top portion of the image) and from Geneious (bottom portion). Both algorithms were run with the setting to display each samples difference from the consensus sequence generated by all the samples. For Geneious this is all the defaults settings. For VV the image was generated using the consensus setting and default trim settings (600 front and back). It is important to note in this image that the two programs found and color the same diagnostic areas (positions within the sequence that differ species to species and are there for useful for species classification). Geneious and VV both identify 7 separate diagnostic areas. The position and consistency of these areas is a 100% match between programs. Given that the two programs produced the same results when provided the identical input provides one can be confident that the basic functionality of VV is correct.



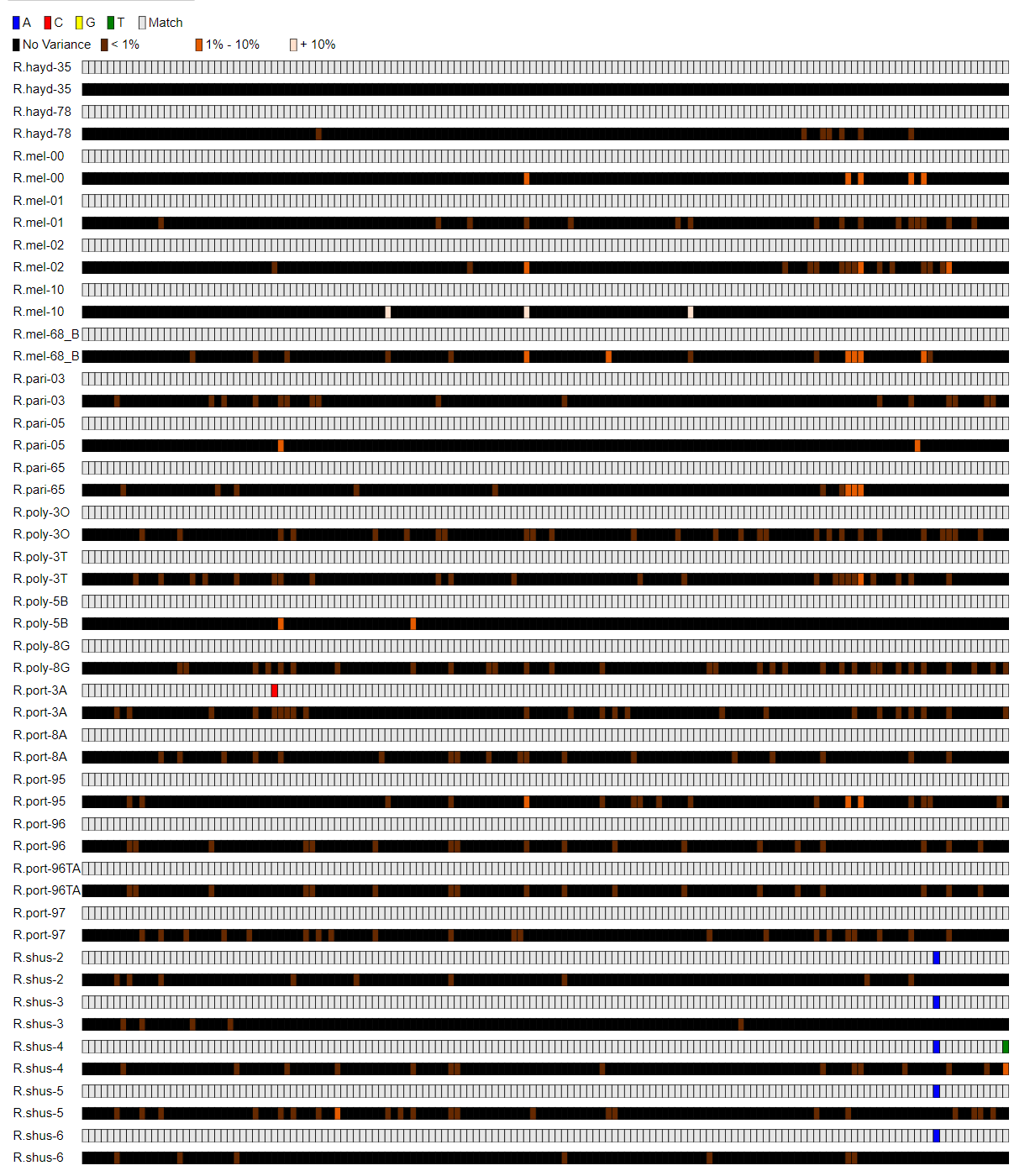
*Figure 3* features two displays of variance, the top image coming from VV the second a hand edited version of a Geneious alignment. The grey and black squares in the Geneious image were manually added to represent variances calculated separately in a previous unpublished study.



*Figure 4* shows the percent variance per kilo base pair for the ITS region and MCM7, RPB1 and RPB2. The three genes the ITS region is compared to are the same evaluated in the original study2 to propose a fungal barcoding region. MCM7 and RPB1 show absolutely no intra genomic variation while RPB2 displays some and the ITS region displays more than double than RPB2. The original study investigating these 4 sequences as barcoding regions determined that the ITS region was the optimal barcoding regions due it how easily it can be amplified using PCR2. However as pointed out by later researchers4,13, the original study failed to address the concerns of intra genomic variation present in highly repeated regions, like the ITS. The results shown here begin fill this gap in the original research demonstrating the ITS regions does have a good deal more variance than the other genes.



*Figure 5* shows the diagnostic and variance images of the ITS1 (the first segment of the ITS region). The visualization shows there are 21 diagnostic areas for distinguishing the 7 species from each other. Of the 21 areas only 3 overlap with an area of variance that falls above the sequencing error rate. One of these areas occurs in R.mel-2 while the other two are found in R.port-95. Both of these individual samples come from species where 5 different samples were provided. R.port-95 and R.mel-2 are both unique compared to their siblings of the same species. Other none of the other samples from the same species contain nearly as much variation as these two and contain now at all in the potentially problematic diagnostics areas.



*Figure 6* shows the diagnostic and variance images of the 5.8S region (the second segment of the ITS region). Within the 5.8S region there is only 1 diagnostic area which overlaps with 0 areas of notable variance.

*Figure 7* will be a multiple sequence alignment displaying variance of the ITS2 region of the rDNA cistron for all provided specimens.

Figure 8

* Variance Image of the single copy genes vs one of the ITSs

*Figures 5-7* will contain the 3 segments of the ITS region. Because of how long the entire ITS region is, I will break it up into its three sub sections for the ease of visualization on paper. These three figures are what will allow us to evaluate the ITS region’s level of intragenomic variation and subsequent use as barcoding region.

In the event that there are areas of consistent intragenomic variation we can compare these to the cross species diagnostic markers to check for overlap. In the event of overlap with key diagnostic markers we may be left to conclude the ITS region is not useful as a barcoding region.

References

1. Bi, Y. *et al.* Chloroplast genomic resources for phylogeny and DNA barcoding: A case study on Fritillaria. *Sci. Rep.* **8,** 1–12 (2018).

2. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci.* **109,** 6241–6246 (2012).

3. Firake, D. M., Sankarganesh, E., Sharma, B., Firake, P. D. & Behere, G. T. DNA barcoding confirmed the occurrence of invasive vegetable leaf miner, Liriomyza sativae Blanchard (Diptera:Agromyzidae) in Northeast India. *J. Asia-Pacific Biodivers.* **11,** 56–60 (2018).

4. Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D. & Thomas, W. K. Intra-Genomic Variation in the Ribosomal Repeats of Nematodes. *PLoS One* **8,** 1–8 (2013).

5. Burrows, M. & Wheeler, D. A block-sorting lossless data compression algorithm. *Algorithm, Data Compression* 18 (1994). doi:10.1.1.37.6774

6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

7. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** (2009).

8. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

9. IGV (Integrative Genomic Viewer). Integrative Genomics Viewer. *Broad Inst.* **29,** 24–26 (2013).

10. Quail, M. A. *et al.* A tale of three next generation sequencingplatforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 1–13 (2012).

11. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11,** 759–769 (2011).

12. Ono, Y., Asai, K. & Hamada, M. PBSIM: PacBio reads simulator - Toward accurate genome assembly. *Bioinformatics* **29,** 119–121 (2013).

13. Kiss, L. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. *Proc. Natl. Acad. Sci.* **109,** E1811–E1811 (2012).