AftrRAD Frequently Asked Questions

Below are some questions frequently asked about AftrRAD. For a list of error messages commonly encountered during AftrRAD runs, and possible solutions, see the document "Common_Errors".

What platforms does AftrRAD run on?

Mac and Linux. Runs are typically faster on Macs. Maybe a PC version in the future?

Are there any additional programs needed to run AftrRAD?

Yes. First, you need a Perl interpreter. This is probably already on your computer. Second, you need R. Good chance you already have this too. Finally, you'll need the alignment programs ACANA and mafft. These (and R, if you need it) are easy downloads, and information for getting them is in the User Manual.

I think I've identified a problem with AftrRAD – what do I do?

We've tried to identify and work out any kinks in the program, but realistically, there are probably still some bugs in it. If you think you might have found a problem, we would greatly appreciate you letting us (and everyone else) know either on the google group or by e-mailing Mike directly at sovic.1@osu.edu. If you can find a solution to the problem and pass it along too, we would appreciate it even more. Anyone providing bug fixes will be rewarded with free access to an even better version of the program for the remainder of their scientific endeavors.

What are the inputs for an AftrRAD run?

The default for AftrRAD is to expect two things: undemultiplexed raw sequence read file(s) in fastq format, and a barcode file for each fastq file (this assumes that barcodes, or molecular identifiers, occur at the beginning of each sequence read). Information for how to format the barcode file is in the User Manual. If your data are already demultiplexed (you have a raw fastq file for each sample), simply put these samples in a directory named 'DemultiplexedFiles' and run AftrRAD.pl with the argument 'dplexedData-1'. These demultiplexed data usually result from Illumina indexed runs.

What type of output files does AftrRAD produce?

There are a few default output files produced from an AftrRAD run. These include a 'SNPMatrix' file which contains all individual SNPs genotyped in each sample, a

'Haplotypes' file that is similar to the SNPMatrix file, but treats multiple SNPs within the same read as a haplotype, and a 'Monomorphics' file, which includes all of the monomorphic loci identified, and read counts for each of these in each sample. Note that the loci printed to these files can be filtered based on the proportion of samples genotyped at each locus, and on the location of the SNPs along the read. Sets of output files can be produced with multiple combinations of these filters. Scripts are provided to further convert these data to input for programs for downstream analyses. Information on each of these formatting scripts is available in the User Manual. We're adding options for formatting as needs arise in our lab. Check the most recent version of AftrRAD for current formatting options available. Contributions of formatting scripts are welcome!

How long does it take to complete an AftrRAD run?

This depends. The number of polymorphic loci in the dataset is likely to be the major factor driving run times. After that, the total number of reads has an effect, and to a lesser degree, the number of samples. The operating system used can have a major effect when there are a large number of polymorphic loci in the dataset. This is because the ACANA aligner has been optimized for Mac OS X systems, but not for Linux, and therefore runs much faster on Macs. If the dataset has relatively low levels of polymorphism (and in turn, relatively few alignments to do), then the differences between the platforms will be tempered. If there are a large number of alignments to do, then runs on Linux can get quite long. As a couple of examples, runs using our rattlesnake data, which have relatively low polymorphism, run on our Mac in somewhere between 20 minutes and a few hours, depending on the number of samples, which range from 10's to 100's (the largest number of samples we've run thus far is ~300). In contrast, the highest polymorphism dataset we've tried (bats) took almost two days on the Mac, and would have taken > 1 week on the Linux. This dataset required > 1 million alignments.

Can AftrRAD analyze paired-end data?

Not currently.

How does AftrRAD handle low-quality base calls?

Low quality base calls are reflected by low Phred scores in the fastq file. AftrRAD is relatively conservative in terms of removing these – more so than other programs such as Stacks and PyRAD. These other programs make efforts to "salvage" information from reads containing low quality calls that AftrRAD does not. Specifically, AftrRAD removes any reads that contain a single base with a Phred score below the chosen threshold. Because of this, sequencing runs with overall low quality may be difficult to analyze in AftrRAD. One option is to set the Phred score threshold very low so that most (or all?) of the reads are retained, and then allow the minDepth parameter to remove error reads. We have not tested this, but it should work OK, as long as average read depths are

relatively high. We do not recommend this approach if average read depths are low (i.e. <10).

How does AftrRAD make genotype calls?

AftrRAD uses a simple binomial test to make genotype calls. The test is applied when a locus has two alleles with nonzero counts in an individual, and helps ensure that any error reads that may have passed through upstream filtering are eliminated. Critical p-values for the tests can be set with a command line argument in Genotype.pl. See the documentation for this part of the script for more information.

How do I evaluate the quality of my RADseq dataset?

There are a lot of factors that can influence the overall "quality" of a RADseq dataset. Another way to look at this is that, unfortunately, there are a lot of ways a RADseq dataset can "go wrong". We've tried to incorporate a number of options into AftrRAD to help give a sense of the quality of RADseq datasets. Check the section of the User Manual "Evaluating Your RADseq Dataset" for specific suggestions on how to assess your dataset.

What if I have barcodes of different lengths at the beginning of my reads?

This shouldn't be a problem. AftrRAD can accept in-line barcodes of different lengths. It will trim the reads from the end accordingly. For example, if some of your barcodes are 5 bp long, and others are 6 bp long, then the reads with the 5 bp barcodes will have an extra base at the end that will not have been sequenced in the reads with the 6 bp barcodes. AftrRAD will trim this extra base off of the 5 bp barcode reads. Note that AftrRAD also trims the barcodes themselves down to the shortest length, so in the hypothetical scenario where you had the 5 bp barcode TAAGA and the separate 6 bp barcode TAAGAC, the program would incorrectly treat all of the reads from these two samples as a single sample.

I see loci in my dataset that have a greater number of indels than the value I set for numIndels. Why?

There are two different alignments performed in the analysis. The numIndels argument applies to the first alignment step, in which pairs of unique reads in the dataset are aligned. The pairwise comparisons that meet the criteria for the number of indels allowed and the percent identity are then assembled into candidate loci. A second alignment is then performed on each locus, which may have more than two alleles. In this alignment step, there is no limit on the number of indels that the aligner puts in the locus, so it is possible that the final loci have a number of indels that exceeds that value set for the numIndels argument.