Bridging the Gap Between Large-Scale Data Sets and Analyses: Semi-Automated Methods to Facilitate Amplified Fragment Length Polymorphism Scoring and Data Analyses

McGreevy Jr TJ[1,2], Markert JA[1,3,4], Grear JS[1], and Nacci DE[1]

[1]US EPA, Atlantic Ecology Division, Population Ecology Branch, Narragansett, RI 02882, [2]University of Rhode Island, Natural Resources Science, Kingston, RI, USA,
[3]US EPA, Ecological Effects Research Division, Population Ecology Branch, Cincinnati, OH, [4]Department of Ichthyology, American Museum of Natural History, New York, NY

ABSTRACT

Conservation biology studies are often focused on non-model organisms that lack previously developed molecular markers. Amplified fragment length polymorphism (AFLP) markers can be developed at a relatively low cost and in a short period of time, which can make them ideal markers for generating large data sets for species at risk. However, manual scoring of AFLP markers is prone to data entry errors, time intensive, and subjective. Recently, the objectivity of scoring AFLP DNA fingerprinting data produced from automated sequencers has been greatly improved with the development of AFLPScore v1.3 (Whitlock et al., 2008). We developed an R script to convert the raw peak intensity data output from GeneMarker® v1.6 (SoftGenetics LLC®, State College, PA) to a format compatible for AFLPScore. We developed a second R script to convert the binary genotype output generated by AFLPScore to a format compatible for AFLP-Surv v1.0 (Vekemans, 2002). We applied this method to investigate the correlation between AFLP genetic diversity values and extinction risk using replicated experimental populations with manipulated levels of genetic diversity subjected to environmental stress. Specifically, the proportion loci polymorphic and expected heterozygosity (H_j) were calculated using AFLP-Surv and a large AFLP data set for mysid shrimp (*Americamysis bahia*). We also demonstrated the reliability of estimating initial H_j values using harmonic-mean effective population size and ending H_j values by comparing results derived from these estimates to experimental data. The two R scripts we developed reduced the opportunity for data entry errors and expedited the analyses of our large AFLP data set. The line code for the R scripts also could be manually adjusted to convert AFLP data between other commonly used computer programs.

AFLP Background and Limitations

- Development: relatively low cost and in a short period of time
- Performance: similar to microsatellites and single nucleotide polymorphisms for addressing most population genetics questions
- Operation: capillary electrophoresis DNA sequencing of fluorescently labeled AFLPs produces output (peak intensities) that must be translated into DNA fingerprint patterns
- Overcoming current limitations: semi-automated systems for analyzing AFLP data can reduce error rates due to subjectivity and increase consistency within and across data sets.

METHODS

Stepwise AFLP data processing

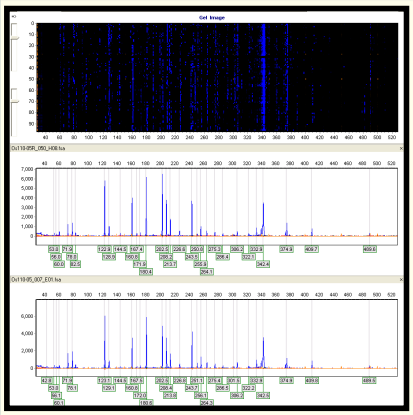
- R script (1) used to convert the raw peak intensity data output from GeneMarker to a format compatible for AFLPScore. AFLPScore was used to minimize the genotype scoring error and maximize the number of AFLP markers retained.
- R script (2) used to convert the binary genotype output generated by AFLPScore to a format compatible for AFLP-Surv.
- AFLP-Surv was used to calculate the proportion loci polymorphic and expected heterozygosity (H_j) using a large AFLP data set from an experimental study.

Application

AFLP data from replicated experimental populations of *Americamysis bahia* (mysid shrimp) with manipulated levels of genetic diversity were processed as described to produce estimates of initial heterozygosity for populations sampled at experiment termination. These simulations were compared with experimentally derived estimates. Analyses were used to investigate extinction risks from the interaction of stress and reduced genetic diversity (see Markert et al. poster).

- After screening, 59 reproducible AFLP markers were identified to assess *A. bahia* genomic diversity (Figure 1).
- For high diversity lines, initial H_j for each population was estimated using the control population's harmonic-mean effective population size (N_e) and ending H_j estimated using AFLP markers (Table 1).
- The reliability of this estimate was demonstrated by comparing the result to simulated lines, which were created by adding *A. bahia* genotypes from two source populations (Table 2).
- The proportion of genetic diversity retained in each population line was estimated using the equation $H_t/H_o = [1-(1/2N_e)]^t$ (Frankham et al., 2004). The variable H_t represents the population's heterozygosity at the second time interval, whereas H_o represents the population's initial heterozygosity. The variable t represents the number of generations.
- Each control population line was established using the same method as a corresponding experimental population line and we assumed their initial H_j values would be the same.
- The ending H_j values for each experimental population line were estimated by multiplying the initial H_j values by their proportion of genetic diversity retained.

AFLP Processing



Two of many replicate AFLP Genotypes used to optimize scoring parameters

1 R script to convert a GeneMarker® v1.6 (SoftGenetics LLC®, State College, PA) file to a format compatible for AFLPScore v1.3 (Whitlock et al., 2008). Comments are preceded by a # and are written in red.

```
#Export your text file from GeneMarker and delete the remarks at the beginning of the file.
#The first line of your file will now be a wrapped list of variable names, but it does not include
#names for the first two columns of data.
#Add field names for the first two variables (e.g., Num & Sample_Name); be sure to include tabs
#and make sure there is only one tab per space.

#The command below reads your file into R and labels the file as f.
f<-read.delim('Type_Your_File_Name.txt',header=T,sep="\t") #reads your file into R
#The command below retains all but the first and last columns into a new object called f.2.
numcole<-length(f[1,])-2
f.2<-f[,2:(numcole+1)]
#The command below saves the new file f.2 with a new name.
write.table(f.2,'M:/Type_Your_New_File_Name.txt',quote=F,sep="\t",row.names=F)
#The file is now converted to a format to conform to AFLPScore.
f<-read.delim('M:/Type_Your_New_File_Name.txt',header=T,sep="\t")
f.sorted<-f[order(f$Sample_Name),]
recs<-length(f.sorted$Sample_Name)
sam<-rep(NA,recs)
isRep<-sam
for(i in 1:recs){
        sam[i]<-strsplit(as.character(f.sorted$Sample_Name[i]),"_")[[1]][1]
        if(nchar(sam[i])==9) {
                isRep[i]<-1
        } else isRep[i]<-2
        sam[i]<-substr(sam[i],1,8)
        sumsam<-sum(sam==sam[i],na.rm=T)
        if(sumsam>1) isRep[i]<-1
}
nfields<-length(names(f.sorted))
f.2<-data.frame(cbind(sam,f.sorted[,2:nfields],isRep))
for(i in 1:recs){
        for(j in 1:recs){
                if(f.2$isRep[j]==1 & f.2$sam[j]==f.2$sam[i]){
                        f.2$isRep[j]<-"1"
                }
        }
}
f.3<-f.2[order(f.2$isRep,f.2$sam),]
f.4<-f.3[,c(1:nfields)]
#The command below deletes the unnecessary files.
rm(f.2,f.sorted,f.3,sam,recs,nfields,isRep)
#The command below write the new file to a text file and save it with a new name.
write.table(f.4,'M:/Type_Your_New_File_Name.txt',quote=F,sep="\t",row.names=F)
#Check the file format to make sure that repeated samples are placed in pairs at the top of the
#data file in alphabetical order.
```


Acknowledgements

We thank Denise Champlin, Ruth Gobell, and Dr. Anne Kuhn for helping maintain and census the mysid shrimp populations. Dr. Mark Bagley, Annette Ross, and Suzanne Jackson helped produce the AFLP data and Dr. Jeffrey Hollister helped create the second R script. Patricia DeCastro of SRA designed this presentation.

Application Results

Table 1: AFLP based measures of Hj were calculated from genotypes collected at the end of the experiment. Detailed weekly census data were used to estimate likely starting H_j (~H_o). Simulated founding Hj was calculated by using 12 individual genotypes drawn from stock populations that had never been through a bottleneck.

	OBSERVED	EXPECTED	
	End of Experiment H_j (~H_t)	Calculated Starting H_j (~H_o)	Simulated H_j
Mean	0.1869	0.1909	0.2059
S.D.	0.0680	0.0179	0.0063

Table 2. Progressive calculations to estimate initial heterozygosity (Hj_i) for each population line using the control population's harmonic-mean effective population size (N_e) and ending expected heterozygosity (Hj_E) value calculated using AFLP markers. Each control population line was established using the same method as a corresponding experimental population line and we assumed their Hj_i values would be the same.

	Week	4	7	10	13				
Population	Tank	Count	Count	Count	Count	N_e	H_t/H_o	Hj_i	Hj_E
Control	3	75	98	73	59	80.97	0.22		
Experimental		4	14	10	6	2	4.77	0.64	

	Week	4	7	10	13			
Population	Tank	Count	Count	Count	Count	N_e	H_t/H_o	Hj_i
Hj_E								
Control	3	75	98	73	59	80.97	0.22	
Experimental	4	14	10	6	4	7.77	0.64	

Week	4	7	10	13				
Population	Tank	Count	Count	Count	Count	N_e	H_t/H_o	Hj_i Hj_E
Control	3	75	98	73	59	80.97	0.22	
Experimental	4	14	10	6	4	7.77	0.64	0.15

2 R script for converting an AFLPScore v1.3 (Whitlock et al., 2008) file to a format compatible with AFLP-Surv v1.0 (Vekemans, 2002). Comments are preceded by a # and are written in red.

```
#Export your genotype text file from AFLPScore and save the file to your working directory.
#The command below reads your file into R and labels the file as f.

f<-read.delim('Type_Your_File_Name.txt',header=T,sep="\t")

#The command below sorts the samples alphabetically.

f.sorted<-f[order(f$Sample_Name),]

#The command below deletes all replicated samples except the first one.

f.2<-f.sorted[!duplicated(f.sorted$Sample_Name),]

#The command below writes you file to your directory.

write.table(f.2,'M:/Type_Your_New_File_Name.txt',quote=F,sep="\t",row.names=F)

#Add the number of populations to the upper left corner, add a tab, and add the number of loci.
#If you had more than one replicated sample, then delete the extra duplicated samples.
```

Conclusions

- The two R scripts we developed reduce subjectivity and expedited the analyses of our large AFLP data sets. The line code for the two R scripts also could be manually adjusted to convert AFLP data between other commonly used computer programs.
- These approaches increase the utility of AFLPs to address important issues in conservation and environmental protection.

References

Frankham R, Ballou JD, Briscoe DA. (2004). Introduction to conservation genetics. Cambridge: Cambridge University Press. 617 p.

Vekemans X. (2002). AFLP-SURV version 1.0. Distributed by the author. Laboratoire de Génétique et Ecologie Végétale, Université Libre de Bruxelles, Belgium. http://www.ulb.ac.be/sciences/lagev/aflp-surv.html

Whitlock R, Hipperson H, Mannerellie M, Butlin K, Burke T. (2008). An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Molecular Ecology Resources 8:725-735.