Evaluación de los filtros escalonados en plink y su efecto en la estructura poblacional de *tepuihyla*.

Abstract

Step filtering techniques in RADseq can have significant effects on population and genetic structure results. The correct selection and application of filters are crucial to ensure data accurately concerning the genetic variability present in a particular population. An 80% of filtering data prioritizes the quality and biological relevance of the SNPs, a 70% and 50% of filtering criteria seek a balance between the conservation of genomic diversity and the elimination of low-quality markers. On the other hand, 30% of filtering data seeks to maximize genomic diversification by retaining a greater number of SNPs, including less common but potentially biologically important variants. To identify how filtering criteria influence population structure results, I used a principal component analysis (PCA) to synthesize genetic similarities and differences between populations. I used different populations of the Auyan-tepui frog (*Tepuihyla edelcae*) from the Canaima National Park in Venezuela.

Key words: de novo assembly, parameter optimization, population genetics, RAD-seq, STACKS.

1. Introducción

Ante la imperante necesidad de comprender como el cambio climático afecta la estructura poblacional y la diversidad genética de las especies (Willis & Bhagwat, 2009), se destaca la importancia de entender la diversidad genética para evaluar el impacto en la salud, la viabilidad y la conservación de las poblaciones (Bertorelle et al., 2022; Robinson, Kyriazis, Yuan, & Lohmueller, 2023). Según la Lista Roja de la Unión Internacional para la conservación de la Naturaleza (IUCN), aproximadamente un 46,6% de las especies evaluadas, que incluyen animales, plantas y hongos están en riesgo (IUCN Standards and Petitions Committee., 2021); lo que sugiere una urgencia para abordar las preocupaciones relacionadas con la diversidad genética.

Los análisis de datos genómicos constituyen una importante herramienta para obtener información sobre la diversidad genética, la estructura poblacional y la viabilidad de especies amenazadas (Robinson et al., 2022). Estos análisis permiten estimar parámetros genéticos, demográficos fundamentales, los cuales son críticos para modelar y predecir el riesgo de extinción de una población (Brook et al., 2000).

Diversos estudios moleculares han sido desarrollados para focalizar las lecturas cortas generadas por las plataformas de secuenciación modernas en posiciones específicas del genoma (Reuter, Spacek, & Snyder, 2015; Van der Auwera et al., 2013). Estos análisis utilizan técnicas de alto rendimiento basadas en la secuenciación de ADN asociada a sitios de restricción (RADseq) (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Shafer et al., 2017).

En estudios específicos, los datos RADseq se han utilizado para identificar genes y procesos biológicos relevantes para la adaptación local en ratones de campo (*Clethrionomys glareolus*) en Gran Bretaña. Un estudio publicado por Marková y colaboradores en 2023 examinó la adaptación evolutiva de estos roedores a partir de los datos RADseq obtenidos para identificar genes y procesos biológicos importantes para la adaptación local. Este trabajo se alinea con otro estudio que explora cómo los datos genómicos pueden ayudar a predecir la vulnerabilidad del cambio climático en distintas especies (Hoffmann, Weeks, & Sgrò, 2021).

En el contexto de los estudios de genética de poblaciones, Stacks emerge como la pieza clave para el ensamblaje de lecturas, la identificación de alelos y genotipos (McKinney, Larson, Seeb, & Seeb, 2017; Rutledge, Devillard, Boone, Hohenlohe, & White, 2015). Este software ha obtenido un amplio reconocimiento, siendo citado en más de dos mil publicaciones y considerado como un componente esencial y popular en el análisis de datos RADseq (Catchen et al., 2013; Rochette, Rivera-Colón, & Catchen, 2019).

Esta plataforma es respaldada por un riguroso marco estadístico. Desempeña un papel crucial en la realización efectiva de análisis genéticos *de novo*, al establecer un puente entre

la complejidad de los datos de secuenciación y la obtención de resultados precisos y confiables en estudios poblacionales a gran escala (Catchen et al., 2011).

PLINK (Phasing and Linking Integrated Network-Based Clustering Application) es un recurso integral que lleva a cabo el filtrado, la imputación de datos, la asociación genética y la visualización de resultados (Purcell et al., 2007; Yang et al., 2012). Esta herramienta proporciona soluciones eficientes para el análisis de grandes conjuntos de datos de secuenciación de regiones amplificadas (RADseq) y otros datos genéticos (Chang et al., 2015).

Los filtros escalonados en la genómica con RADseq permiten la eliminación de muestras o marcadores no deseados, mejorando la calidad de los datos y simplificando los análisis posteriores (Andrews et al., 2016; Eaton & Ree, 2013). Además, Pedersen y colaboradores en 2021 resaltan el uso de PLINK en la identificación de variantes genéticas y la reducción de ruido genético en los datos de secuenciación.

Este estudio explora cómo los filtros escalonados pueden proporcionar una mayor comprensión de la estructura poblacional de *Tepuihyla*, un fascinante género endémico de la región de Pantepui en la Gran Sabana venezolana (Mijares-Urrutia, Manzanilla-Puppo, & La Marca, 1999). *Tepuihyla* es un género que habita en cimas montañosas aisladas llamadas tepuis (Kok, Ratz, Marco, Aubret, & Means, 2015), cuya región geológica brinda condiciones ecológicas únicas e históricamente han sido vistas como barreras potenciales al flujo de genes (Aubrecht, Barrio-Amorós, Breure, Brewrer-Carías, et al., 2012). Por lo tanto, *Tepuihyla* es un modelo ideal para estudiar la divergencia entre poblaciones altamente aisladas.

Para realizar esta investigación, se necesitó de un considerable poder de cómputo, el cual fue aprovechado mediante el uso de clústers bioinformáticos como EC2 de Amazon, un servicio en línea que permite diseñar una máquina virtual (Kumar, Kumar, Divakar, & Gokul, 2017; Stigler, 2018; Wilkins, 2019), y Hoffman2 de la Universidad de California, Los Ángeles (UCLA), los cuales desempeñan un papel esencial en el estudio. Estas herramientas proporcionaron la capacidad requerida para analizar de manera eficiente y precisa grandes

conjuntos de datos genómicos, permitiendo una exploración exhaustiva de la estructura genética de *Tepuihyla*. Este trabajo de titulación busca esclarecer la estructura poblacional entre cuatro poblaciones muestreadas en distintas cimas, conectando las herramientas de bioinformática, el poder de cómputo y la riqueza biológica de los tepuis venezolanos, ofreciendo una perspectiva sobre la interacción entre genética, biología computacional y geografía en el estudio de la evolución de las poblaciones.

2. Métodos

2.1. *Muestras, preparación de bibliotecas y secuenciación*

Esta sección fue previamente elaborada por Patricia E. Salerno en el año 2013, durante un estudio llevado a cabo en la región geológica del Escudo Guayanés. El muestreo se realizó en cuatro cumbres del Parque Nacional Canaima, incluyendo una en Auyán-tepui y tres en el Macizo de Chimantá (Eruoda-tepui, Churí-tepui y Abakapá-tepui). Se secuenciaron un total de 97 individuos de *Tepuihyla edelcae* de estas cuatro cumbres. Los detalles sobre los códigos y las localidades del muestreo se encuentran en la tabla del Anexo 1.

Se generaron bibliotecas de ADN asociado al sitio de restricción de doble digestión (ddRADseq) bajo el protocolo descrito en Peterson et al., (2012). Se extrajo ADN con el kit de extracción de ADN genómico de sangre y tejidos Viogene, siguiendo las instrucciones asociadas.

La normalización inicial del extracto de ADN se estableció en 8 ng/µl (para un total inicial de 200 ng). Se realizaron digestiones dobles con las enzimas SphI y MspI y seleccionaron un rango de tamaño de fragmento de 430 a 470 pb (sin incluir adaptadores). Para la separación de los fragmentos se utilizó Blue Pippin (agarosa al 2%, Sage Science, Beverly, MA, EE. UU.). Se utilizaron adaptadores prediseñados (P1 flex y P2) y tres cebadores de PCR compatibles con las enzimas de restricción. Las instrucciones del termociclador para la PCR phusion fue el mismo que el descrito por Peterson et al. (2012), con un total de 12 ciclos. Las secuencias se obtuvieron en Illumina Hiseq 2500 como lecturas de extremos emparejados, cuyo objetivo de lecturas fue de aproximadamente 40 millones por grupo de 48 individuos.

2.2. *Resumen del pipeline bioinformático utilizado*

El pipeline presentado resume los pasos seguidos para el desarrollo de esta investigación (Figura 1). El flujo de trabajo descrito incluye el uso de herramientas bioinformáticas diseñadas para facilitar el análisis de datos genéticos. Stacks es una suite de software que permite la identificación y agrupación de secuencias de ADN en loci (Catchen et al., 2011), mientras que PLINK es una herramienta de análisis de datos genéticos que permite la aplicación de filtros escalonados para limpiar el conjunto de datos (Purcell et al., 2007).

Por otro lado, PGDSpider es una herramienta de conversión de datos automatizada para programas de genómica y genética de poblaciones (Lischer & Excoffier, 2012). Por último, R es un entorno de programación estadística y gráfica que permite realizar análisis avanzados (R Core Team, 2017), incluido el Análisis de Componente Principales (PCA) con la biblioteca `adegenet` (Jombart, 2008; Jombart & Ahmed, 2011).

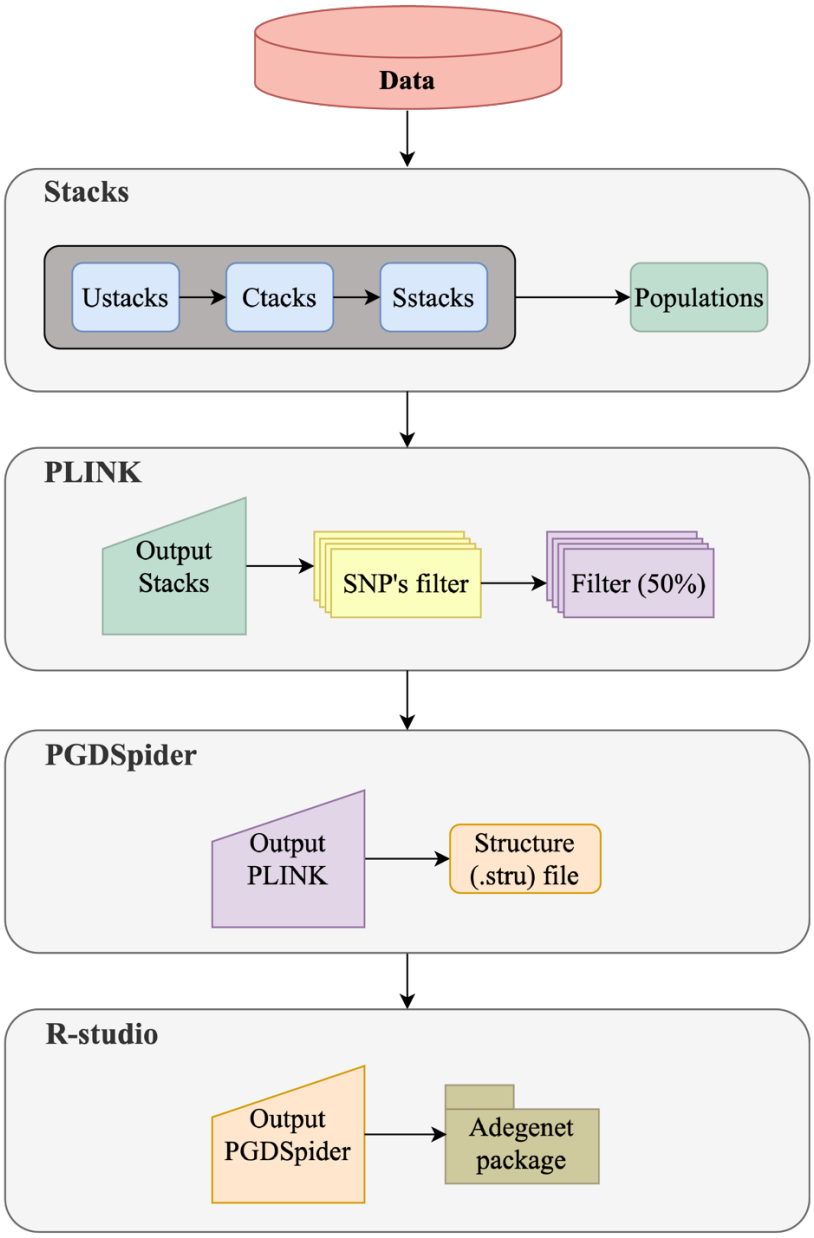


Figura 1. Resumen del pipeline bioinformático.

2.3. *Ingreso al servidor y montaje del bucket S3 en una instancia*

Se creó una instancia en EC2 de Amazon con 16 GB de RAM y 4 unidades de cómputo para comprender los módulos de la tubería de Stacks. El proceso de montaje implicó la configuración del almacenamiento en la nube de Amazon S3 (Simple Storage Service) y la conexión del bucket S3.

Desde el terminal, se usó el siguiente código:

``

ssh -i keystacks2.pem ubuntu@[DNS] #Ingreso al servidor

s3fs oreos3 bucket/ -o passwd_file=$HOME/.passwd-s3fs,allow_other,uid=1001,gid=1001 #Montaje del bucket

` `

Nota: El DNS cambia cada vez que se abre el servidor.

2.4. *Genotipado de novo en Stacks, módulos por separado*

La construcción de esta propuesta de ejecución de módulos por separado en Stacks se desarrolló con un único script bash. El código fue tomado y adaptado de la guía de usuario de Stacks con los primeros tres pasos de la canalización *de novo* (ustacks, cstacks, sstacks). Los análisis de la canalización del genotipado se realizaron usando la versión de Stacks 2.6.0.

En este trabajo, debido a la falta de un genoma de referencia, se utiliza el enfoque *de novo* (Catchen et al., 2011). De acuerdo a la figura 1, el primer paso del proceso del pipeline corresponde a `ustacks`, que construye loci de un solo extremo a partir de secuencias de lectura corta. El segundo módulo es `cstacks`, que permite la creación de un catálogo a partir de cualquier conjunto de muestras procesadas por `ustacks`. Este catálogo genera un conjunto de loci de consenso al fusionar alelos. El ultimo modulo, llamado `sstacks`, relaciona los loci de individuos con el catálogo creado por `cstacks`. En el caso de una población general, todas las muestras se compararían con el catálogo utilizando `sstacks` para identificar los alelos presentes en la población (Catchen et al., 2011, 2013).

El código utilizado para ejecutar Stacks paso por paso fue el siguiente:

``

#!/bin/bash

```
samples=$1 #Carpeta de los datos de entrada demultiplexados
output=$2 #Salida UStacks

id=1 #contador ID
for sample in "$samples"/*; do
filename=$(basename "$sample" ".fq.gz") #Elimina la extensión  y se guarda en la variable asignada

#Paso ustacks
ustacks -f "$sample" -o "$output" -i "$id" --name "$filename" -M 4 -p 4 #El parámetro M puede cambiar de valores

id=$((id+1))
done

#Paso cstacks
popfile=$3 #Archivo de poblaciones
cstacks -n 6 -P "$output" -M "$popfile" -p 4 #El parámetro p corresponde al poder de computo

#Paso sstacks
sstacks -P $src/stacks/ -M $src/popmaps/popfile -p 8
```

Para ejecutar este script se usó la siguiente línea de código:

```
sh run_stacks path_to_samples output_path* popmap
```

Nota: La ruta de salida debe existir previamente.

2.5. *Genotipado de novo en Stacks, versión condensada*

La construcción de esta versión condensada de Stacks se desarrolló en una sola línea. El código fue tomado y adaptado de la guía de usuario de Stacks con los primeros tres pasos de la canalización *de novo*, más el último programa denominado `populations`.

Para acceder al clúster de Hoffman2, ingresé el usuario y contraseña.

``
ssh user@hoffman2.idre.ucla.edu #Inicio de sesión como usuario
``

Dada mi experiencia previa ajustando los valores de los parámetros en el paso mencionado anteriormente, opté por seguir parámetros específicos para ejecutar el pipeline de manera integral. Las variables que consideré importantes para este proceso fueron:

Parámetros	Valores usados
-m: Profundidad mínima de cobertura requerida para crear un stack (predeterminado 3).	3
-M: Distancia máxima (en nucleótidos) permitida entre stacks (predeterminado 2).	4
-n: Número de discrepancias permitidas entre loci de muestra al crear el catálogo (predeterminado 1).	5

El programa ofrecía una serie de opciones para el filtrado de datos, seleccioné las siguientes:

- -R: corresponde al porcentaje mínimo de individuos en todas las poblaciones necesarios para procesar un locus, por lo que el valor ideal fue de 0.1.
- --write-random-snp: restringe el output de datos a un SNP aleatorio por locus.

El paso final corresponde al programa `populations`, que se encarga de analizar una población de muestras individuales, calcula estadísticas genéticas de poblaciones y exporta

en diversos formatos estándar de salida (Catchen et al., 2011). Los parámetros de descarga para este trabajo fueron: `--structure`, `--plink` y `--vcf`.

El código condensado fue el siguiente:

```
denovo_map.pl -T 16 -m 3 -M 4 -n 5 -o ./stacks/ --samples Tepuihyla_data --popmap tepuihyla_popmap97.txt  -X "populations:--vcf --plink --structure --write-random-snp -R 0.1"
```

2.6. *Filtrado escalonado de matriz de SNPs*

El proceso de filtrado de la matriz de SNPs se realizó utilizando PLINK versión 1.07 y se ejecutó desde el clúster de Hoffman2. Se empleó un procedimiento basado en la guía de usuario, adaptado para encontrar la lista blanca de loci e individuos que se conservaron luego de aplicar los filtros en dos etapas principales.

Para ello, desde el terminal se navegó a la carpeta que contiene los resultados de Stacks y cargamos el módulo de PLINK de la siguiete manera:

```
ssh user@hoffman2.idre.ucla.edu  #Inicio de sesión como usuario
module load plink  #Carga del módulo
plink  #Ejecución del bin
```

El primer paso implica establece el umbral de exclusión de variantes genéticas utilizando la opción `--geno`. Este proceso involucra la aplicación de cuatro filtros, cada uno desarrollado con los siguientes porcentajes:

No. de filtrado	Porcentaje %
1ro	0.8
2do	0.7
3ro	0.5
4to	0.3

Para eliminar loci con datos faltantes en exceso, se utilizó el siguiente comando:

``

plink --file populations.plink --geno 0.2 --recode --out new_plink-b --allow-extra-chr
``

Nota: El valor del parámetro `geno` varía de acuerdo con el porcentaje deseado.

El segundo paso se llevó a cabo utilizando la opción `--mind 0.5` en base a los outputs del paso anterior. Esta opción permite excluir individuos que contienen más del 50% de los datos (SNPs) faltantes. De esta manera, eliminamos individuos que podrían sesgar la interpretación de las inferencias poblacionales las cuales son altamente sensibles a los datos faltantes.

El siguiente script genera archivos limpios y optimizados para análisis estadísticos y asociaciones genéticas. El código empleado fue:

``

plink --file new_plink-b --mind 0.5 --recode --out new_plink-c --allow-extra-chr
``

Nota: El input del análisis cambia en cada ejecución.

2.7. *Identificación de la frecuencia alélica (PCA)*

Para llevar a cabo el análisis de componentes principales (en español ACP, en inglés, PCA) se requirió un paso adicional. Los archivos de salida de PLINK, que se encontraban en formato `.ped` y `.map` se transformaron en formato `structure (.stru)` mediante el software PGDSpider version 2.1.1.5 (May 2018). La figura 3 muestra la vista desde el software durante el proceso de transformación de formatos.

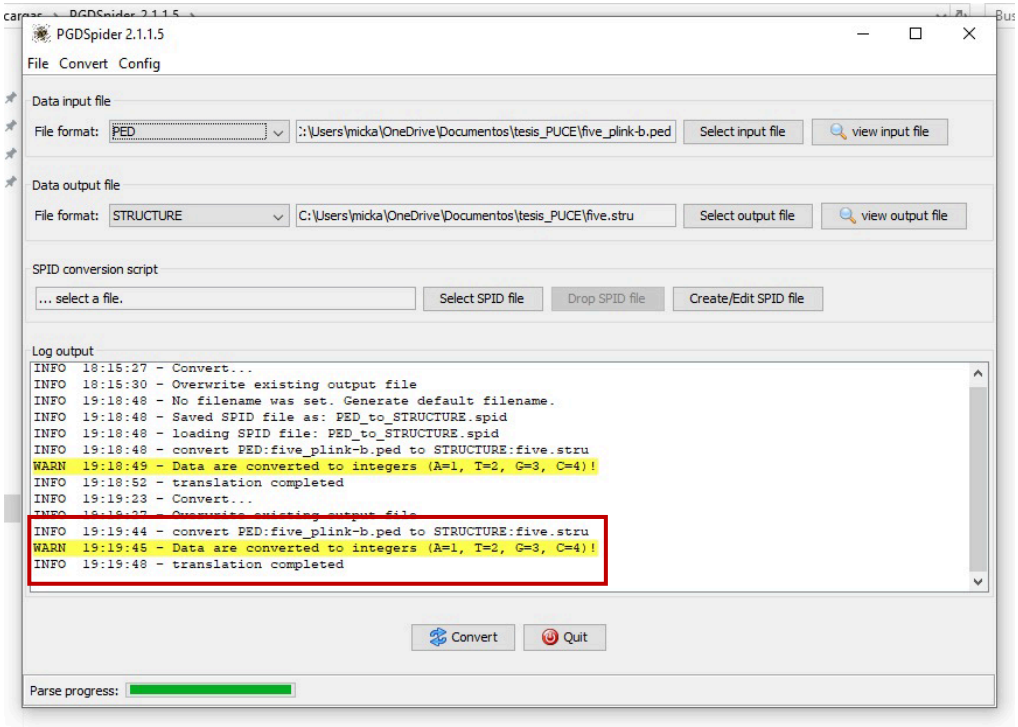


Figura 2. Vista desde el software PGDSpider. Proceso finalizado.

Una vez finalizada la transformación de los archivos, se procedió a importarlos a R-studio utilizando la versión 4.3.2 de R. Para llevar a cabo el análisis de PC (componentes principales), se empleó la biblioteca `adegenet` en su versión 2.1.10.

La biblioteca `adegenet` forma parte de un paquete homónimo diseñado específicamente para el análisis multivariado de marcadores genéticos. Esta herramienta es esencial en el ámbito genético, ya que posibilita la realización de PCA y otras técnicas estadístcas para explorar y visualizar datos genéticos de manera efectiva (Jombart, 2008; Jombart & Ahmed, 2011).

Se usó el comando `find.clusters` (también en adegenet) para identificar grupos de individuos y compararlos con los grupos a priori basados en la membresía de la población. Este comando transforma los datos en PC y luego utiliza un algoritmo k-means con un número creciente de clústers (k).

Se seleccionó el valor más bajo del Criterio de Información Bayesiano (BIC) para elegir el k más adecuado. El primer análisis en `adegenet` se realizó con una matriz de loci original que constaba de 97 individuos y 169216 SNPs, el segundo análisis con 80 individuos 507 SNPs, el tercer análisis con 61 individuos y 20105 SNPs, y el cuarto con 52 individuos y 42852 SNPs.

El código empleado para instalar, cargar la bibloteca `adegenet`, y ejecutar el PCA fue el siguiente:

``

```
#install.packages("seqinr")
#install.packages("ade4", dependencies = TRUE)
#install.packages("hierfstat")
#install.packages("car")
#install.packages("factoextra")
#install.packages("adegenet")
#install.packages("ggplot2")

#Cargar bibliotecas
library("adegenet")
library("ape")
library("pegas")
library("seqinr")
library("ggplot2")
library("hierfstat")
library("factoextra") #to get eigen values
```

```
#Establecer directorio de trabajo
setwd("/Users/mickaelagallo/Documents/TesisPUCE/R_stru")

#Abrir archivo structure
#Primer filtro
myFile <- read.structure("popu.stru") #97 inds and 107630 SNPs
myFile <- read.structure("new.stru") #80 inds and 507 SNPs
myFile <- read.structure("three.stru") #77 inds and 3680 SNPs, marker 0
myFile <- read.structure("five.stru") #61 inds and 20105 SNPs, marker 0
myFile <- read.structure("seven.stru") #52 inds and 42852 SNPs

#Segundo filtro
myFile <- read.structure("new1.stru") #80 inds and 507 SNPs
myFile <- read.structure("three1.stru") #77 inds and 3680 SNPs, marker 0
myFile <- read.structure("five1.stru") #61 inds and 20105 SNPs, marker 0
myFile <- read.structure("seven1.stru") #52 inds and 42852 SNPs

X<-scaleGen(myFile, NA.method="mean")

#PCA
pca1<-dudi.pca(X,cent=FALSE,scale=FALSE,scannf=FALSE,nf=3)
myCol<-c(colors()[503],colors()[94],colors()[257],colors()[566])
pop_jimmy<-pop(myFile)

#Graficar PCA
s.class(pca1$li,pop(myFile),col=myCol)
add.scatter.eig(pca1$eig[1:20],3,2,2, posi = "bottomright") #para agregar los eigenvalues  |
la posi = "topleft"
eig.val <- get_eigenvalue(pca1) #para obtener los porcentajes
eig.val
```

```
#DAPC
dapc2<-dapc(X,pop(myFile))
scatter(dapc2,col=c(colors()[503],colors()[94],colors()[257],colors()[566]))

#COMPOPLOT
compoplot(dapc2,col=myCol)

#BIC
foo.BIC <- find.clusters(myFile, n.pca=200, choose=FALSE)
plot(foo.BIC$Kstat, type="o", xlab="number of clusters (K)", ylab="BIC",col="blue",
main="Detection based on BIC")
points(2, foo.BIC$Kstat[2], pch="x", cex=2)
mtext(3, tex="'X' indicates the actual number of clusters")
```

Nota: El data set (.stru) cambia en cada análisis.

3. Resultados

3.1. *Resultados de los filtros*

El filtrado en PLINK muestran el impacto de cada filtro en la cantidad de SNPs eliminados y retenidos. El análisis genómico de *T. edelcae* revela una disminución en el porcentaje de SNPs presentes en el material genético al aumentar el umbral de genotipos faltantes. Por ejemplo, al aplicar un filtro del 80% del genoma, se excluyeron 107123 SNPs, reteniendo únicamente 507. Así mismo, una selección del 30% provocó una retención de tan solo 42852 SNPs y una omisión de 64778, mostrando que puede haber un cambio significativo en el genoma al aplicar un % de filtro determinado, ver Tabla 1.

Tabla 1. Matriz de loci original con 97 individuos y 107,630 loci presentes.

No. de filtrado	Porcentaje %	SNPs eliminados	SNPs retenidos	Filtro individuos (50%)		
				Tasa de genotipado	Individuos eliminados	Individuos retenidos

Sin filtro	0	0	107630	0.30	0	97
1ro	0.8	107,123	507	0.97	17	80
2do	0.7	103,950	3,680	0.92	20	77
3ro	0.5	87,525	20,105	0.88	36	61
4to	0.3	64,778	42,852	0.79	45	52

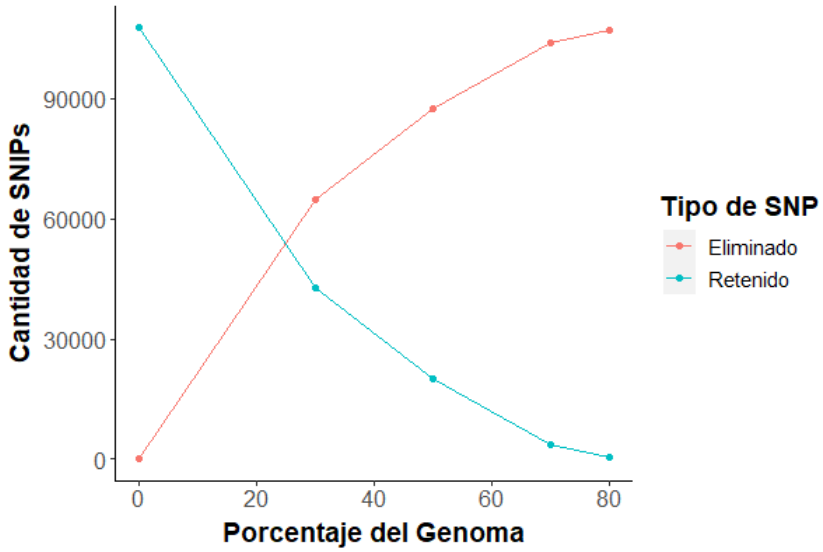


Figura 3. Prueba de filtros en PLINK.

Esta relación entre los SNPs retenidos y eliminados puede visualizarse en la Figura 3. Donde, la cantidad de SNPs omitidos es mayor al 50% a partir del 25%. Al tener una selección del 30%, los especímenes representados por el genoma elegido corresponden a 52 individuos, como se nota en la tabla 1. Este hallazgo resulta novedoso porque da un indicio de que existe una diferenciación genómica en al menos el 40% de los ejemplares muestreados en las cuatro cumbres del Parque Nacional Canaima.

3.2. *Estimación de la estructura poblacional*

La diferenciación genómica de *T. edelcae* anteriormente mencionada fue corroborada al estimar su estructura poblacional. El análisis de componentes principales y el compoplot

evidenciaron que esta estructura genética de la especie se encuentra diversificada en al menos dos clústers. Por ejemplo, cuando se hace el PCA para el primer filtro de los SNPs, tres grupos fueron formados, ver Figura 6. En cambio, cuando se hace el PCA en la eliminación de solamente el 30% de los SNPs, dos elipses fueron obtenidas, ver Figura 9. Para los demás casos, la tendencia de 2 a 3 agrupaciones fue repetida, ver Figura 7 y Figura 8.

El análisis compoplot de la Figura 5 fue una segunda confirmación de esta diversificación hallada en *T. edelcae*. Principalmente porque sobresalen patrones de las poblaciones de Auyán-tepui, Abakapá-tepui y una mezcla entre Churí y Eruoda tepui.

En resumen, la estructura poblacional de esta especie podría estar dada a partir de su genoma, con un filtro de la eliminación del ruido de al menos un 50% de SNPs. En consecuencia, se cabría esperar que *T. edelcae* posea como mínimo 20000 SNPs como especie. Las cumbres de Auyán-tepui y las tres en el Macizo de Chimantá (Eruoda-tepui, Churí-tepui y Abakapá-tepui) juegan un rol esencial para la diversificación de *Tepuihyla*.

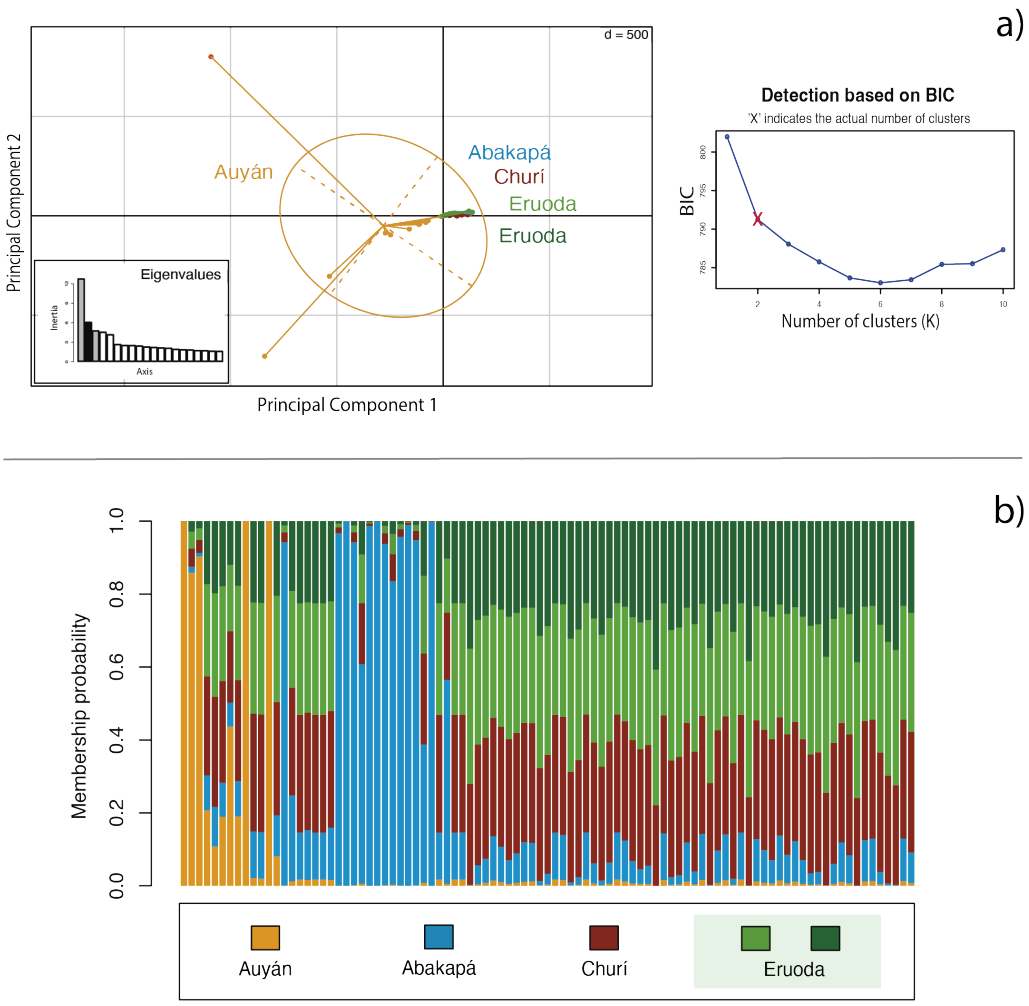
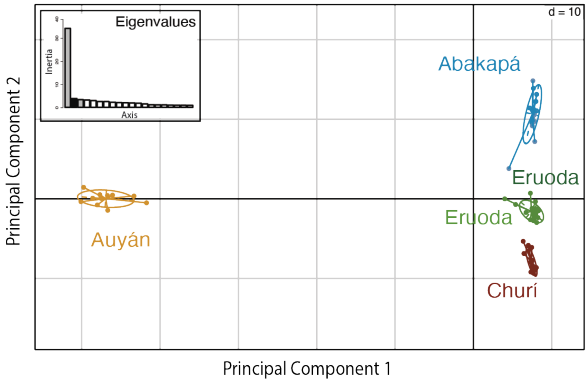
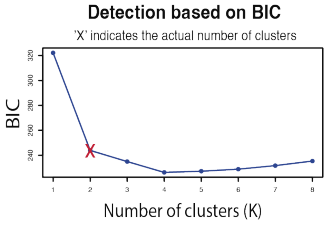


Figura 4. Estructura de la población de *Tepuihyla* (97 individuos y 107630 SNPs). (a) PCA (lado izquierdo) y BIC (lado derecho) de la matriz principal sin filtro. (b) En el compoplot cada barra representa un individuo. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.

Segundo filtro (50%)
SNPs: 507
Ind: 80

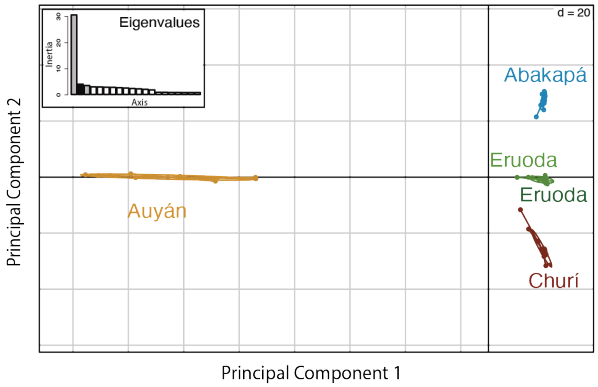
a)

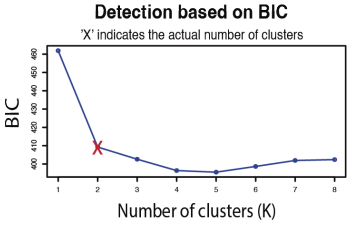




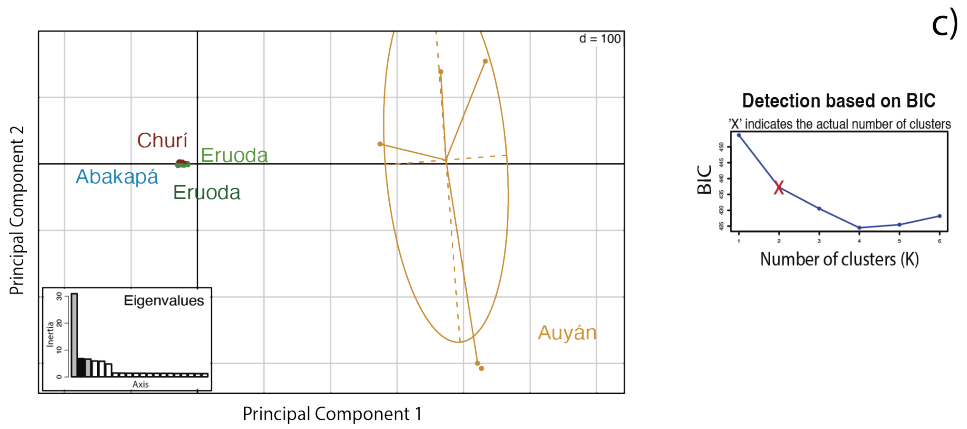
Segundo filtro (50%)
SNPs: 3680
Ind: 77

b)





Segundo filtro (50%)
SNPs: 20105
Ind: 61



Segundo filtro (50%)
SNPs: 42852
Ind: 52

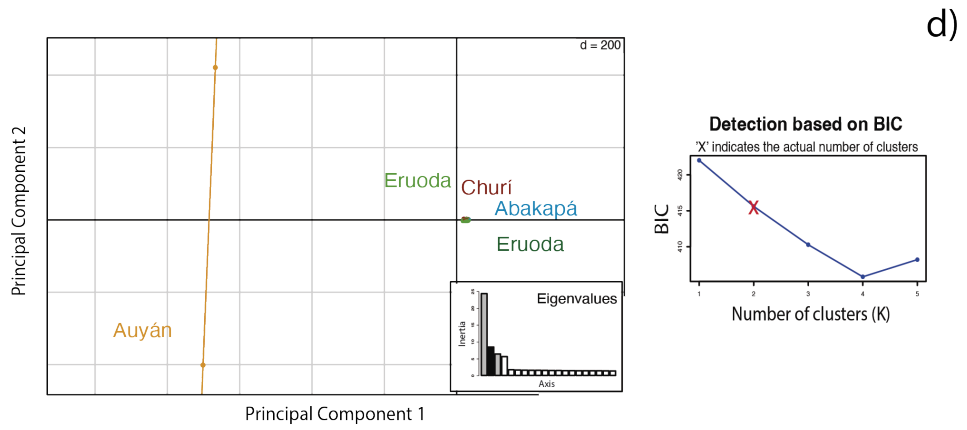


Figura 6. Estructura de la población de *Tepuihyla.* Segundo filtro de exclusión de individuos con más del 50% de los datos (SNPs) faltantes a partir de PCA (lado izquierdo) y BIC (lado derecho). (a) Filtro de 0.2, (b) Filtro de 0.3, (c) Filtro de 0.5 y (d) Filtro de 0.7. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.

4. Discusión

4.1. *Filtros*

La aplicación de filtros escalonados en técnicas como RADseq puede tener efectos significativos en la inferencia de la estructura poblacional y genética (Peterson et al., 2012). Al seleccionar fragmentos cortos de ADN basados en criterios como el tamaño, se puede afectar la representatividad de la diversidad genética de la población estudiada; lo que influye en la recuperación de regiones genómicas específicas y en la saturación de regiones compartidas entre individuos (Hohenlohe, Amish, Catchen, Allendorf, & Luikart, 2011; Pearman, Urban, & Alexander, 2022).

La correcta selección y aplicación de filtros es crucial para garantizar que los datos reflejen con precisión la variabilidad genética presente en la población estudiada (Peterson et al., 2012). En este estudio se optó por un filtrado escalonado específico (80%, 70%, 50% y 30%) para la identificación de SNP, donde cada nivel de filtrado tiene argumentos particulares en relación con la diversificación del genoma y la especificidad de los datos, ver Tabla 1.

El filtro al 80%, aunque riguroso, presenta un impacto positivo en la diversificación del genoma al priorizar la calidad y relevancia biológica de los SNPs retenidos. Al eliminar marcadores propensos a errores o de baja variabilidad genética, se obtiene un conjunto más selectivo de SNPs, mejorando la confiabilidad de los resultados y reduciendo la probabilidad de interpretaciones sesgadas. Este enfoque también contribuye a la especificidad al eliminar SNPs que podrían introducir ruido en los datos (O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018; Pearman et al., 2022).

Por otro lado, los umbrales de filtrado al 70% y 50% representan un equilibrio entre la conservación de la diversidad genómica y la eliminación de marcadores de baja calidad. Estos niveles permiten la retención de un número considerable de SNP, proporcionando una visión más completa de la variabilidad genética. Sin embargo, pueden afectar a la especificidad de los datos, ya que la inclusión de marcadores menos informativos influye en la precisión de las interpretaciones genómicas (Peterson et al., 2012).

Finalmente, el filtro del 30%, al retener un mayor número de SNPs, busca maximizar la diversificación del genoma. Con este enfoque se puede capturar una variabilidad genética extensa al conservar una proporción sustancial de variantes, incluso aquellas con frecuencias alélicas más bajas o menos comunes, pero potencialmente importantes desde el punto de vista biológico (Stuart, Edwards, Sherwin, & Rollins, 2023).

4.2. *Estructura poblacional*

La selección de cuatro cumbres distintas en el Parque Nacional Canaima para el muestreo de *Tepuihyla edelcae*, incluyendo Auyán-tepui, Eruoda-tepui, Churí-tepui y Abakapá-tepui (Lasso & Señaris, 2018; Señaris & Rojas-Runjaic, 2020), ha proporcionado una base sólida para investigar la estructura poblacional de esta especie de rana arborícola. La diversificación genómica se contextualiza en un entorno geográfico y topológico único, con cada tepui presentando condiciones climáticas, topográficas y geográficas particulares.

Los tepuis forman un ecosistema discontinuo de islas conocidas como Pantepui (Salerno, Señaris, Rojas-Runjaic, & Cannatella, 2015). Este singular ecosistema de extrema topografía alberga una alta tasa de endemicidad de varios taxones, especialmente ranas (McDiarmid & Donnelly, 2005).

Eruoda-tepui, con una altitud de 2698 metros, presenta un entorno climático distinto caracterizado por condiciones térmicas entre 17 - 19 °C y de alrededor de 31.73 mm/día. Su altitud elevada influye en la composición genómica de las poblaciones de *Tepuihyla edelcae* que lo habitan (Aubrecht, Barrio-Amorós, Breure, Brewer-Carías, et al., 2012; Mijares-Urrutia et al., 1999).

Auyán-tepui, con una altitud de 2535 metros y Churí-tepui con una altitud de 2500 metros experimentan temperaturas entre 19 - 20 °C y alrededor de 22.52 mm/día de precipitación. Estas condiciones climáticas desempeñan un papel significativo en la adaptación genómica de *Tepuihyla edelcae* en este tepui en particular (Kok et al., 2015).

Abakapá-tepui, con una altitud de 2400 metros, también ofrecen ambientes únicos. Alcanzando hasta 18.45 mm/día de precipitación y con temperaturas que oscilan entre 21 – 22 °C (Aubrecht, Barrio-Amorós, Breure, Brewrer-Carías, et al., 2012).

La inclusión de individuos de diferente tepuis no solo permite explorar la posible variabilidad genética entre poblaciones, sino que también destaca cómo las condiciones climáticas específicas de cada cumbre han esculpido la estructura poblacional de *Tepuihyla edelcae* a lo largo del tiempo (McDiarmid & Donnelly, 2005; Señaris & Rojas-Runjaic, 2020). Estos datos climáticos específicos proporcionan una perspectiva más precisa sobre la relación entre la topografía, las condiciones climáticas y la diversificación genómica en cada tepui.

La identificación de grupos genéticos representa un objetivo fundamental en el campo de la genética de poblaciones (Miller, Cullingham, & Peery, 2020). Básicamente, esto implica sintetizar las similitudes y diferencias genéticas entre poblaciones de la manera más concisa posible. Estos métodos de agrupamiento se respaldan en algoritmos bayesianos, los cuales demandan múltiples suposiciones previas, involucran modelos evolutivos complejos y demandan extenso tiempo de cálculo (Jombart, Devillard, & Balloux, 2010; Kalinowski, 2011).

Durante años, las técnicas multivariadas en ecología se han utilizado para investigar cómo las variables ambientales abióticas influyen en la composición biótica de los ecosistemas (Legendre & Legendre, 2012). Recientemente, como una alternativa al análisis bayesiano, se ha propuesto el uso del análisis de componentes principales (PCA) y algoritmos de agrupamiento (Lee, Abdool, & Huang, 2009; Liu & Zhao, 2006; Patterson, Price, & Reich, 2006; Price et al., 2006). El PCA destaca por su capacidad para identificar estructuras genéticas en grandes conjuntos de datos de manera eficiente y sin suposiciones sobre el modelo genético poblacional (Jombart et al., 2010). No obstante, es importante tener en cuenta que el PCA es sensible a la presencia de datos faltantes (Yi & Latch, 2022).

Las figuras del PCA mostraron una clara diferenciación genómica de *Tepuihyla edelcae*, especialmente al aplicar diferentes filtros de SNPs. Estas agrupaciones reflejaron la diversidad genética y la estructura poblacional de la especie. Al variar los porcentajes de

filtrado de SNPs en el PCA, se identificaron clústers genéticos que representan subpoblaciones o grupos genéticamente distintos dentro de *Tepuihyla*, evidenciando así su diversificación genómica.

En algunas instancias del análisis de PCA se observó la presencia de admixture, indicando una posible mezcla genética entre poblaciones o la existencia de individuos con ancestros genéticos diversos. Este fenómeno puede sugerir procesos como hibridación o migración genética en la historia evolutiva de la especie (Kok et al., 2015)..

Un estudio realizado por Yi & Latch en 2022 señala que, a pesar de la falta de datos, utilizar un filtro menos estricto puede ser más efectivo para identificar la estructura poblacional en el PCA. Sin embargo, al aplicar filtros más rigurosos y eliminar un mayor porcentaje de SNPs, se evidenció una mayor estructura poblacional en *Tepuihyla edelcae*, lo que nuevamente resalta la diversificación genómica de esta especie.

5. Conclusiones

La utilización de filtros escalonados en el estudio de la estructura genética del género *Tepuihyla* es fundamental para la identificación de SNPs relevantes y la mejora de la confiabilidad de los resultados. El filtro al 80% prioriza la calidad y relevancia biológica de los SNPs retenidos, lo que reduce la probabilidad de interpretaciones sesgadas y contribuye a la especificidad de los datos. Los filtros al 70% y 50% permiten la retención de un número considerable de SNPs, proporcionando una visión más completa de la variabilidad genética, aunque pueden afectar a la especificidad de los datos. El filtro al 30% busca maximizar la diversificación genómica, capturando una variabilidad genética extensa y conservando una proporción sustancial de variantes, incluidas aquellas con frecuencias alélicas más bajas o menos comunes, pero potencialmente importantes desde el punto de vista biológico.

La selección de cuatro cumbres distintas en el Parque Nacional Canaima para el muestreo de *Tepuihyla edelcae* ha proporcionado una base sólida para investigar la estructura poblacional de esta especie de rana arborícola. La diversificación genómica se contextualiza en un entorno geográfico y topológico único, con cada tepui presentando condiciones climáticas, topográficas y geográficas particulares. Estos datos climáticos específicos

proporcionan una perspectiva más precisa sobre la relación entre las condiciones climáticas y la diversificación genómica en cada tepui.

El análisis de componentes principales (PCA) y algoritmos de agrupamiento se han utilizado para describir estructuras genéticas en conjuntos de datos extensos, prescindiendo de suposiciones sobre el modelo genético poblacional subyacente. Sin embargo, estos análisis son altamente sensibles a datos faltantes y pueden ser menos precisos que los análisis bayesianos, lo que destaca la importancia de seleccionar cuidadosamente los protocolos de secuenciación respecto al filtrado de datos para estudios de conservación.

6. Referencias

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81-92. https://doi.org/10.1038/NRG.2015.28

Aubrecht, Barrio-Amorós, Breure, Brewrer-Carías, Derka, Fuentes-Ramos, … Vlček. (2012). *Venezuelan Tepuis: their caves and biota*. Bratislava, Eslovaquia: Acta Geologica Slovaca, AGEOS.

Aubrecht, R., Barrio-Amorós, C. L., Breure, A. S. H., Brewer-Carías, C., Derka, T., Fuentes-Ramos, O. A., … Vlček, L. (2012). *Venezuelan tepuis: Their caves and biota. Acta Geologica Slovaca - Monograph*. Bratislava, Eslovaquia: Acta Geologica Slovaca, AGEOS.

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., … van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics 2022 23:8*, *23*(8), 492-503. https://doi.org/10.1038/s41576-022-00448-x

Brook, B. W., O'Grady, J. J., Chapman, A. P., Burgman, M. A., Resit Akçakaya, H., & Frankham, R. (2000). Predictive accuracy of population viability analysis in conservation biology. *Nature*, *404*(6776), 385-387. https://doi.org/10.1038/35006050

Catchen, J., Amores, A., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3 Genes|Genomes|Genetics*, *1*(3), 171-182. https://doi.org/10.1534/G3.111.000240

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124-3140. https://doi.org/10.1111/mec.12354

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/S13742-015-0047-8/2707533

Eaton, D. A. R., & Ree, R. H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology*, *62*(5), 689-706. https://doi.org/10.1093/SYSBIO/SYT032

Hoffmann, A. A., Weeks, A. R., & Sgrò, C. M. (2021). Opportunities and challenges in assessing climate change vulnerability through genomics. *Cell*, *184*(6), 1420-1425. https://doi.org/10.1016/J.CELL.2021.02.006

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular ecology resources*, *11 Suppl 1*(SUPPL. 1), 117-122. https://doi.org/10.1111/J.1755-0998.2010.02967.X

IUCN Standards and Petitions Committee. (2021). The IUCN red list of threatened species. Recuperado 20 de febrero de 2024, de IUCN website: https://www.iucn.org/resources/conservation- tools/iucn- red-listthreatened- species

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403-1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070-3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*. https://doi.org/10.1186/1471-2156-11-94

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity*, *106*(4), 625-632. https://doi.org/10.1038/HDY.2010.95

Kok, P. J. R., Ratz, S., Marco, T., Aubret, F., & Means, D. B. (2015). Out of taxonomic limbo: a name for the species of Tepuihyla (Anura: Hylidae) from the Chimantá Massif, Pantepui region, northern South America. *Salamandra*, *51*(4), 283-314. Recuperado de https://ut3-toulouseinp.hal.science/hal-02966053

Kumar, V. D. A., Kumar, V. D. A., Divakar, H., & Gokul, R. (2017). Cloud enabled media streaming using Amazon Web Services. *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017 - Proceedings*, 195-198. https://doi.org/10.1109/ICSTM.2017.8089150

Lasso, C. A., & Señaris, J. C. (2018). *Vi . Fauna Silvestre Del Escudo Guayanés*.

Lee, C., Abdool, A., & Huang, C. H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC bioinformatics*, *10 Suppl 1*(Suppl 1). https://doi.org/10.1186/1471-2105-10-S1-S73

Legendre, P., & Legendre, L. (2012). Numerical Ecology Ch 6 - Multidimensional qualitative data. *Developments in Environmental Modelling*, *24*, 337-424. Recuperado de http://www.sciencedirect.com/science/article/pii/B9780444538680500083

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics (Oxford, England)*, *28*(2), 298-299. https://doi.org/10.1093/BIOINFORMATICS/BTR642

Liu, N., & Zhao, H. (2006). A non-parametric approach to population structure inference using multilocus genotypes. *Human genomics*, *2*(6), 353-364. https://doi.org/10.1186/1479-7364-2-6-353

Marková, S., Lanier, H. C., Escalante, M. A., da Cruz, M. O. R., Horníková, M., Konczal, M., … Kotlík, P. (2023). Local adaptation and future climate vulnerability in a wild

rodent. *Nature Communications 2023 14:1*, *14*(1), 1-11. https://doi.org/10.1038/s41467-023-43383-z

McDiarmid, R. W., & Donnelly, M. A. (2005). The Herpetofauna of the Guayana Highlands: Amphibians and Reptiles of the Lost World. *Ecology and evolution in the tropics: a herpetological perspective*, 1-11.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, *17*(3), 356-361. https://doi.org/10.1111/1755-0998.12649

Mijares-Urrutia, A., Manzanilla-Puppo, J., & La Marca, E. (1999). Una nueva especie de Tepuihyla (Anura: Hylidae) del noroeste de Venezuela, con comentarios sobre su biogeografía. *Revista de Biología Tropical*, *47*(4), 1099-1110. Recuperado de http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S0034-77441999000400046&lng=en&nrm=iso&tlng=es

Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity*, *125*(5), 269. https://doi.org/10.1038/S41437-020-0348-2

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular ecology*, *27*(16), 3193-3206. https://doi.org/10.1111/MEC.14792

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*(12), 2074-2093. https://doi.org/10.1371/JOURNAL.PGEN.0020190

Pearman, W. S., Urban, L., & Alexander, A. (2022). Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular Ecology Resources*, *22*(7), 2599. https://doi.org/10.1111/1755-0998.13646

Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., … Quinlan, A. R. (2021). Effective variant filtering and expected candidate variant yield in studies of rare human disease. *npj Genomic Medicine 2021 6:1*, *6*(1), 1-8. https://doi.org/10.1038/s41525-021-00227-3

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/JOURNAL.PONE.0037135

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909. https://doi.org/10.1038/NG1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, *81*(3), 559. https://doi.org/10.1086/519795

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, *58*(4), 586-597. https://doi.org/10.1016/J.MOLCEL.2015.05.004

Robinson, J., Kyriazis, C. C., Nigenda-Morales, S. F., Beichman, A. C., Rojas-Bracho, L., Robertson, K. M., … Morin, P. A. (2022). The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science (New York, N.Y.)*, *376*(6593), 635-639. https://doi.org/10.1126/SCIENCE.ABM1742

Robinson, J., Kyriazis, C. C., Yuan, S. C., & Lohmueller, K. E. (2023). Deleterious Variation in Natural Populations and Implications for Conservation Genetics. *Annual review of animal biosciences*, *11*(1), 93-114. https://doi.org/10.1146/ANNUREV-ANIMAL-080522-093311

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737-4754. https://doi.org/10.1111/mec.15253

Rutledge, L. Y., Devillard, S., Boone, J. Q., Hohenlohe, P. A., & White, B. N. (2015). RAD sequencing and genomic simulations resolve hybrid origins within North American Canis. *Biology letters*, *11*(7). https://doi.org/10.1098/RSBL.2015.0303

Salerno, P. E., Señaris, J. C., Rojas-Runjaic, F. J. M., & Cannatella, D. C. (2015). Recent evolutionary history of Lost World endemics: Population genetics, species delimitation, and phylogeography of sky-island treefrogs. *Molecular Phylogenetics and Evolution*, *82*(PA), 314-323. https://doi.org/10.1016/j.ympev.2014.10.020

Señaris, C., & Rojas-Runjaic, F. J. M. (2020). Amphibians and Reptiles of Venezuelan Guayana: Diversity, Biogeography and Conservation. En *Neotropical Diversification: Patterns and Processes* (pp. 571-633). Springer, Cham. https://doi.org/10.1007/978-3-030-31167-4_22

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907-917. https://doi.org/10.1111/2041-210X.12700

Stigler, M. (2018). Amazon Web Services. *Beginning Serverless Computing*, 41-81. https://doi.org/10.1007/978-1-4842-3084-8_3

Stuart, K. C., Edwards, R. J., Sherwin, W. B., & Rollins, L. A. (2023). Contrasting Patterns of Single Nucleotide Polymorphisms and Structural Variation Across Multiple Invasions. *Molecular Biology and Evolution*, *40*(3). https://doi.org/10.1093/MOLBEV/MSAD046

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, *11*(1110), 11.10.1. https://doi.org/10.1002/0471250953.BI1110S43

Wilkins, M. (2019). *Learning Amazon Web Services (AWS)*. Addison-Wesley Professional. Recuperado de https://books.google.com/books?hl=es&lr=&id=HvifDwAAQBAJ&oi=fnd&pg=PT24&dq=Amazon+Web+Services+in+Action&ots=-2cjTXmhK6&sig=oJZf_TODZV0NEFSR2kZhiklPFow

Willis, K. J., & Bhagwat, S. A. (2009). Biodiversity and Climate Change. *Science*, *326*(5954), 806-807. https://doi.org/10.1126/SCIENCE.1178838

Yang, B., Wu, Y. J., Zhu, M., Fan, S. B., Lin, J., Zhang, K., … Dong, M. Q. (2012). Identification of cross-linked peptides from complex samples. *Nature Methods 2012 9:9*, *9*(9), 904-906. https://doi.org/10.1038/nmeth.2099

Yi, X., & Latch, E. K. (2022). Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Molecular Ecology Resources*, *22*(2), 602-611. https://doi.org/10.1111/1755-0998.13498

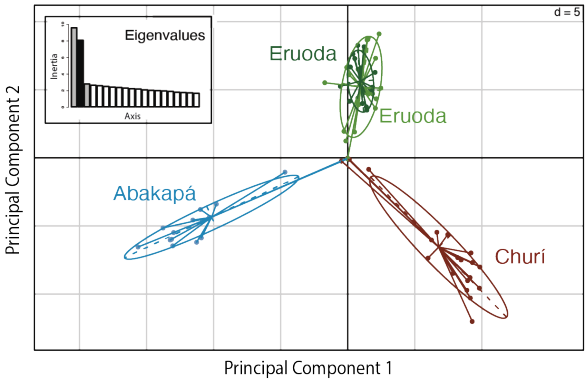
Anexos (Material suplementario)

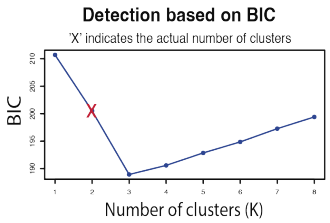
Anexo 1. Individuos y localidades. TNHC-FS = Texas Natural History Collection Field Series, Universidad de Texas; PS = Patricia Salerno (muestras depositadas en el MHNLS, Museo de Historia Natural La Salle).

Localidad general	N° individuos	ID de campo	Coordenadas	Localidad específica
Auyán	13	Au_05824-05836	5°46.599'N 62°32.251'W	Campamento el Oso, Auyán-tepui, Parque Nacional Canaima, Estado Bolívar, Venezuela
Abakapá	25	Ab_365-371, Ab_387-404	5°11.497'N 62°18.939'W	Abakapá-tepui, Macizo de Chimantá, Parque Nacional Canaima, Estado Bolívar, Venezuela
Churí	20	Ch_330-339, Ch_345-346, Ch_357-364	5°15.257'N 62°00.472'W	Churí-tepui, Macizo de Chimantá, Parque Nacional Canaima, Estado Bolívar, Venezuela
Eruoda	39	Er_410-411, Er_446-453, Er_455-466, Er_R-01-17	5°22.525'N 62°05.674'W	Eruoda-tepui, Macizo de Chimantá, Parque Nacional Canaima, Estado Bolívar, Venezuela

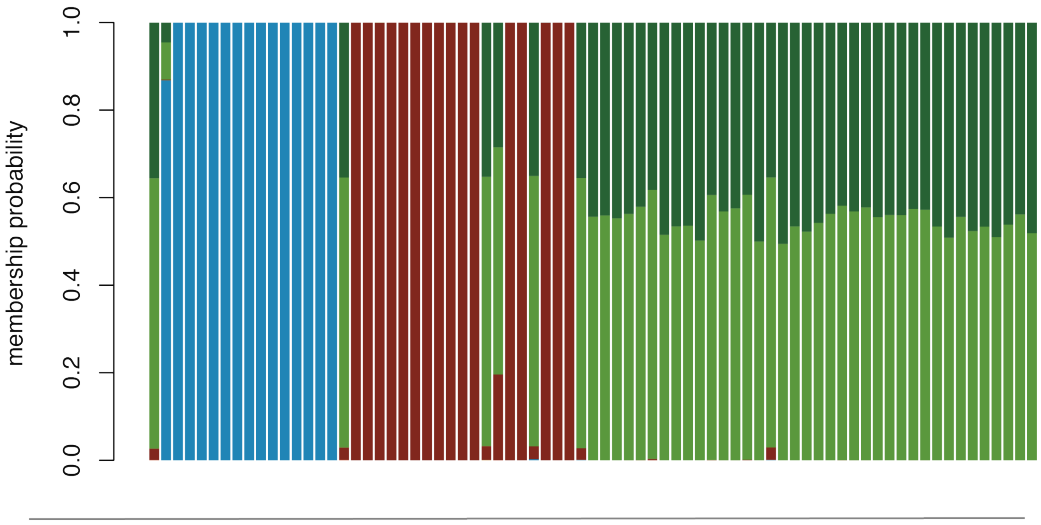
Primer filtrado (0.2)
SNPs: 507
Ind: 80

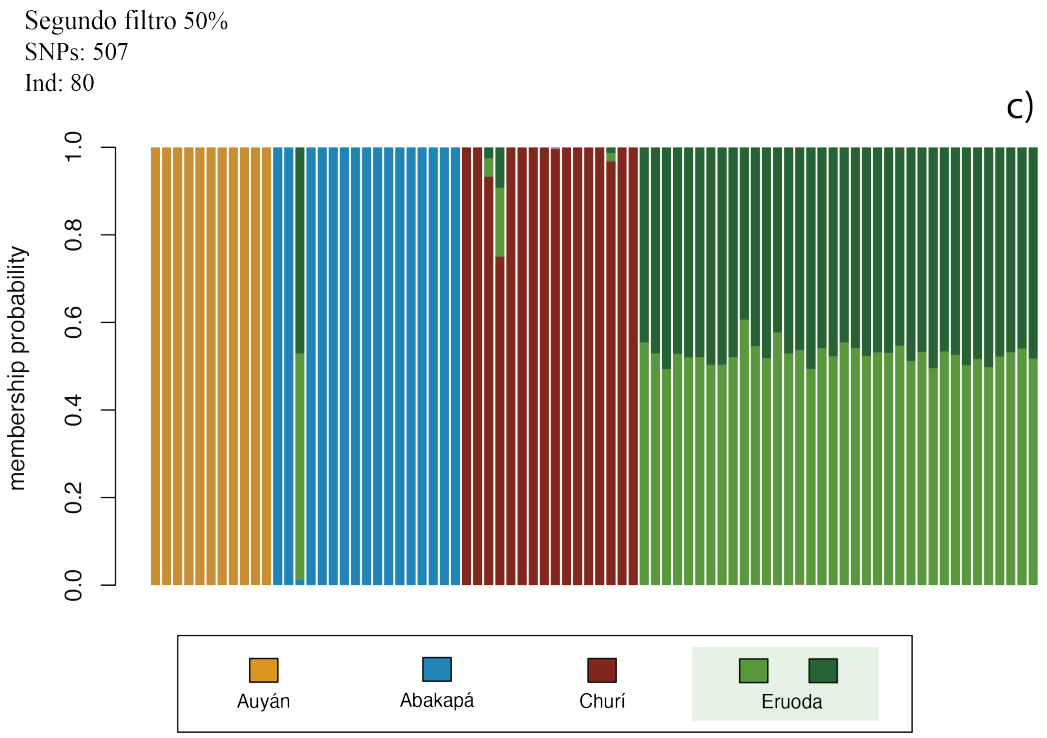
a)





b)

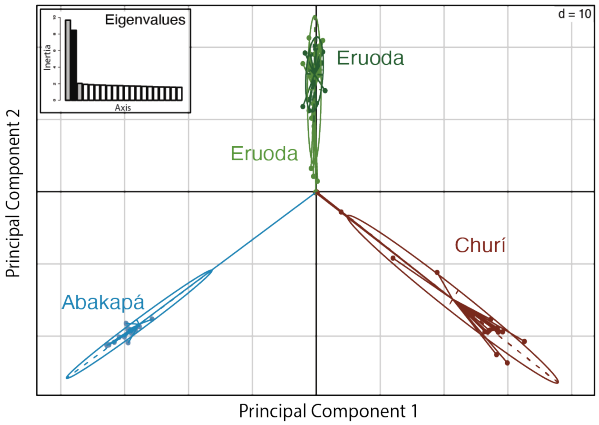


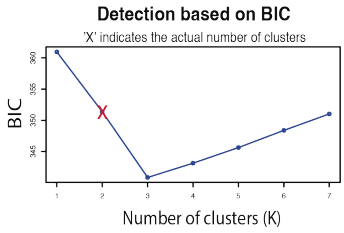


Anexo 2: Estructura de la población de *Tepuihyla.* Segundo filtro de exclusión de individuos con más del 50% de los datos (SNPs) faltantes a partir de PCA (lado izquierdo) y BIC (lado derecho). (a) Filtro de 0.2, (b) Filtro de 0.3, (c) Filtro de 0.5 y (d) Filtro de 0.7. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.

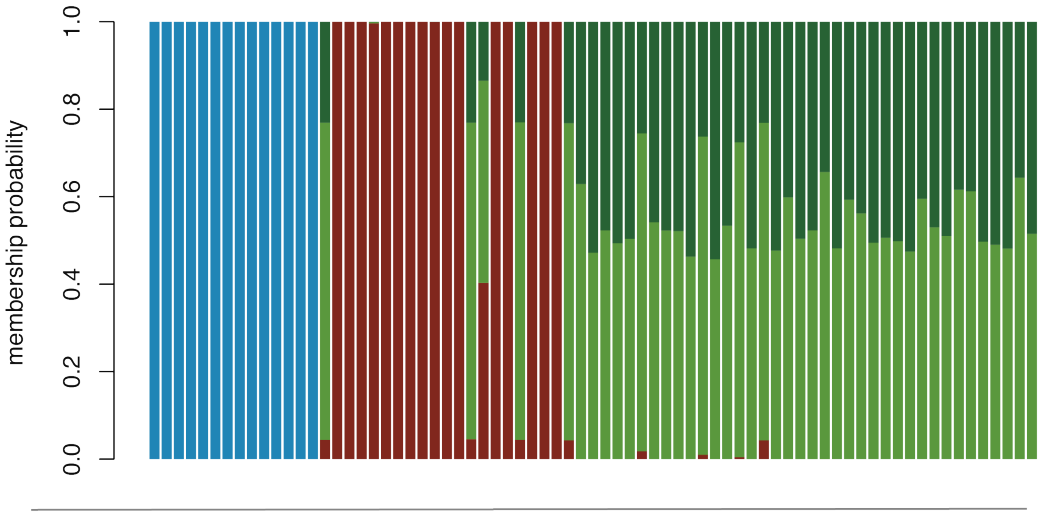
Segundo filtrado (0.3)
SNPs: 3680
Ind: 77

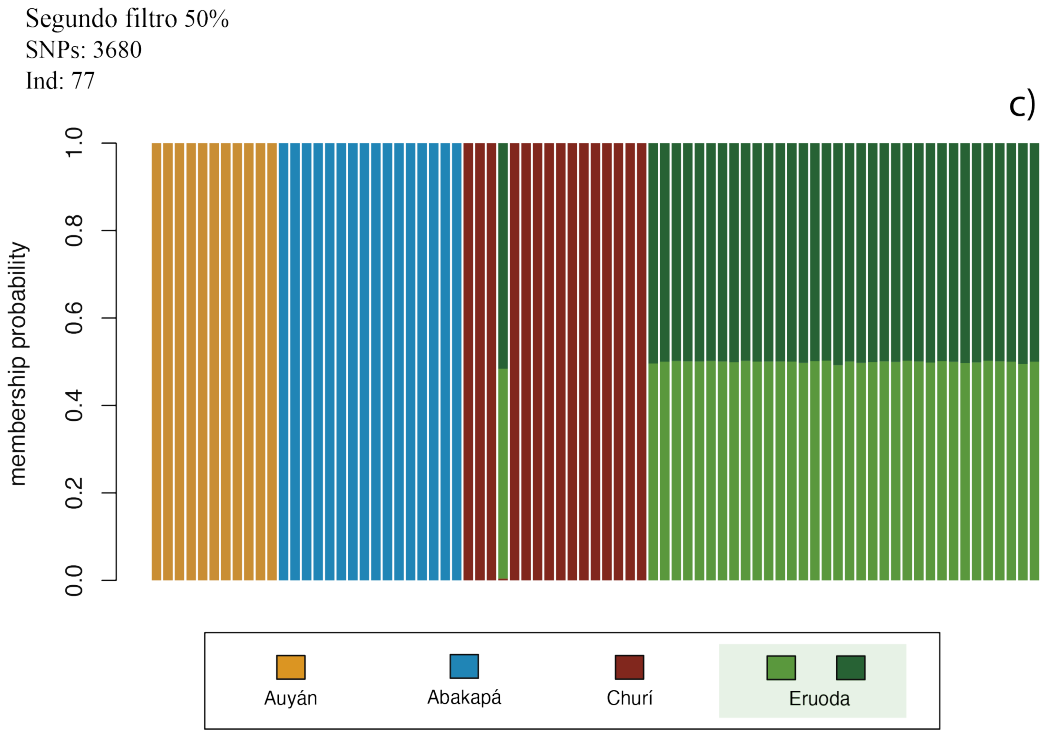
a)





b)

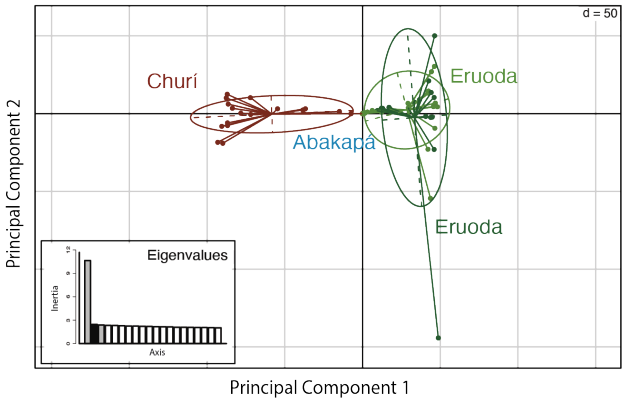


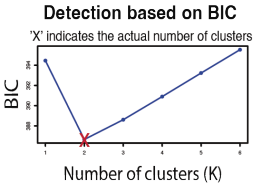


Anexo 3: Estructura de la población de *Tepuihyla.* Segundo filtro de exclusión de individuos con más del 50% de los datos (SNPs) faltantes a partir de PCA (lado izquierdo) y BIC (lado derecho). (a) Filtro de 0.2, (b) Filtro de 0.3, (c) Filtro de 0.5 y (d) Filtro de 0.7. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.

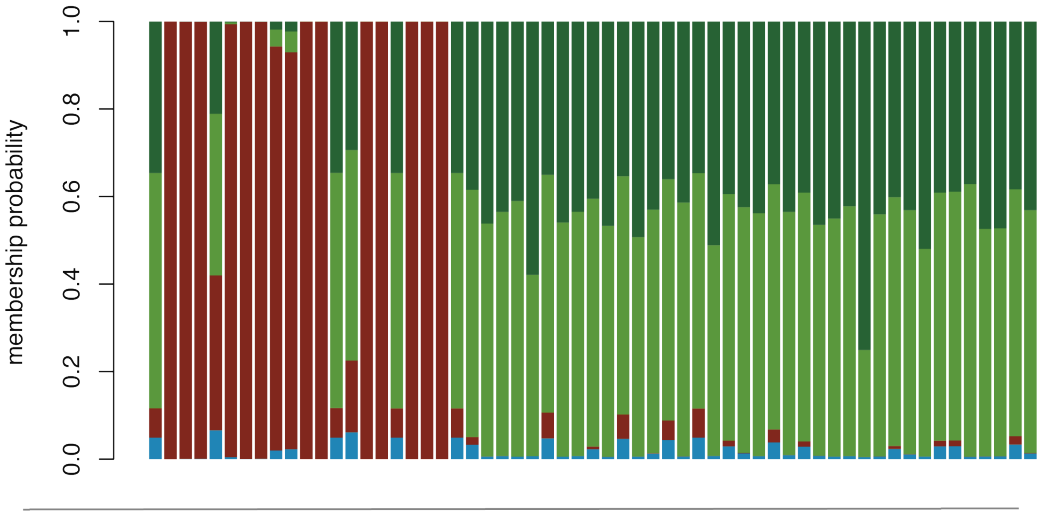
Tercer filtrado (0.5)
SNPs: 20105
Ind: 61

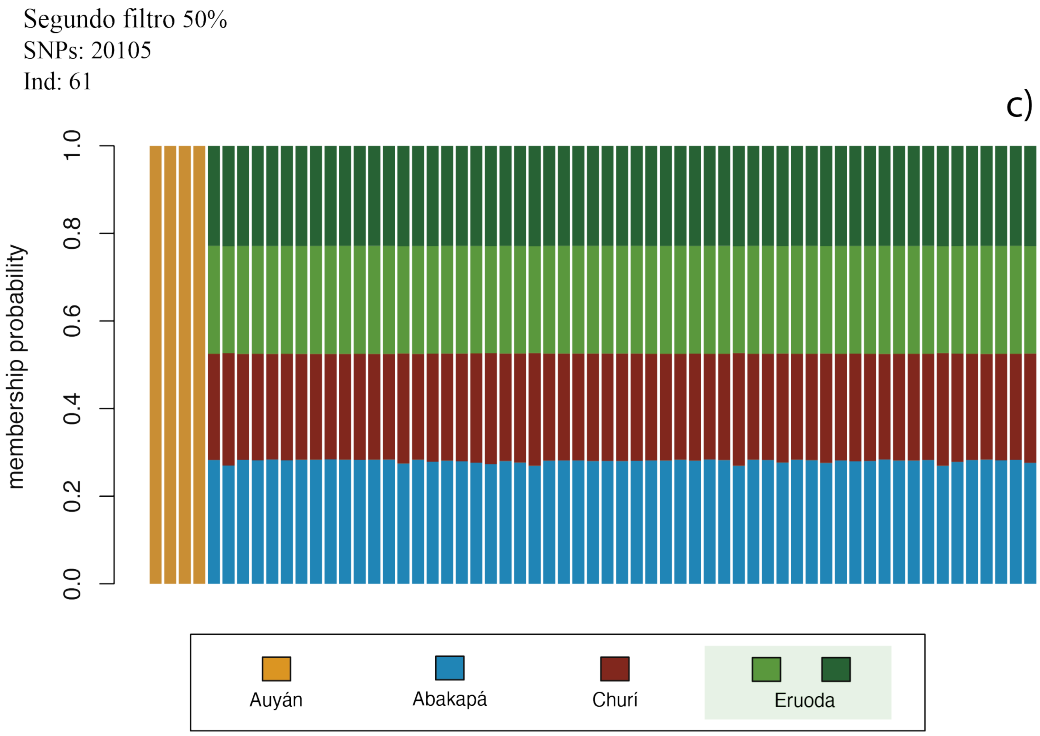
a)





b)

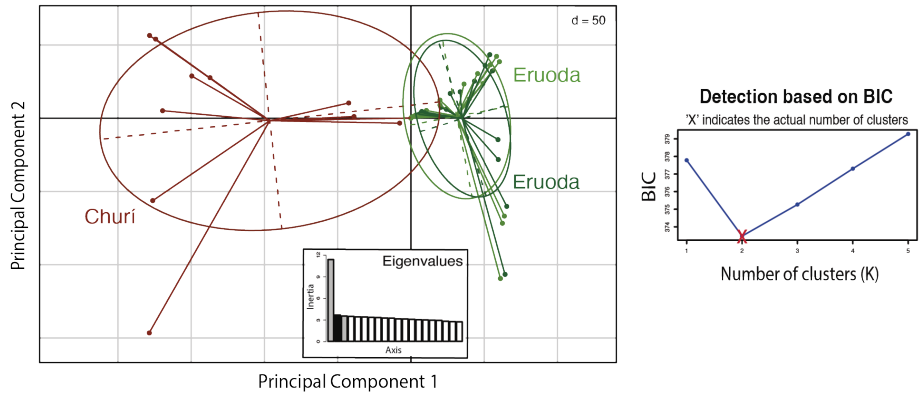




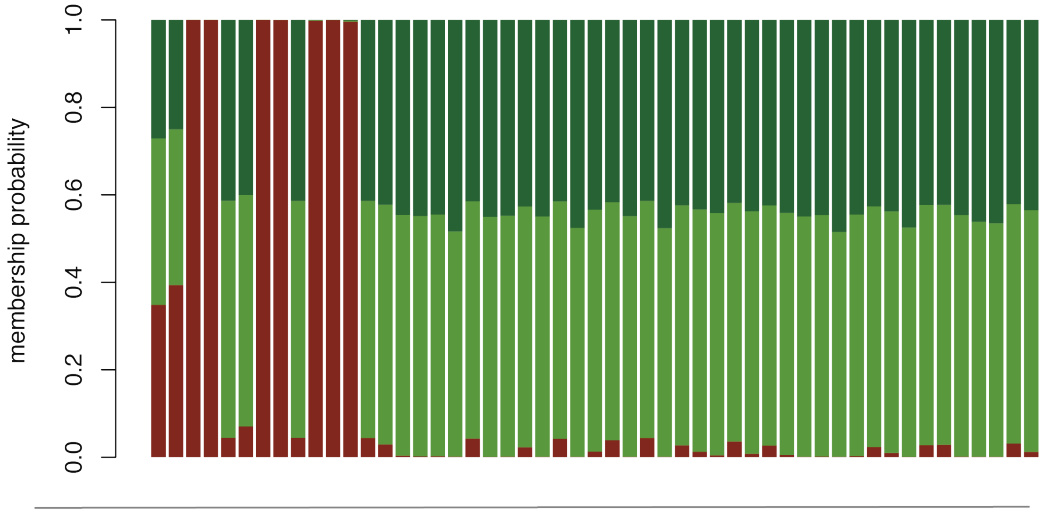
Anexo 4: Estructura de la población de *Tepuihyla.* Segundo filtro de exclusión de individuos con más del 50% de los datos (SNPs) faltantes a partir de PCA (lado izquierdo) y BIC (lado derecho). (a) Filtro de 0.2, (b) Filtro de 0.3, (c) Filtro de 0.5 y (d) Filtro de 0.7. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.

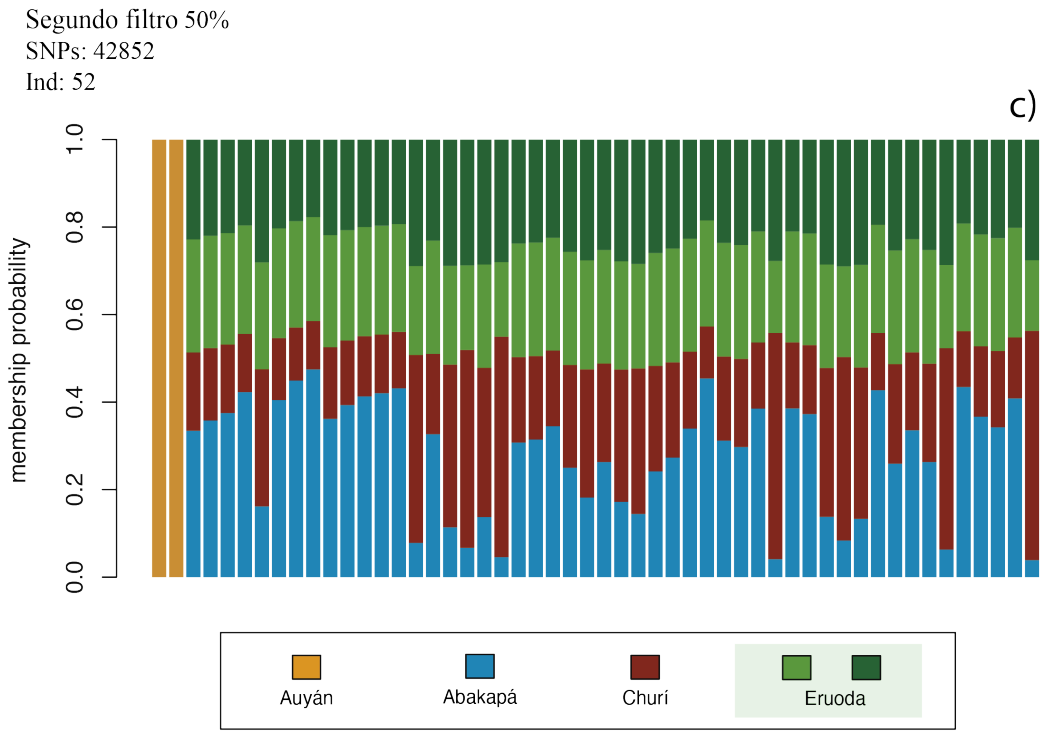
Cuarto filtrado (0.7)
SNPs: 42852
Ind: 52

a)



b)





Anexo 5: Anexo 2: Estructura de la población de Tepuihyla. Segundo filtro de exclusión de individuos con más del 50% de los datos (SNPs) faltantes a partir de PCA (lado izquierdo) y BIC (lado derecho). (a) Filtro de 0.2, (b) Filtro de 0.3, (c) Filtro de 0.5 y (d) Filtro de 0.7. Los colores fueron asignados para cada población de la siguiente manera: naranja para la localidad Auyán-tepui, celeste para la localidad de Abakapá-tepui, la localidad Churí-tepui está representada con el color vino y Eruoda-tepui con la gama del color verde para sus dos poblaciones en etapa adulta y larvaria, respectivamente.