Análisis de datos Géneticos Poblacional de Cáncer Colorrectal en México

Edison Vázquez

Centro de Investigación en Matemáticas
Unidad Monterrey

Contenido

Avances y tareas por realizar

TAREAS POR REALIZAR:

- Introducir las bases de datos obtenidas en la primera corrida del análisis de ancestralidad y observar los mapas de calor de ancestralidad. `check`
- Revisar softwares que nos permitan recuperar regiones especificas en el genotipado de los datos. `Check`
- Asociar las regiones de los SNP's con la ancestralidad de los individuos.`En proceso - 80%`

ADMIXTURE SOFTWARE

Varios autores han usado diferentes métodos para la estimación de ancestría Veronica et al. (2006); Justo et al. (2017); Nuri et al. (2014). Los softwares de mayor uso, especificamente diseñados para realizar *admixture mapping* en la literatura son los programas `STRUCTURE`, `MALDsoft`, `ADMIXMAP`, `ANCESTRYMAP` y `ADMIXTURE`. Este último calcula las estimaciones mucho mas rápido usando un algoritmo número de optimización. Especificamente ADMIXTURE es una herramienta de software para la estimación de máxima verosimilitud de ancestros individuales de conjuntos de datos genotipo SNP multilocus. ADMIXTURE utiliza un enfoque de relajación de bloques (**block relaxation**) para actualizar alternativamente la frecuencia de alelos y los parámetros de la fracción ascendente. Cada actualización de bloques se maneja resolviendo una gran cantidad de problemas de optimización convexos independientes, que se abordan usando un algoritmo de programación cuadrático secuencial rápido. La convergencia del algoritmo se acelera utilizando un novedoso método de aceleración *cuasi Newton*. El algoritmo supera a los algoritmos EM y los métodos de muestreo MCMC por un amplio margen Alexander et al. (2009). Una vista general de los programas para la estimación de ancestria se puede observar en la tabla 1.

Programa	Global/local	Sistema operativo
STRUCTURE	Global/Local	Windows/Linux/Mac
frappe	Global	Windows/Linux/Mac
ADMIXTURE	Global	Linux/Mac
EIGENSTRAT/smartpca	Global	Linux
ipPCA/EigenDev	Global	Windows/Linux(MatLab)
GEMTools	Global	Windows/Linux
PLINK	Global	Windows/Linux/Mac/C/C+
LAMP	Local/Global	Windows/Linux
SABER	Local/Global	Linux
HAPMIX	Local/Global	Unix/Linux/Windows
ANCESTRYMAP	Local/Global	Unix/Linux

Cuadro: Descripción de programas para la estimación de ancestria local o globlal Yushi et al. (2013)

ADMIXTURE SOFTWARE

A grandes rasgos, el programa ADMIXTURE estima la probabilidad para los genotipos observados basandose en las proporciones de ascedencia y frecuencias de alelos de población. Esto lo realiza simultaneamente, mientras estima las frecuencias de alelos de pobación lo hace junto a la estimación de las proporciones de ascedencia.

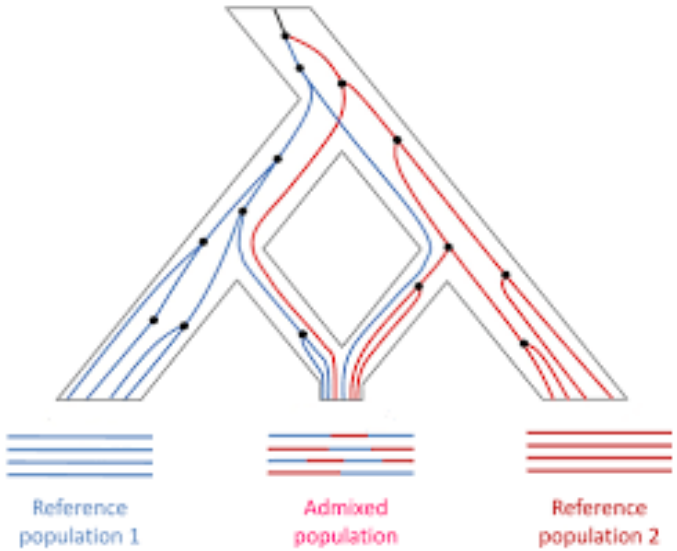

Figura: "La figura muestra un árbol de población (en gris) y un árbol de genes (en azul y rojo) que rastrea la historia evolutiva de dos poblaciones ancestrales y su correspondiente población mezclada. Debido a ILS, los haplotipos específicos de la población de referencia 1 (líneas rojas) podrían fluir y mezclarse con la población 2 y viceversa." Kai et al. (2017)

ESTIMANDO EL MEJOR K

Para la estimación de la ancestría es necesario conocer apriori las poblaciones que existen dentro de la muestra. En este caso se usa el método de *cross-validation* para identificar el valor de **K**. Este procedimiento divide los genotipos observados en v=5 (por defecto) folds de aproximadamente el mismo tamaño. El procedimiento enmascara (es decir, convierte a "MISSING") todos los genotipos, para cada fold a su vez. Osea, para cada fold, el conjunto enmascarado G resultante es usado para calcular las estimaciones $\tilde{\theta} = (\tilde{Q},\tilde{P})$.

Ante este procedimiento se corrió el primer análisis para estimar el número de poblaciones dentro de nuestros datos.

# de poblaciones	# de iteracciones	Tiempo de ejecución (min)
2	33	311
3	59	645
4	73	922
5	94	1322
	TOTAL	3200 (53 hrs.)

Cuadro: Ejecución para la búsqueda del mejor K

ESTIMANDO EL MEJOR K

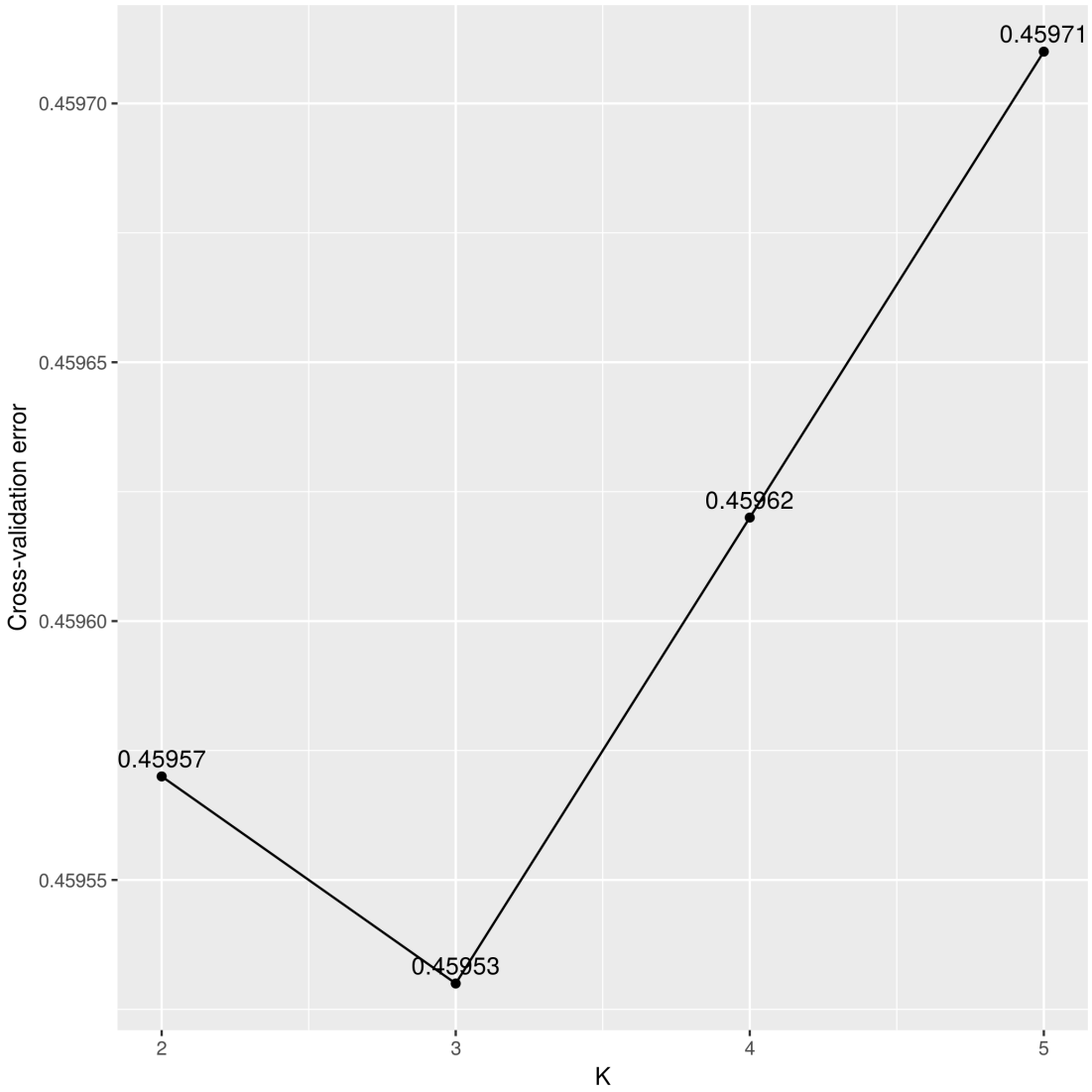

Figura: Error método de cross-validation

Gráfica de colores K=3

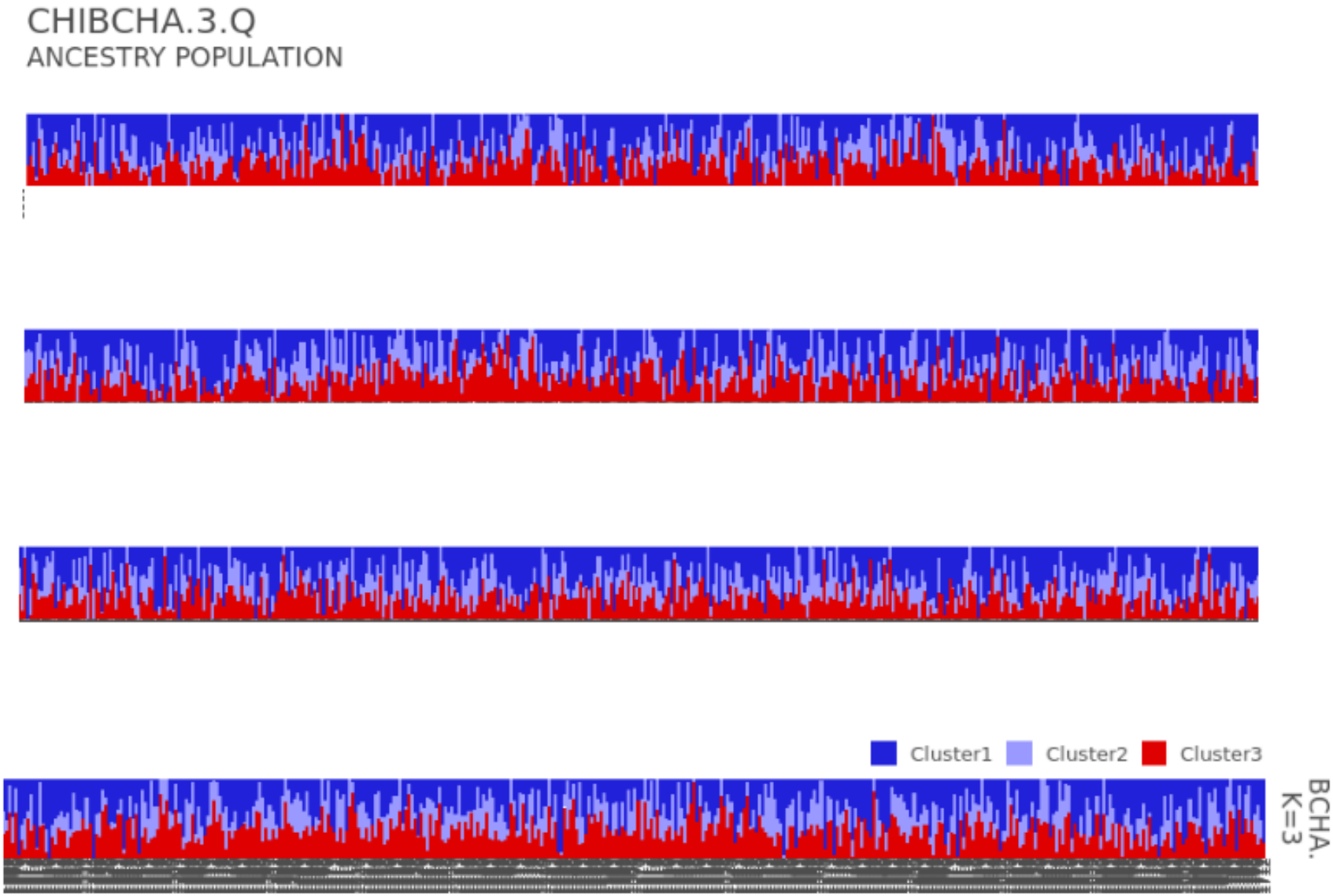

Figura: Gráfica de color de ancestria

Relacionando los cluster con población geografica

Para relacionar los colores con poblaciones se requiere correr el proyecto HAPMAP. El objetivo del Proyecto Internacional HapMap era crear un mapa de haplotipos del genoma humano. A menudo conocido como el HapMap, éste describe los patrones comunes de la variación genética humana.
El HapMap proporciona un recurso clave que los investigadores pueden usar para encontrar genes que afectan a la salud, la enfermedad y las respuestas a los medicamentos y los factores ambientales. La información producida por el proyecto está ahora disponible gratuitamente en bases de datos públicas a los investigadores alrededor del mundo.

Relacionando los cluster con población geografica

Las siguientes muestras de poblaciones fueron estudiadas en este proyecto:

```
CEU Utah residents with Northern and Western European ancestry from the CEP
collection
CHB Han Chinese in Beijing, China
CHD Chinese in Metropolitan Denver, Colorado
GIH Gujarati Indians in Houston, Texas
JPT Japanese in Tokyo, Japan
LWK Luhya in Webuye, Kenya
MXL Mexican ancestry in Los Angeles, California
MKK Maasai in Kinyawa, Kenya
TSI Toscani in Italia
YRI Yoruba in Ibadan, Nigeria
```

TAREAS POR REALIZAR

- Asociar las regiones de los SNP's con la ancestralidad de los individuos. `En proceso - 80%`
- Asociar los SNP's que tienen relacioón con el CCR y observar su procedencia de ascedencia.

Referencias

M. Veronica, V. Adan, C. Emily, et al. Admixture in mexico city: implications for admixture mapping of 2 diabetes genetic risk factors. *Hum Genet*, 120, 2006.

L. Justo, B. Felix, G. Rosa, et al. Subtypes of native american ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in chile. *PLoS Genet*, 13, 2017.

K. Nuri, P. Alvaro, G. Barbara, et al. Human and helicobacter pylori coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences*, 111(4), 2014.

D. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 2009.

L. Yushi, N. Toru, L. Shuguang, et al. Softwares and methods for estimating genetic ancestry in human populations. *Hum Genomics*, 7(1), 2013.

Y. Kai, Z. Ying, N. Xumin, et al. Models, methods and tools for ancestry inference and admixture analysis. *Quantitative Biology*, 5(3), 2017.

GRACIAS