Flujo de trabajo bioinformático para estudios de asociación de genoma completo en plantas

6 de septiembre de 2019

Resumen

Presentamos un flujo de trabajo blgioinformático para estudios de asociación de genoma completo (GWAS) que permite realizar el proceso completo de GWAS en plantas. El flujo de trabajo está dividido en diferentes etapas correspondientes al proceso general de GWAS. En cada etapa se tienen una o más técnicas que pueden seleccionarse para realizar esa etapa del proceso. El flujo de trabajo se ejecuta en modo línea de comandos, ya sea completamente o etapa por etapa. Para la implementación se utilizan diferentes tipos de software entre programas, librerías y scripts en diferentes lenguajes, los cuales están disponibles gratuitamente y que son de amplio uso en proyectos GWAS. Este software se envuelve sobre otra capa de software que oculta los detalles y pone una interfaz de comandos muy simple de usar para personas no expertas en computación.

1. Flujo de Trabajo GWAS

1.1. Formatos de Archivos de Entrada

Se necesitan dos archivos de entrada: uno con la información del genotipo/fenotipo y otro con la información de los SNPs. Estos archivos están son de tipo texto y sigen el formato estándar establecido por el software PLINK [1]. El primer archivo, con extensión *.ped,* contiene la información del genotipo con 6 columnas iniciales fijas seguidas después por la información del SNP, así: (1) ID de la famila, (2) ID del individuo, (3) ID parental, (4) ID maternal, (5) Genero, y (6) Fenotip (Valor cuantitativo ó Estado de afectación). Las columnas 7 y 8 contienen el código de los alelos observados en el primer SNP, las columnas 9 y 10 para los alelos en del seguno SNP y así sucesivamente.

El segundo archivo, con extensión *.map,* contiene la información del SNP y su posición, una línea por cada SNP y con las siguientes 6 columnas: (1) Número del cromosoma, (2) Identificador de variante, (3) Distancia genética (en Morgans), y (4) Posición física (en pares de bases).

Generalmente, los anteriores archivos son muy grandes y se necesita transformarlos de tipo texto a binario para procesarlos más fácilmente. Al transformalos a binario se generan tres archivos con extensiones *.bed*, *.bid*, y *.fam*, que sigen también los formatos establecidos por PLINK[1]. Los archivos *.bed* contienen el registro del genotipo de cada individuo y sus primeros seis campos son: (1) ID de la famila, (2) ID del individuo, (3) ID parental, (4) ID maternal, (5) Genero, y (6) Estado de afectación. Los campos siguientes del registro corresponden a información del genotipo. Los archivos *.bim* contienen la información del SNP y su mapa de posiciones a través de 6 campos: (1) Número del cromosoma, (2) Identificador de variante, (3) Distancia genética (en Morgans), (4) Posición física (en pares de bases), (5 y 6) Alelo 1 y alelo 2. Y los archivos *.fam* contienen la información del pedigree/fenotipo, así: (1) ID de la famila, (2) ID de la muestra, (3) ID parental, (4) ID maternal, (5) Genero, y (6) Fenotipo.

1.2. Preprocesamiento

En esta etapa se realiza un procesamiento de los datos para excluir individuos y SNPs que no cumplan algunas de las siguientes condiciones:

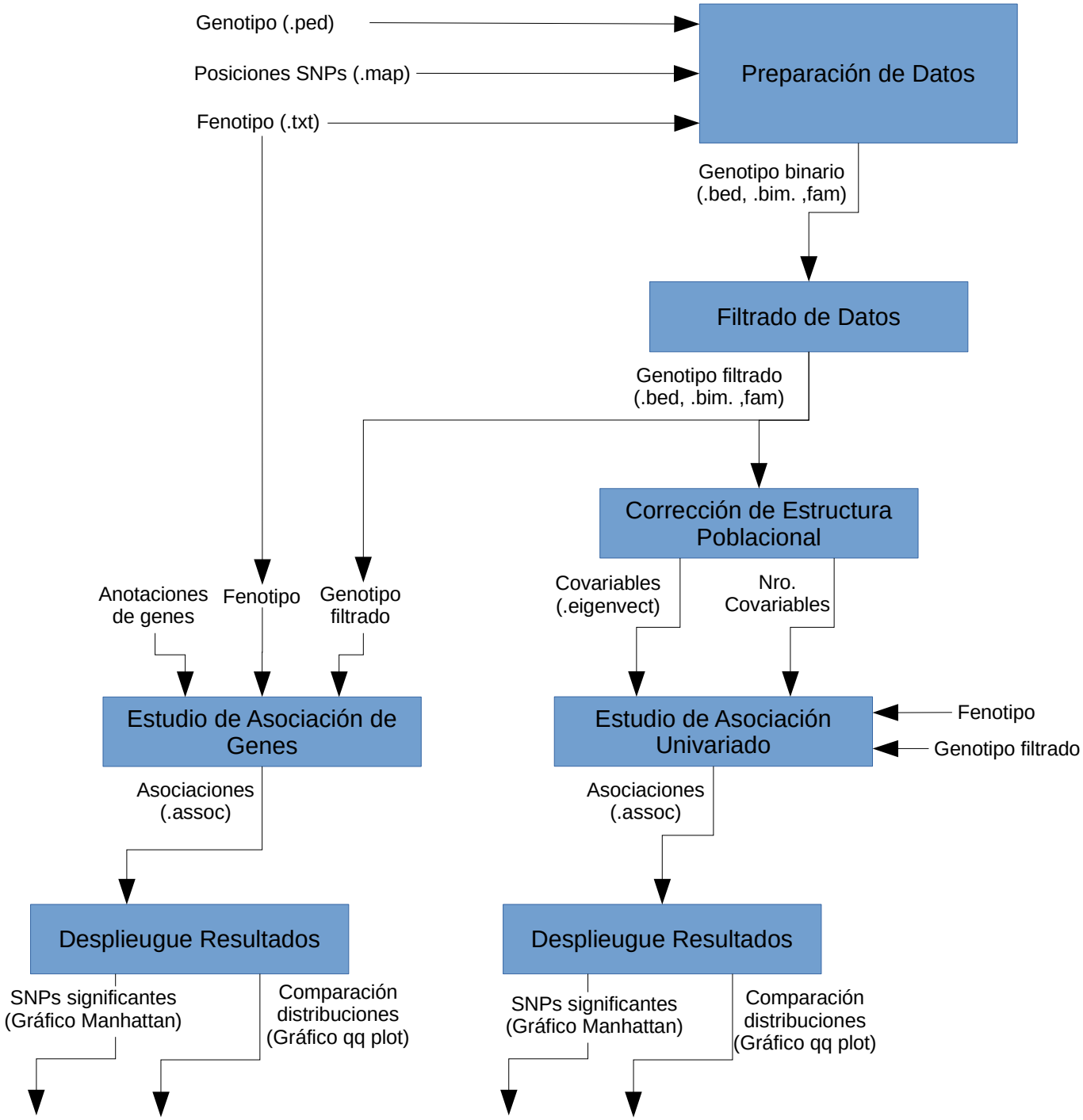


Figura 1: Flujo de trabajo GWAS para plantas.

FAM1	NA06985	0	0	1	1	A	T	T	T	G	G	C	C	A	T	T	T	G	G	C	C
FAM1	NA06991	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA06993	0	0	1	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G	C	T
0	NA06994	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA07000	0	0	2	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G	C	T
0	NA07019	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA07022	0	0	2	1	C	T	T	T	G	G	0	0	C	T	T	T	G	G	0	0
0	NA07029	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
FAM2	NA07056	0	0	0	2	C	T	T	T	A	G	C	T	C	T	T	T	A	G	C	T
FAM2	NA07345	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C

Figura 2: **Archivo de genotipo formato PLINK**. Columnas: (1) ID de la famila, (2) ID del individuo, (3) ID parental, (4) ID maternal, (5) Genero, y (6) Fenotipo (Valor cuantitativo ó Estado de afectación). Las columnas 7 y 8 contienen el código de los alelos observados en el primer SNP, las columnas 9 y 10 para los alelos en del seguno SNP y así sucesivamente.

21	rs11511647	0	26765
X	rs3883674	0	32380
X	rs12218882	0	48172
9	rs10904045	0	48426
9	rs10751931	0	49949
8	rs11252127	0	52087
10	rs12775203	0	52277
8	rs12255619	0	52481

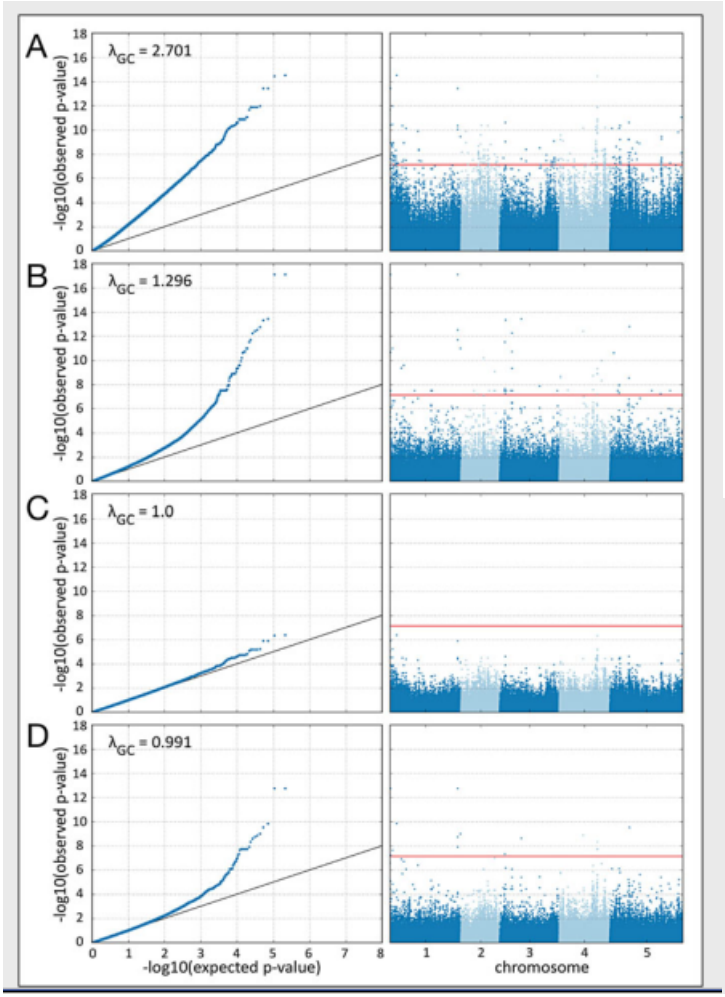
Figura 3: Archivo posiciones SNPs formato PLINK. Columnas: (1) Número del cromosoma, (2) Identificador de variante, (3) Distancia genética (en Morgans), y (4) Posición física (en pares de bases).

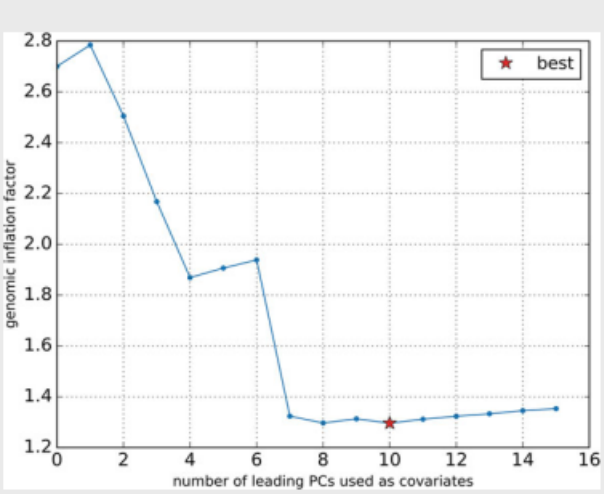
FID	IID	FT_Field
8353	8353	91.9233
6911	6911	111.225
6906	6906	114.658
8213	8213	126.735
7081	7081	136.856
6970	6970	136.865
6971	6971	138.002
6961	6961	139.717
6988	6988	143.812
6904	6904	145.313
6933	6933	145.696
6910	6910	148.116

Figura 4: Formato archivo fenotipo.

- Filtro de equilibrio de Hardy-Weinberg (HWE), donde se descarta los SNPs que se desvian del equilibrio de Hardy-Weinberg con un *p-valor* de la prueba estadística $<$ 1e-10.
- Filtro de fecuencia del alelo menos común (MAF), que descarta los SNPs con una MAF $<$ 1 %.
- Filtro de perdida de genotipos por individuo, que descarta individuos que tienen perdida de genotipos en más del 10 % de los SNPs.
- Filtro de perdida de genotipos por SNP, que descarta los SNPs que tienen una perdida de genotipo en más del 10 % de los individuos.

1.3. Corrección Estructura Poblacional





2. Implementación Python

Referencias

[1] S Purcell. PLINK (1.07). Documentation. http://zzz.bwh.harvard.edu/plink/dist/plink-doc-1.07.pdf. Accessed 28 Agosto, 2010.