Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Genetics and population analysis

MultiGWAS: Integrando múltiples herramientas para realizar GWAS en tetraploides

Luis Garreta [1,*], Paula Reyes [1] and Ivania Cerón [1,*]

[1] Colombian Agricultural Research Corporation (Agrosavia), Kilómetro 14, Vía a Mosquera, 250047, Colombia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Muchas de las plantas son tetraploides sin embargo la mayoria de herramientas para realizar estudios de asociación de genoma completo (GWAS) están desarrolladas para especies tetraplodies.
Results: Presentamos aquí un flujo de trabajo que realiza GWAS sobre plantas tetraplodes utilizando cuatro herramientas de GWAS: dos diseñadas para especias tetraploides, y dos diseñadas para diploides. La herramienta tomo como datos de entrada un archivo de genotipo y otro de fenotipo, los preprocesa y los ejecuta en paralelo en las cuatro herramientas. De los resultados finales de cada herramienta se construye un resumen tanto gráfico como de forma tabular para que le ayuden al investigador a realizar un mejor análisis.
Availability: Tanto la herramienta como su código fuente son de libre acceso y se puede descargar del repositorio: https://github.com/agrosavia/gwas-polypilene
Contact: lgarreta@agrosavia.co
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Although many important crops are polyploids and new genomic data of polyploid plants is available, most of the GWAS software tools has been developed for diploid organisms. Currently, two software tailored for polyploid organisms are the R package GWASpoly and the SHEsis tool. However diploid GWAS software has been also used to analyze polyploids organisms by «diploidizing» the marker data (Simko I. and Jones, 2006), from these, plink and tassel are two of the most widely used software packages for GWAS, the first tipically used for humans and animals, and the second for plants.

The goal of this tool is to do GWAS analyses using multiple GWAS software in order to pool and integrate information from their results in such a way that could help researchers in the selection of true associations across GWAS.

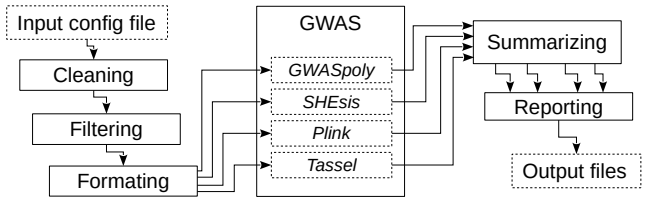
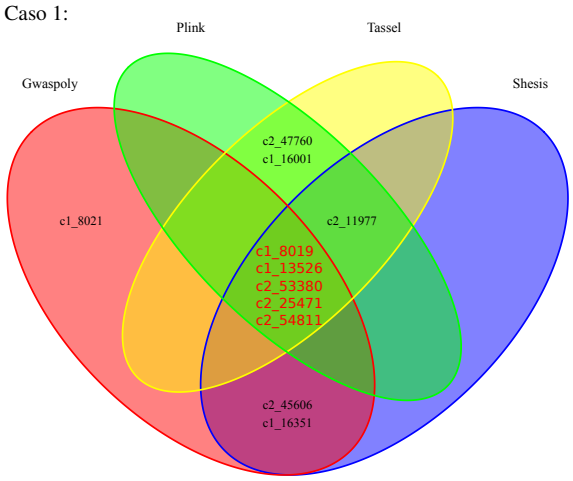


Fig. 2. Multi-GWAS workflow.

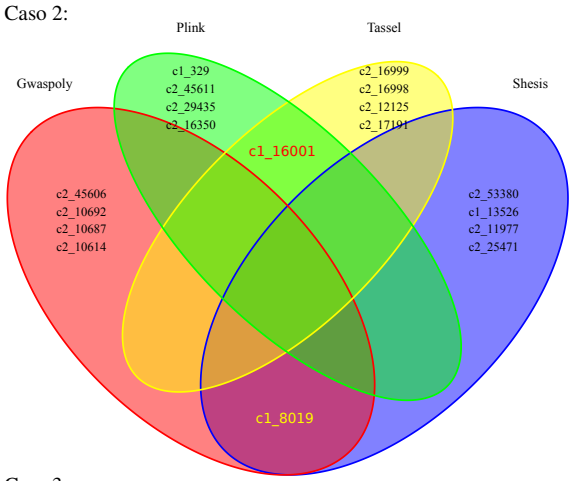
2 Métodos

El control de calidad de los datos se lleva a cabo a través de filtros que se aplican tanto al genotipo y al fenotipo y que buscan eliminar individuos y marcadores que no cumplan ciertos criterios y que pueden sesgar el análisis. Algunos filtros se aplican por defecto, como la eliminación de duplicados (tanto de individuos como de marcadores) y la eliminación de individuos sin información genotípica. Otros filtros se pueden seleccionar o ajustar sus valores a través del archivo de configuración, entre estos: frecuencia del alelo menor, porcentaje de individuos/marcadores con perdida de información, y marcadores en equilibrio de Hardy-Weinber.

Caso 1:



Caso 2:



Caso 3:

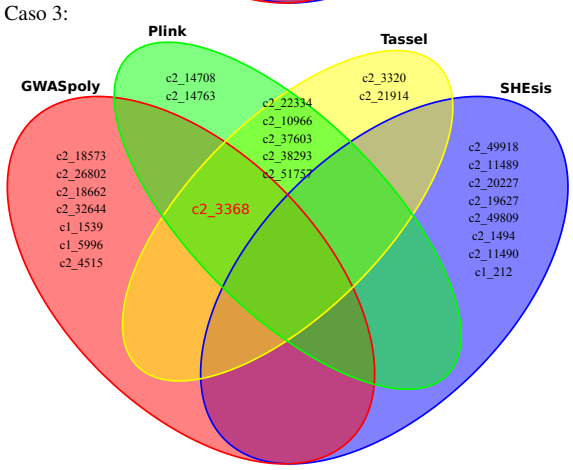


Fig. 1. MultiGwasTool aplicado en tres rasgos de poblaciones de papa. En rojo los marcadores comúnes. Caso1: las cuatro herramientas encuentran marcadores comunes. Caso 2: Las dos herramientas diploides encuentran un marcador común, mientras que las dos poliploides encuentran un marcador común diferente. Caso 3: Tres de las cuatro herramientas encuentran marcadores comunes.

MultiGWAS ejecuta GWAS usando dos modelos estadísticos de los cuatro propuestos por Sharma et al.: modelo Naive y modelo Completo[1]. El modelo Naive no realiza corrección de estructura poblacional ni considera las relaciones de parentesco; mientras que el moodelo Completo realiza la corrección de estructura poblacional mediante el cálculo de los 10 primeros componentes principales que se ajustan como efectos fijos en cada herramienta; y las relaciones de parentesco entre individuos que se ajustan de acuerdo al algoritmo utilizado por defecto en cada herramienta. GWASpoly calcula el parentesco mediante la matriz K como MM^T, donde M es la matriz de genotipos centrada (individuos x marcadores) (Rosyara *et al.*, 2016); en Plink, primero se calculan los individuos relacionados mediante el algoritmo KING-Robust implementado en la herramienta king (Manichaikul *et al.*, 2010) y luego se los remueve de los datos del genotipo; y Tassel que calcula el parentesco mediante el método Centered-Identity by State (Centered-IBS) de Endelman and Jannink (Endelman and Jannink, 2012).

2.1 GWAS

3 Implementación

MultiGWAS está implementado en R como una aplicación que se ejecuta en línea de comandos, ya sea en una terminal en ambiente Linux o Mac. La entrada consiste en un único archivo de configuración donde se especifica los nombres de los archivos de entrada (genotipo y fenotipo), el modelo de GWAS a ejecutar (Naive o Completo), y la especificación de los valores de los distintos filtros para aplicar en el genotipo y fenotipo.

MultiGWAS integra cuatro herramientas de GWAS que deben estar previamente instaladas. Esto se puede llevar ya sea de forma automática, a través de la ejecución de un script propio de configuración que acompaña a MultiGWAS; o de forma manual, seguiendo las instrucciones de instalación propias de cada herramienta: GWASPoly (https://potatobreeding.cals.wisc.edu/software); Plink utiliza tres programas binarios: plink1.9 (https://www.cog-genomics.org/plink), plink2.0 https://www.cog-genomics.org/plink/2.0), y king (`http://people.virginia.edu/~wc9c/KING`). Tassel (https://www.maizegenetics.net/tassel),

4 Resultados and Discussion

En la figura 1 se muestran tres resultados producidos por MultiGWAS al ejecutarse sobre dos conjuntos de datos: el genotipo y fenotipo recolectados como parte del proyecto Solanaceae Coordinated Agricultural Project (SolCAP) y usados para probar el software de GWASpoly (Rosyara *et al.*, 2016); y el genotipo y fenotipo recolectados en el estudio de la Colección Central Colombiana de Papa (CCC) (Berdugo-Cely *et al.*, 2017).

En el primer caso, figura 1.A, las cuatro herramientas coinciden en un conjunto grande de marcadores significativos, ya que el modelo GWAS aplicado fue Naive, sin ningún control de estructura poblacional ni consideración de relaciones de parentesco, y con la posibilidad de encontrar asociaciones falsas. Entre esos marcadores comunes, el primero de la lista y de mayor puntaje es el c1_8019, reportado por Rosyara et al. (Rosyara *et al.*, 2016) como rasgo cuantitativo más significante. En el segundo caso, figura 1.B, se ejecutó MultiGWAS sobre el mismo conjunto de datos pero ahora con el modelo Completo de GWAS, el cual filtra los posibles falsos negativos reduciendo el número de marcadores comunes a dos, c1_16001 y c1_8019, el primero encontrado como más significativo solo por las herramientas diploides (Plink y Tasse), y el segundo encontrado solo para las herramientas poliploides (GWASpoly y SHEsis). Este último es el mismo marcador c1_8019 encontrado por las cuatro herramientas en el caso anterior. El tercer caso, figura 1.C, el análisis se realizó sobre el conjunto de datos de la CCC (ver métodos) y muestra como tres de las cuatro herramientas encuentran un marcador común. En este caso el modelo utilizado fue el Completo, que lo soportan estas tres herramientas y que permite filtrar falsas asociaciones. Sin embargo, la herramienta SHEsis

[1] Excepto la herramienta SHEsis que solo soporta el modelo Naive.

soporta solo el modelo Naive, sin control de los posibles falsos positivos y por lo tanto con resultados muy diferentes a los de las otras herramientas.

References

Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., and Yockteng, R. (2017). Genetic diversity and association mapping in the Colombian Central Collection of Solanum tuberosum L. Andigenum group using SNPs markers. *PLOS ONE*, **12**(3), e0173039.

Endelman, J. B. and Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, **2**(11), 1405–1413.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22), 2867–2873.

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, **9**(2), 0.

Sharma, S. K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., and Bryan, G. J. (2018). Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, **8**(10), 3185–3202.

Simko I., K. G. H. and Jones, R. W. (2006). Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics*, **173**, 2237–2245.