

## Flujo de trabajo bioinformático para estudios de asociación de genoma completo (GWAS) en plantas de papa tetraploide

Luis Garreta, Paula Reyes, Ivania Cerón

Corporación colombiana de investigación agropecuaria, AGROSAVIA

5 de Noviembre del 2019

- 1 Agrosavia
- 2 Propuesta
- 3 GWAS
- 4 Pipeline
- 5 Resultados preliminares
- 6 Implementación

# AGROSAVIA: Corporación Colombiana de Investigación Agropecuaria

**Agrosavia** (antes Corpoica) es una entidad pública, financiada principalmente por el ministerio de Agricultura, con sedes en todo el país.



## Misión

Contribuir al cambio técnico para mejorar la productividad y competitividad del sector agropecuario colombiano.

# Bancos de Germoplasma para alimentación y agricultura

AGROSAVIA desde el año 1994, tiene a cargo el Sistema de Bancos de Germoplasma de la Nación Colombiana, el cual engloba tres subsistemas:

- Vegetal,
- Animal y
- Microorganismos.



**Bancos de Germoplasma:** Repositorios destinados a la conservación de la diversidad genética de diferentes especies de interés agropecuario y de importancia económica para el país (**Patrimonio nacional**).

# Banco de Germoplasma de Papa: Colección Central Colombiana de Papa



## Trabajo parcial: Pipeline para GWAS en tetraploides

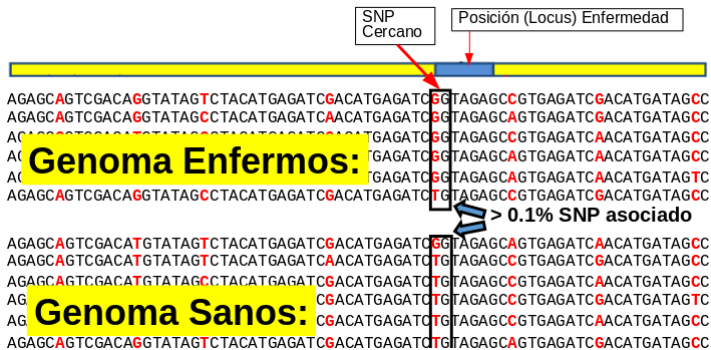
- Automatización del proceso de GWAS
- Se reutiliza software (librerías, aplicaciones), , no se implementa modelos estadísticos.
- Se integra varias herramientas: GWASpoly, plink, tassell
- Resultados previos, sin análisis

## Objetivo general:

Identificar una estrategia óptima para asociar SNPs y características de interés agronómico en papas nativas diploides y tetraploides utilizando algoritmos de selección genómica (SG) y GWAS (Genome-Wide Association Studies).

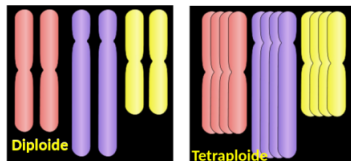
# GWAS: Estudios de Asociación de Genoma Completo

Los GWAS implican la búsqueda de marcadores (SNPs) a través del genoma de varios individuos. Estos SNPs pueden indicar variaciones genéticas asociadas con un rasgo particular (fenotipo, enfermedad).



- Uno de los cultivos más importantes: *4o en el mundo*.

- Dos poblaciones (papas nativas):  
*diploide y tetraploide*.



- La principal especie cultivada *Solanum tuberosum (tetraploide)*
- Una de las mayores subespecies es la *Andigena*: (e.g papa amarilla)
- Sufren varias enfermedades: *gota*, *polilla guatemalteca*

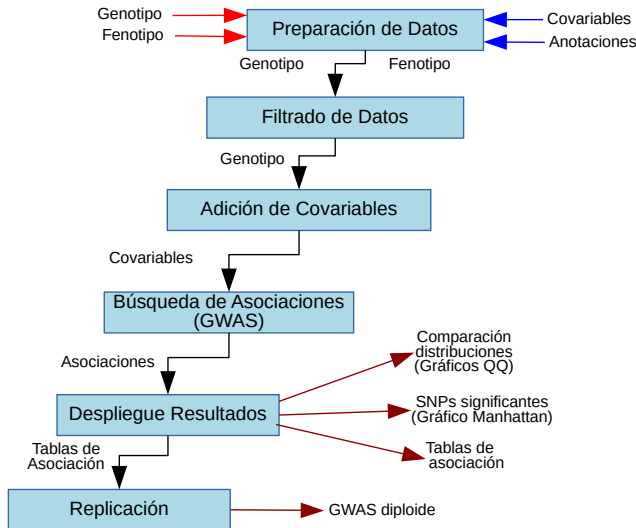


- Los GWAS son **ampliamente usados en especies diploides** y varios paquetes de software se han desarrollado basados en varios modelos estadísticos.
- Sin embargo, software GWAS **para poliploides es muy escaso**, y los que existen son librerías para usarlas dentro de un lenguaje de programación.

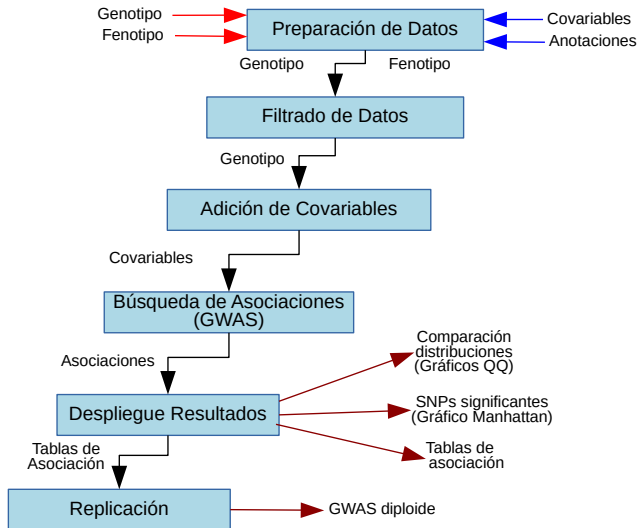
Nombre	Especie	Tipo
PLINK	Diplo	Programa
TASSEL	Diplo	Programa
GAPIT	Diplo	Programa
FaST-LMM	Diplo	Programa
GWASpoly	Poly	Librería de R
....		

Yan et al, 2019. Effects of input data quantity on genome-wide association studies (GWAS). Int. J. Data Mining and Bioinformatics, Vol. 22, No. 1, 2019

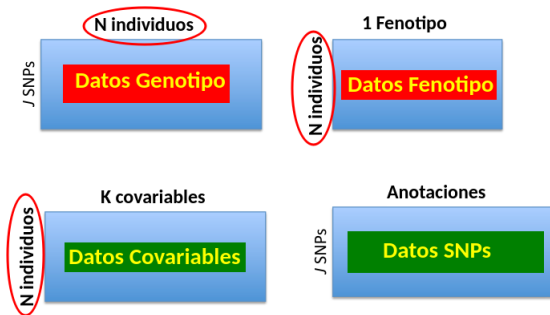
# Flujo de Trabajo o Pipeline para GWAS en Tetraploides (Caso Papa)



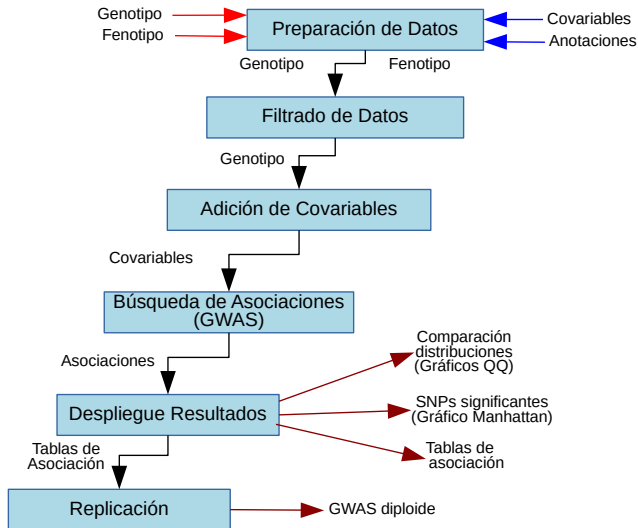
# Proceso 1: Preparación de Datos



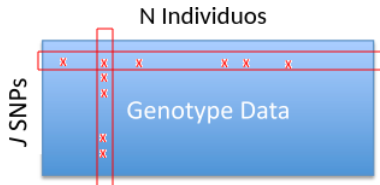
- Archivos de diferentes fuentes
- Individuos no coinciden o no están



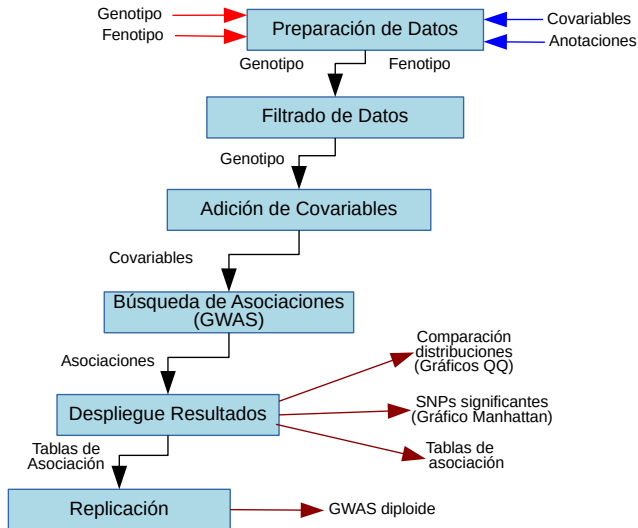
# Proceso 2: Filtrado de Datos



- Filtrado por genotipos perdidos (invalidos)
  - ▶ Muchos SNPs inválidos
  - ▶ Muchos individuos inválidos
- Filtrado por frecuencia del alelo menor ( $<0.05$ )



# Proceso 3: Adición de Covariables



- Evitar Falsos Positivos

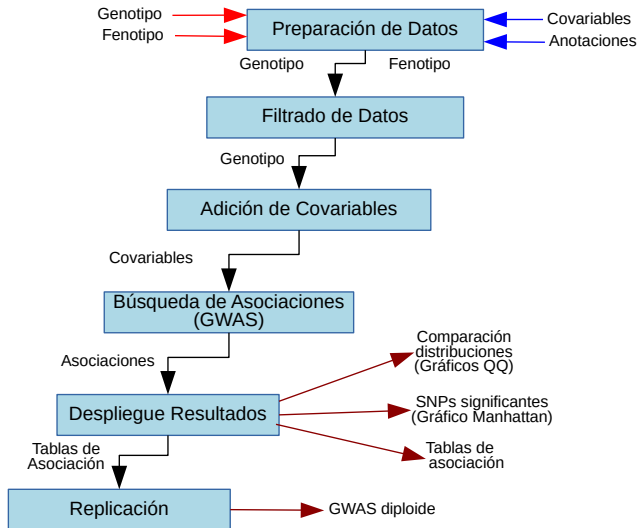
- ▶ Toma de entrada el archivo con la estructura poblacional, si se conoce
- ▶ Se calcular el parentesco (kinship)

- Cuatro tipos de GWAS:

- ▶ **Naive:** Sin covariables
- ▶ **Kinship:** Cálculo de parentesco
- ▶ **Structure:** Archivo Estructura poblacional
- ▶ **Kinship+Structure**



# Proceso 4: Búsqueda de Asociaciones



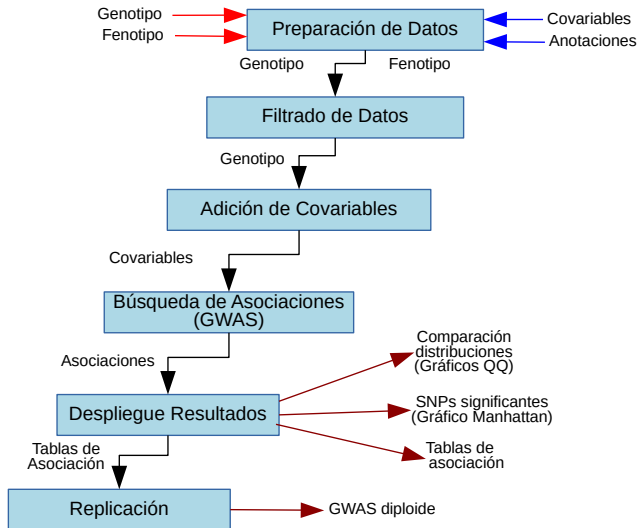
- Implementado sobre la librería de R *GWA Spoly*<sup>1</sup>
  - ▶ Asociaciones basadas en *Q+K Linear Mixed Model*<sup>1</sup>:

$$y = ZQv + Zu + ZS\tau + \varepsilon$$

- ▶ Que tiene en cuenta:
  - ★ La estructura poblacional (Q) y el parentesco (K)
  - ★ Y es adecuado para fenotipos cuantitativos (e.g. grosor tallo).
- ▶ Donde:
  - ★ *y*: vector de fenotipos observados
  - ★ *v*: vector de efectos poblacionales
  - ★ *u*: efecto poligenico aleatorio (K matriz de parentescos)
  - ★ *τ*: efectos de los SNPs
  - ★ *ε*: efectos residuales aleatorios
  - ★ *S*: Depende del modelo genético que se asuma: General, aditivo, simplex-dominante, duplex-dominante

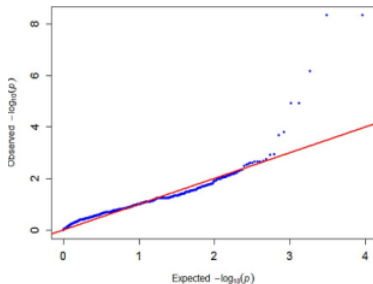
<sup>1</sup> Rosyare et al. *Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato, The Plant Genome*, 2016.

# Proceso 5: Despliegue de Resultados (Plots)

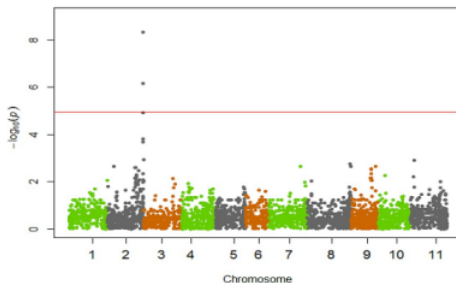


# Proceso 5: Despliegue de Resultados (Plots)

## QQ-Plots



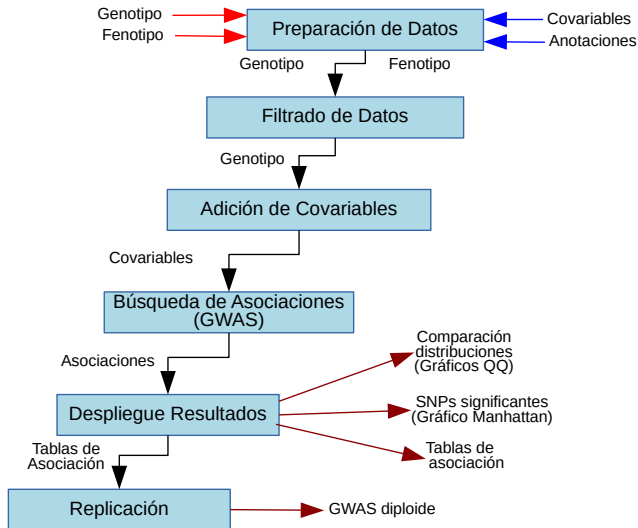
## Manhattan Plots



## Tablas de Asociación

Ploidy	Type	GC	Model	Score	Threshold	Effect	Marker	Chr	Position	Ref	Alt	Annotation	Arabidopsis
4	Kinship+Structure	1.10	1-dom-ref	5.14	4.44	0.57	snp_c2_40748	12	12654160	0	1	Entaticopeptide repeat-containing prote	AT2G37230.1
4	Kinship+Structure	1.10	1-dom-ref	4.79	4.44	0.64	snp_c1_7325	2	42764457	0	1	POM30	AT5G67500.1
4	Kinship+Structure	1.01	1-dom-alt	6.35	4.35	-0.78	snp_c1_3484	4	71945107	0	1	Cinnamoyl-CoA reductase	AT2G33590.1
4	Kinship+Structure	1.01	1-dom-alt	5.55	4.35	-0.69	snp_c2_10568	4	71954528	0	1	4-hydroxy-3-methylglutaryl coenzyme A redu	AT1G76490.1
4	Kinship+Structure	1.01	1-dom-alt	5.35	4.35	-0.59	snp_c2_36664	1	535454	0	1	Adagio protein 3	AT1G68050.1
4	Kinship+Structure	1.01	1-dom-alt	4.98	4.35	-0.59	snp_c2_4875	1	83177518	0	1	Conserved gene of unknown function	AT5G65810.1
4	Kinship+Structure	0.80	2-dom-ref	9.1	4.48	-0.76	snp_c2_2998	9	58689196	0	1	Gene of unknown function	No Hit
4	Kinship+Structure	0.80	2-dom-ref	4.84	4.48	0.55	snp_c1_10855	7	9998099	0	1	DNA repair protein RAD51 homolog	AT5G20850.1
4	Kinship+Structure	0.80	2-dom-ref	4.65	4.48	-0.44	snp_c2_24064	6	54713088	0	1	Bell-like homeodomain protein 2	No Hit
4	Kinship+Structure	0.80	2-dom-ref	4.61	4.48	0.42	snp_c1_11907	9	51386555	0	1	ATP binding protein	AT5G35960.1
4	Kinship+Structure	0.71	2-dom-alt	6.58	4.44	0.72	snp_c2_32854	5	1E+07	0	1	Signal transducer	AT1G67900.1
4	Kinship+Structure	0.71	2-dom-alt	4.91	4.44	-0.55	snp_c1_15787	11	6943511	0	1	Pectinesterase 51	AT5G09760.1
4	Kinship+Structure	0.71	2-dom-alt	4.55	4.44	-0.45	snp_c1_6992	6	53677174	0	1	Conserved gene of unknown function	AT1G08760.1
4	Kinship+Structure	0.47	general	12.35	4.72	NA	snp_c1_11246	11	4745919	0	1	Seed maturation protein PM36	AT3G16990.1
4	Kinship+Structure	0.47	general	11.41	4.72	NA	snp_c2_15929	7	43192596	0	1	Conserved gene of unknown function	AT2G04280.1
4	Kinship+Structure	0.47	general	9.64	4.72	NA	snp_c2_49245	8	55651378	0	1	UDP-sulfoquinovose synthase	AT4G33030.1
4	Kinship+Structure	0.47	general	9.64	4.72	NA	snp_c2_41463	8	35922019	0	1	Conserved gene of unknown function	AT2G26110.1
4	Kinship+Structure	0.47	general	9.27	4.72	NA	snp_c1_5895	2	48564909	0	1	Acetyl-CoA synthetase	AT5G36880.2
4	Kinship+Structure	0.47	general	9.18	4.72	NA	snp_c1_12386	2	37411847	0	1	Zinc finger protein	AT1G30970.1
4	Kinship+Structure	0.47	general	8.51	4.72	NA	snp_c1_821	8	374494529	0	1	Conserved gene of unknown function	AT4G32820.1
4	Kinship+Structure	0.47	general	7.91	4.72	NA	snp_c2_2998	9	58689196	0	1	Gene of unknown function	No Hit
4	Kinship+Structure	0.47	general	6.37	4.72	NA	snp_c2_12601	7	53106007	0	1	3-hydroxybutyryl-CoA dehydrogenase	AT3G15290.1
4	Kinship+Structure	0.47	general	5.75	4.72	NA	snp_c2_32854	5	1E+07	0	1	Signal transducer	AT1G67900.1
4	Kinship+Structure	0.47	general	5.57	4.72	NA	snp_c2_12404	7	52544646	0	1	Conserved gene of unknown function	AT1G78810.1
4	Kinship+Structure	0.47	general	5.1	4.72	NA	snp_c1_6869	3	41992129	0	1	CDPK	AT1G12580.1
4	Kinship+Structure	0.47	general	4.75	4.72	NA	snp_c2_3512	5	51477749	0	1	Conserved gene of unknown function	AT3G04350.1

# Proceso 6: Replicación



- El estudio se puede replicar independientemente?
  - ▶ Individuos no incluidos en el estudio
  - ▶ Software distinto

Genotipo	
Tetra	Diplo
AAAA	AA
AAAB	AB
AABB	
ABBB	
BBBB	BB

- El pipeline realiza la replicación mediante la transformación del genotipo tetraploide a diploide
- Nueva ejecución del estudio pero ahora para diploides
  - ▶ Librería GWASpoly (Interno)
  - ▶ *Software PLINK (Externo)*

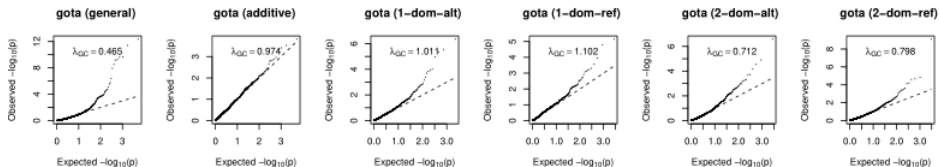
# Resultados preliminares GWAS sobre papa tetraploide: Datos

- Datos involucrados en el estudio:
  - ▶ Datos de 605 muestras o accesiones de papa del grupo Andigena del CCC de Agrosavia
  - ▶ Datos del genotipo de 4700 SNPs (Agrosavia)
  - ▶ Valores del fenotipo de resistencia a gota (LateBlight) de las 605 muestras (Agrosavia)
  - ▶ Datos de estructura poblacional del trabajo de Berdugo et al 2017<sup>1</sup>.
  - ▶ Anotaciones de 8300 SNPs de un 8K SNPArray tomados de la iniciativa SOLCAPs ([http://solcap.msu.edu/potato\\_infinium.shtml](http://solcap.msu.edu/potato_infinium.shtml)).

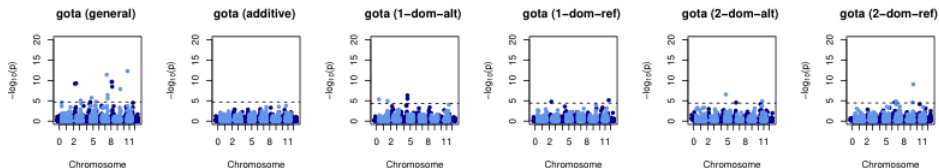
<sup>1</sup>. Berdugo-Cely J, Valbuena RI, Sánchez-Betancourt E, Barrero LS, Yockteng R. Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. Li X-Q, ed. PLoS One. 2017;12(3):e0173039. doi:10.1371/journal.pone.0173039



# Resultados preliminares GWAS sobre papa tetraploide: Plots



## GWAS 4-ploidy with Kinship+Structure for gota trait



# Resultados preliminares GWAS sobre papa tetraploide: Tabla de Asociaciones

Ploidy	Type			GC		Model	SNPs	Score	Threshold	Chrom	Position
4	Kinship+Structure	1.011	1-dom-alt	6.35	4.35	-0.78	snp c1 3484	4	71945107	Cinnamoyl-CoA reductase	AT2G33590.1
4	Kinship+Structure	1.011	1-dom-alt	5.55	4.35	-0.69	snp c2 10568	4	71954528	3-hydroxy-3-methylglutaryl coenzyme A synthase	AT1G76490.1
4	Kinship+Structure	1.011	1-dom-alt	5.35	4.35	-0.59	snp c2 36664	1	535454	Adagio protein 3	AT1G68050.1
4	Kinship+Structure	1.011	1-dom-alt	4.98	4.35	-0.59	snp c2 4875	1	83177518	Conserved gene of unknown function	AT5G65810.1
4	Kinship+Structure	1.102	1-dom-ref	5.14	4.44	0.57	snp c2 40748	12	12654160	Pentatricopeptide repeat-containing protein	AT2G37230.1
4	Kinship+Structure	1.102	1-dom-ref	4.79	4.44	0.64	snp c1 7325	2	42764457	POM30	AT5G67500.1
4	Kinship+Structure	0.798	2-dom-ref	9.1	4.48	-0.76	snp c2 2998	9	58689196	Gene of unknown function	No Hit
4	Kinship+Structure	0.798	2-dom-ref	4.84	4.48	0.55	snp c1 10855	7	9998099	DNA repair protein RAD51 homolog	AT5G20850.1
4	Kinship+Structure	0.798	2-dom-ref	4.65	4.48	-0.44	snp c2 24064	6	54713088	Bell-like homeodomain protein 2	No Hit
4	Kinship+Structure	0.798	2-dom-ref	4.61	4.48	0.42	snp c1 11907	9	51386555	ATP binding protein	AT5G35960.1
4	Kinship+Structure	0.712	2-dom-alt	6.58	4.44	0.72	snp c2 32854	5	10418984	Signal transducer	AT1G67900.1
4	Kinship+Structure	0.712	2-dom-alt	4.91	4.44	-0.55	snp c1 15787	11	6943511	Pectinesterase 51	AT5G09760.1
4	Kinship+Structure	0.712	2-dom-alt	4.55	4.44	-0.45	snp c1 6992	6	53677174	Conserved gene of unknown function	AT1G08760.1
2	Kinship+Structure	0.683	1-dom-alt	6.8	4.42	-0.9	snp c1 3484	4	71945107	Cinnamoyl-CoA reductase	AT2G33590.1
2	Kinship+Structure	0.683	1-dom-alt	5.72	4.42	-0.76	snp c2 10568	4	71954528	3-hydroxy-3-methylglutaryl coenzyme A synthase	AT1G76490.1
2	Kinship+Structure	0.683	1-dom-alt	4.93	4.42	-0.63	snp c2 4875	1	83177518	Conserved gene of unknown function	AT5G65810.1
2	Kinship+Structure	0.683	1-dom-alt	4.61	4.42	-0.56	snp c2 36664	1	535454	Adagio protein 3	AT1G68050.1
2	Kinship+Structure	0.683	1-dom-alt	4.48	4.42	-0.52	snp c2 49495	2	37017928	Protease C56	AT3G02720.1
2	Kinship+Structure	0.666	1-dom-ref	5.31	4.53	0.6	snp c2 40748	12	12654160	Pentatricopeptide repeat-containing protein	AT2G37230.1
2	Kinship+Structure	0.635	additive	5.14	4.79	-0.47	snp c2 12601	7	53106007	3-hydroxybutyryl-CoA dehydrogenase	AT3G15290.1
2	Kinship+Structure	0.635	additive	4.79	4.79	0.54	snp c2 49245	8	35651378	UDP-sulfoquinovose synthase	AT4G33030.1

- Lenguaje R
- Librería de R GWASpoly
- Interacción modo línea de comandos
- Integración Software Plink y Tassel

- Solo un software para tetraploides
- Cambio de tetra a diplo se pierde información
- Resultados con los otros softwares no son los mismos
- No se como comprobar si lo que encuentro es lo correcto

Gracias!