UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Introducción al software TASSEL para el análisis masivo de datos de genotipado

Apellidos, nombre	Arrones Olmo, Andrea (anarol1@etsiamn.upv.es) Vilanova Navarro, Santiago (sanvina@upvnet.upv.es) Gramazio, Pietro (piegra@upv.es) Prohens Tomás, Jaime (jprohens@btc.upv.es) Plazas Ávila, Mariola (maplaav@btc.upv.es)
Departamento	Instituto de Conservación y Mejora de la Agrodiversidad Valenciana/Biotecnología
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

Un paso esencial en muchos proyectos de investigación genética es el genotipado de las muestras humanas, animales, vegetales o microbianas que se están estudiando. Las tecnologías de genotipado se basan en la identificación de diferencias entre las secuencias genómicas que puedan dar lugar a cambios importantes en el fenotipo. Como resultado, las plataformas de genotipado de alto rendimiento generan una enorme cantidad de información, para cuyo correcto análisis se requieren algunos fundamentos de bioinformática y Big Data. Existen diferentes programas informáticos para el análisis masivo de datos brutos, pero a menudo se basan en líneas de comandos, que distan mucho de ser intuitivas. En este artículo vamos a utilizar el software Trait Analysis by aSSociation, Evolution and Linkage (TASSEL), ya que es una herramienta muy útil y fácil de usar que permite un análisis visual de las secuencias genotipadas.

2 Objetivos

Una vez que el alumnado se lea con detenimiento este documento, será capaz de:

- Analizar los datos brutos que se reciben de un servicio de genotipado.
- Encontrar el sistema de filtrado e imputación que más se ajusta a las muestras de partida.
- Identificar la existencia de agrupaciones de muestras mediante un análisis de componentes principales (PCA).
- Realizar un estudio de asociación de genoma completo (GWAS) para identificar regiones genómicas asociadas a caracteres de interés.
- Entender el proceso de identificación de regiones o genes candidatos responsables de un fenotipo concreto.

3 Introducción

El genotipado es una tarea rutinaria en el campo de la genética que se realiza con el fin de conocer la información genética de una muestra de partida. Las plataformas de genotipado generan unos archivos de datos con una representación parcial del genoma, es decir, únicamente devuelven algunos nucleótidos (adenina, timina, citosina o guanina) de la secuencia completa de ADN. Estas posiciones se denominan **polimorfismos de nucleótido único (SNPs)**. Los SNPs son variaciones en la secuencia de ADN que ocurren cuando se altera un solo nucleótido de la secuencia del genoma (Imagen 1). Estos SNPs permiten diferenciar individuos entre sí, ya que pueden presentar distintos nucleótidos o SNPs en una posición concreta. A cada posible variante se le denomina alelo.

Imagen 1. Ilustración conceptual de polimorfismos de nucleótido único (SNPs).

Sin embargo, los sistemas de genotipado no tienen una eficacia del 100% y pueden generar datos en algunas posiciones con baja confianza. Es por ello por lo que todas las posiciones o SNPs van asociados a un valor de calidad y un flujo de análisis de secuencias se inicia generalmente con un proceso previo de **filtrado de los SNPs**. También ocurre que, en ocasiones, los sistemas de genotipado no logran determinar ciertas posiciones de una muestra. La **imputación de alelos** es una técnica estadística que permite inferir los genotipos de los SNPs no identificados. Estas herramientas reconstruyen la secuencia en base a una referencia o a las posiciones vecinas más cercanas.

El manejo de softwares específicos para el análisis de las secuencias en bruto es esencial para enfrentarse a un ensayo de genotipado. El software TASSEL ver. 5.0 es muy intuitivo ya que los estadísticos se calculan paso a paso y los resultados se visualizan gráficamente. Además de poder realizar un primer cribado de las secuencias, este software presenta un gran número de funcionalidades. Entre otras, TASSEL integra herramientas para realizar **análisis de componentes principales (PCA)** con el objetivo de observar la distribución de los datos. Un PCA es una proyección de un conjunto de datos en términos de nuevas variables o componentes no correlacionadas que se ordenan por la cantidad de varianza que explican. En esta representación gráfica de los datos es posible identificar la presencia de patrones dentro de una nube de puntos, es decir, permite distinguir grupos separados entre los individuos de estudio. Sin embargo, la funcionalidad principal por la que TASSEL es ampliamente conocido es por ser utilizado para realizar **estudios de asociación de genoma completo (GWAS)**. El GWAS combina datos genotípicos y fenotípicos para identificar variantes genómicas asociadas estadísticamente a un carácter. El algoritmo busca SNPs a lo largo del genoma que presenten una diferencia constante entre dos fenotipos extremos. Una vez se asocia una región genómica a un carácter de interés, se puede profundizar en la búsqueda de genes responsables de dicho carácter dentro de la región candidata.

4 Desarrollo

Los proyectos de genotipado masivo están desafiando los métodos existentes de análisis de datos brutos. Debido a la gran cantidad de datos generados, el análisis se realiza generalmente en línea de comandos ya que los programas de interfaz gráfica no son capaces de soportar grandes cantidades de información. En este sentido, el programa TASSEL ver. 5.0 implementa un software mejorado que acepta archivos pesados y ofrece diversas posibilidades de manipulación de datos y visualización de resultados. Además, es

un software libre (https://tassel.bitbucket.io/) que puede descargarse en cualquier ordenador.

En este artículo docente se proporcionan los conocimientos necesarios para poder entender cómo llevar a cabo el análisis de datos procedentes de un proyecto de genotipado masivo utilizando el software TASSEL. El contenido se ha estructurado para poder reproducir paso a paso cada uno de los análisis imprescindibles de un proyecto de genotipado.

4.1 Filtrado e imputación de SNPs

Para este primer apartado, los estudiantes tendrán a su disposición un archivo en formato ".vcf" que incluye todas las posiciones de SNPs genotipadas de una colección de individuos y un archivo ".txt" con sus respectivos fenotipos para diferentes caracteres de interés. En caso de que se disponga de un proyecto de genotipado propio, hay diferentes bases de datos donde se pueden conseguir estos datos de forma gratuita. Una de las bases de datos más grande y completa es la del GenBank (https://www.ncbi.nlm.nih.gov/genbank/).

El primer paso es descargar y abrir el archivo ".vcf" en TASSEL para visualizar las secuencias. A continuación, las muestras se someterán a un proceso de filtrado de los SNPs y a la imputación de los datos faltantes (Imagen 2). El programa TASSEL es muy intuitivo y presenta dos pestañas en la parte superior izquierda, una de filtrado [*Filter*] y una de imputación [*Impute*]. Existen diferentes técnicas de filtrado en función de la frecuencia alélica o las posiciones heterocigotas, entre otras, y diferentes métodos de imputación como "LD KNNi", "FILLIN" o "FSFHap". El objetivo de este apartado es manipular las secuencias y probar distintas combinaciones hasta identifica el método que más se ajusta a los datos.

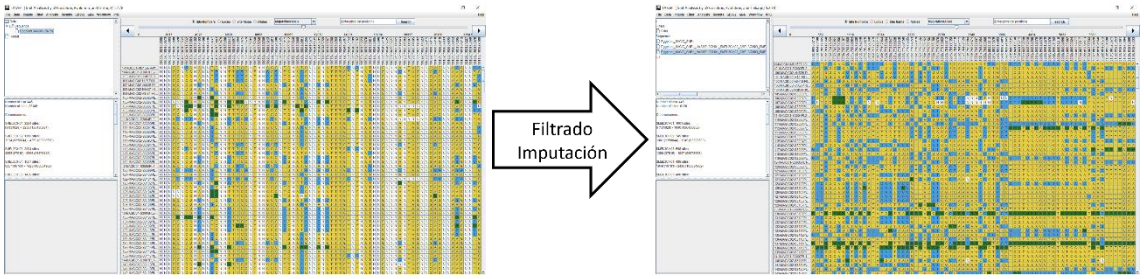

Imagen 2. Diferencias entre los SNPs o posiciones genotipadas antes (izquierda) y después (derecha) del filtrado y la imputación.

4.2 Identificación de grupos en los datos de partida

Una vez se disponga de las secuencias finales, se pueden realizar distintos análisis para identificar con qué tipo de secuencias se está trabajando. Un primer análisis consiste en realizar un PCA para observar la distribución de los datos. Este análisis es importante cuando se trabaja con poblaciones para determinar la presencia o ausencia de estructura poblacional, ya que su existencia podría interferir en los resultados finales.

En primer lugar, es necesario construir una matriz de distancias [*Analysis > Relatedness > Distance Matrix*] y realizar un escalado multidimensional (MDS) para obtener los valores de las componentes principales (PCs) [*Analysis > Relatedness > MDS*]. Una vez generadas

ambas tablas, se puede estudiar el porcentaje de variabilidad explicado por cada una de las PCs. A continuación, se puede representar gráficamente los resultados en un diagrama de dispersión XY combinando diferentes PCs para identificar patrones o agrupaciones en la nube de puntos [*Results > Chart > XYScatter > X: PC1, Y1: PC2*] (Imagen 3).

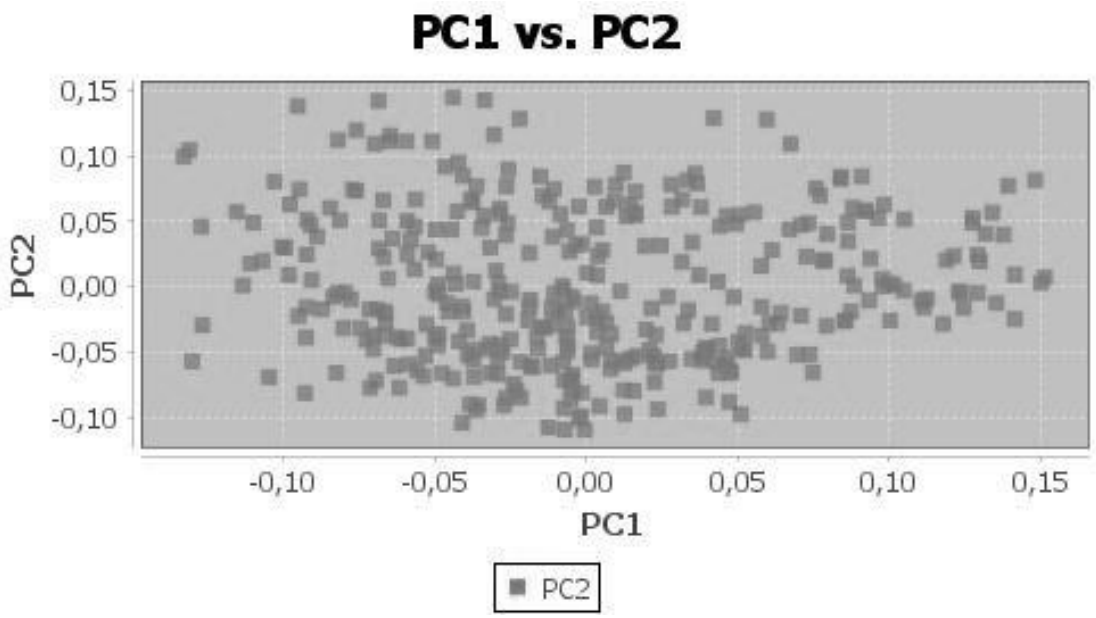

Imagen 3. Representación gráfica de un análisis de componentes principales (PCA) que muestra el diagrama de dispersión de la primera y segunda componentes principales (PCs).

4.3 Búsqueda de regiones genómicas asociadas a un carácter de interés

El segundo análisis que realizar consiste en identificar regiones genómicas asociadas a un carácter de interés mediante un GWAS. Es en este apartado es cuando se utiliza el archivo ".txt" de fenotipado. Combinando la información de genotipado, fenotipado y MDS, se analizan los datos por medio del modelo lineal general (GLM) de asociación [*Analysis > Association > GLM*]. Los resultados generados se pueden representar gráficamente en un diagrama Manhattan para cada carácter de interés [*Results > Manhattan Plot*] (Imagen 4).

En este tipo de representación gráfica, las coordenadas genómicas se muestran a lo largo del eje x. En el eje y, se muestra el logaritmo negativo del valor *p* de asociación para cada uno de los SNPs. Por lo tanto, cada punto del gráfico es equivalente a cada uno de los SNPs estudiados y distribuidos a lo largo del genoma. Los diferentes colores de cada bloque suelen mostrar la extensión de los cromosomas. Dado que las asociaciones más fuertes tienen los valore *p* más pequeños, sus logaritmos negativos son mayores y se muestran como picos más altos.

Este gráfico permite comprender cómo se relacionan regiones genómicas con caracteres específicos a través de la identificación de posiciones de SNPs significativas. Para ello, se establecen valores de corte o umbrales mediante la corrección de pruebas múltiples de Bonferroni y la tasa de descubrimientos falsos (FDR). Los SNPs con logaritmo negativo del valor *p* por encima de estos umbrales se declararán significativamente asociados a un carácter.

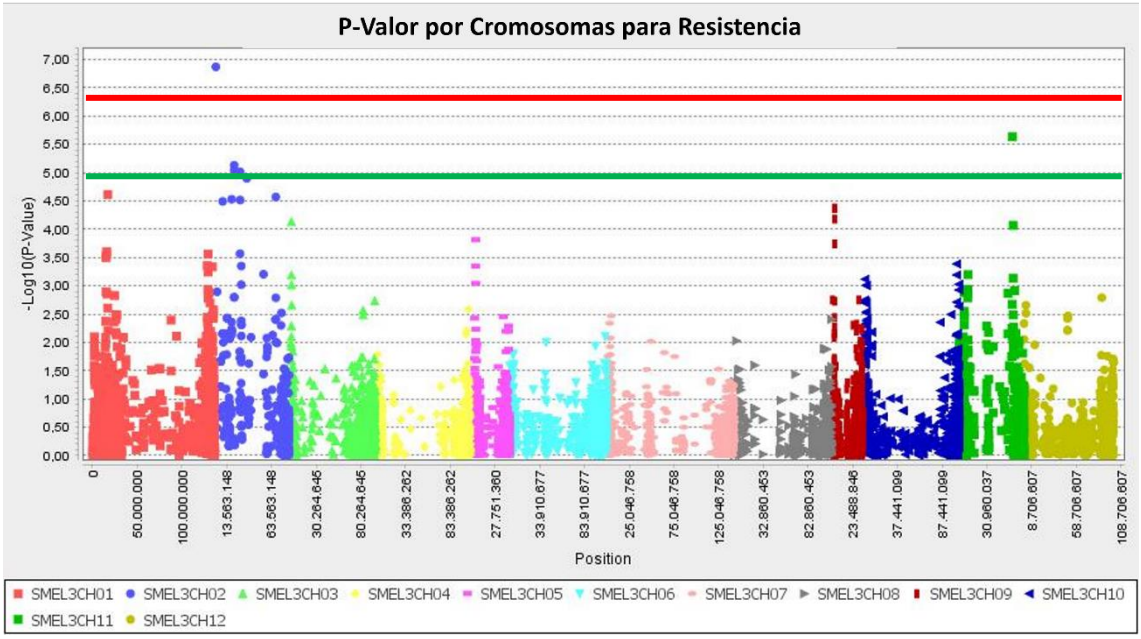

Imagen 4. Ejemplo de un diagrama Manhattan para un carácter concreto. Las líneas horizontales roja y verde representan los umbrales de significación de Bonferroni y la tasa de descubrimientos falsos (FDR), respectivamente.

5 Cierre

A lo largo de este objeto de aprendizaje hemos introducido diferentes conceptos teóricos que combinados con fundamentos bioinformáticos y Big Data proporcionan a los alumnos las competencias necesarias para el análisis de datos brutos de genotipado con ayuda de un software específico. Se ha elegido el software TASSEL por ser fácil de usar y por mostrar los resultados gráficamente, lo cual es crucial para una mejor comprensión y refuerzo del conocimiento adquirido. Este software tiene numerosas funcionalidades, por lo que sería posible utilizarlo para otros fines o para el análisis de muestras de distintos orígenes.

6 Bibliografía

Benjamini, Y.; Hochberg, Y.: "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal statistical society: series B (Methodological), 1995, 57(1), pág. 289–300.

Bradbury, P. J.; Zhang, Z.; Kroon, D. E.; Casstevens, T. M.; Ramdoss, Y.; Buckler, E. S.: "TASSEL: software for association mapping of complex traits in diverse samples", Bioinformatics, 2007, 23(19), pág. 2633–2635.

Chan, A. W.; Hamblin, M. T.; Jannink, J.-L.: "Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data", PloS one, 2016, 11(8), e0160733.

Hu, Z. L.; Ying, Y. L.; Huo, M. Z.; Kong, X. F.; Yu, X. D.; Zhang, J. R.; Long, Y. T.: "A Course of Hands-On Nanopore Experiments for Undergraduates: Single-Molecule Detection with Portable Electrochemical Instruments", Journal of Chemical Education, 2020, 97(12), pág. 4345–4354.