Universidad Miguel Hernández de Elche

Easymap: un programa que facilita la cartografía de mutaciones mediante secuenciación masiva

Samuel Daniel Lup

Tutores:

José Luis Micol Molina

David Wilson Sánchez

Unidad de Genética

Instituto de Bioingeniería

Máster en Biotecnología y Bioingeniería

Curso académico 2015-2016

JOSÉ LUIS MICOL MOLINA, Catedrático de Genética de la Universidad Miguel Hernández de Elche, y

DAVID WILSON SÁNCHEZ, Contratado posdoctoral de Genética de la Universidad Miguel Hernández de Elche,

HACEMOS CONSTAR:

Que el presente trabajo ha sido realizado bajo nuestra dirección y recoge fielmente la labor realizada por Samuel Daniel Lup como Trabajo de Fin del Máster en Biotecnología y Bioingeniería. Las investigaciones reflejadas en esta memoria se han desarrollado íntegramente en la Unidad de Genética del Instituto de Bioingeniería de la Universidad Miguel Hernández de Elche.

David Wilson Sánchez José Luis Micol Molina

Elche, 30 de junio de 2017.

I.- RESUMEN Y PALABRAS CLAVE

Las búsquedas de mutantes basadas en la genética directa han resultado útiles para la identificación de muchos genes y continúan siendo herramientas poderosas para la disección de la función y las interacciones génicas en los organismos modelo. La secuenciación masiva ha revitalizado las tediosas estrategias genéticas para la identificación de mutaciones causantes de un fenotipo de interés. La cartografía mediante secuenciación combina la secuenciación masiva con las estrategias clásicas de análisis de ligamiento y logra la identificación rápida de mutaciones puntuales. Existen programas que analizan los datos de secuenciación masiva y determinan la posición de la mutación que causa un fenotipo de interés. Sin embargo, estos programas son poco amigables, requieren *software* adicional para completar sus análisis o son relativamente caros. Hemos creado Easymap, un programa que automatiza el flujo de trabajo desde los datos brutos hasta las mutaciones candidatas. Easymap ofrece dos tipos de cartografía: de segregantes agrupados para las mutaciones puntuales y de secuencias señalizadas para las inserciones grandes, como las que causan los transposones y el ADN-T. Hemos cartografiado con Easymap mutaciones de Arabidopsis a partir de lecturas de secuenciación masiva simuladas.

Palabras clave: Easymap; cartografía mediante secuenciación; análisis de ligamiento; bioinformática.

Forward genetic screens have identified many genes and continue to be powerful tools for the dissection of gene action and interactions in model species. Moreover, massive sequencing has revitalized the time-consuming genetic approaches to identify the mutation causing a phenotype of interest. Mapping-by-sequencing combines next-generation sequencing with classical mapping strategies and allows rapid identification of point mutations. Programs are available that can analyze whole-genome sequencing data to map the position of the causal mutations for a specific phenotype, but these programs are complicated to install or use, require additional software to perform their analyses, or require the user to purchase expensive licenses. We created the Easymap program, which simplifies the data analysis workflow from raw reads to candidate mutations. Two main workflow types are available: bulked segregant mapping for point mutations, and tagged-sequence mapping for large insertions such as transposons or T-DNAs. We successfully tested Easymap as a tool to identify Arabidopsis mutations from simulated massive sequencing reads.

Keywords: Easymap; mapping-by-sequencing; linkage analysis; bioinformatics.

II.- CONCLUSIONES

En este Trabajo de Fin de Máster hemos desarrollado Easymap, un programa informático para facilitar la cartografía de mutaciones puntuales o insercionales mediante secuenciación masiva, válido para Arabidopsis o cualquier otro organismo modelo. Easymap se ha diseñado para reducir al máximo la intervención del usuario, al que no se le exige estar familiarizado con la bioinformática. El programa cuenta con una interfaz gráfica aún en desarrollo. Su sencillo manejo contrasta con su robustez y flexibilidad, que hemos demostrado analizando datos simulados.

Easymap permite estudiar dos tipos de mutaciones de naturaleza muy distinta: las transiciones GC→AT causadas por el EMS, cuya identificación debe hacerse mediante análisis del ligamiento a marcadores, y las inserciones grandes, que se localizan usando como referencia su propia secuencia. La naturaleza modular de Easymap facilitará la implementación de nuevas funciones en versiones futuras.

La mutagénesis al azar seguida del aislamiento de mutantes y la identificación posterior de los genes mutados ha sido y sigue siendo una estrategia de uso común para la disección genética de procesos biológicos. En consecuencia, creemos que Easymap puede ser muy útil para las comunidades científicas de las especies modelo, cuyos genomas han sido totalmente secuenciados y anotados, como el díptero *Drosophila melanogaster*, el nematodo *Caenorhabditis elegans* y la angiosperma Arabidopsis, entre otros. Un desarrollo previsible de Easymap es la generación de un flujo de trabajo específico para las mutagénesis de segundos sitios (second site mutagenesis), en auge creciente para la búsqueda de interactores genéticos de un gen de interés. La peculiaridad de este tipo de experimentos radica en que se mutageniza una estirpe mutante, razón por la que es necesario identificar dos mutaciones: la preexistente, de posición conocida, y la nueva, que interacciona con la anterior, de posición desconocida.

Durante el desarrollo de Easymap he aprendido cuáles son los distintos abordajes experimentales a la cartografía de las mutaciones que causan un fenotipo de interés. También he consolidado mis conocimientos sobre las tecnologías de secuenciación masiva. Además, me he familiarizado con la programación en Python y Bash, lenguajes esenciales en bioinformática. Por último, he aprendido a utilizar con soltura los programas de dominio público para el alineamiento de lecturas y genotipado de individuos y poblaciones, que son indispensables para el análisis de datos de secuenciación masiva.