Programa de Doctorado en Bioingeniería
Universidad Miguel Hernández de Elche

Desarrollo de herramientas bioinformáticas para cartografía génica mediante secuenciación masiva

Samuel Daniel Lup Haruta

Director de la tesis
José Luis Micol Molina

Elche, 2022

JOSÉ LUIS MICOL MOLINA, Catedrático de Genética de la Universidad Miguel Hernández de Elche (UMH)

HAGO CONSTAR:

Que el presente trabajo ha sido realizado bajo mi dirección y recoge fielmente la labor desarrollada por el Graduado Samuel Daniel Lup Haruta para optar al grado de Doctor. Las investigaciones reflejadas en esta memoria se han desarrollado íntegramente en la Unidad de Genética del Instituto de Bioingeniería de la UMH, según los términos y condiciones definidos en el Plan de Investigación del doctorando, y cumpliendo los objetivos propuestos de forma satisfactoria y lo establecido en el Código de Buenas Prácticas de la UMH.

José Luis Micol Molina

Elche, 11 de julio de 2022

II.- RESUMEN

Los escrutinios basados en la genética directa fueron y siguen siendo poderosas herramientas para la identificación de genes y la disección de su actividad e interacciones, tanto en Arabidopsis como en otras especies modelo vegetales y animales. La secuenciación masiva de ADN ha revitalizado los abordajes genéticos clásicos a la identificación de las mutaciones que causan un fenotipo de interés, que se basaban en el lento y tedioso análisis iterativo del ligamiento a un número relativamente bajo de marcadores moleculares. La cartografía mediante secuenciación (mapping-by-sequencing) combina la secuenciación masiva con el análisis de ligamiento clásico para identificar rápidamente mutaciones puntuales. Al igual que en los análisis de ligamiento convencionales, en la cartografía mediante secuenciación se requiere una población cartográfica fenotipada. Sin embargo, no es necesario obtener genotipos individuales, ya que los miembros de las poblaciones cartográficas son genotipados en conjunto. En la cartografía mediante secuenciación pueden usarse como marcadores los polimorfismos de un solo nucleótido inducidos por un mutágeno químico, resultando prescindibles los exocruzamientos del mutante a estudio por una estirpe genéticamente distante de su parental silvestre.

Hemos realizado simulaciones con el fin de facilitar el diseño de experimentos de cartografía mediante secuenciación para la identificación de mutaciones puntuales inducidas químicamente. En dichas simulaciones determinamos en primer lugar qué tecnología de secuenciación masiva —de las que producen lecturas cortas— es más adecuada para las regiones del genoma de Arabidopsis ricas en genes, y la profundidad de lectura mínima para detectar polimorfismos eficazmente. Concluimos que las lecturas simples y las pareadas son igualmente aptas, y que con una profundidad de lectura de 40-50× se alcanza el mejor compromiso entre coste y precisión. También hemos simulado experimentos de cartografía mediante secuenciación para establecer los efectos del tamaño de la población cartográfica y la profundidad de lectura sobre la resolución alcanzada: ambos factores la limitan. Hemos realizado exocruzamientos virtuales para evaluar la utilidad de la obtención de máscaras de polimorfismos naturales, que integran las variantes detectadas en distintos mutantes, aislados tras una misma mutagénesis. Estas máscaras son útiles para filtrar los polimorfismos que distinguen la secuencia genómica de referencia de las de otros accesos silvestres. Hemos evaluado la viabilidad de los cruzamientos entre mutantes portadores de mutaciones puntuales, recesivas y no alélicas para obtener una población cartográfica que permitiese identificarlas simultáneamente. También hemos usado simulaciones para poner a

punto un protocolo de cartografía de inserciones de ADN-T o transposones, basado en el desapareamiento de lecturas pareadas, cuya eficacia hemos demostrado incluso con profundidades de lectura muy bajas y agrupando varios mutantes insercionales. Estas simulaciones se han revelado útiles para el diseño de experimentos reales.

La combinación de las tecnologías de secuenciación masiva y las estrategias clásicas de cartografía génica ha facilitado considerablemente el establecimiento de relaciones entre genotipo y fenotipo. Se han desarrollado muchos programas bioinformáticos para extraer información de las lecturas de secuenciación masiva de ADN. Sin embargo, su uso requiere conocimientos de bioinformática y la redacción de rutinas que encadenen varias de estas herramientas de dominio público. Con el propósito de hacer accesible la cartografía mediante secuenciación a investigadores sin conocimientos de bioinformática, hemos desarrollado Easymap, un programa amigable que parte de lecturas sin procesar para identificar mutaciones candidatas con una intervención mínima del usuario.

Hemos desarrollado dos versiones de Easymap, la primera de las cuales incluía flujos de trabajo para cartografiar mutaciones puntuales en poblaciones de segregantes agrupados, así como inserciones grandes de ADN, como el ADN-T y los transposones. También incorporaba un módulo de simulación de experimentos de secuenciación masiva de ADN. Hemos añadido a Easymap v.2 flujos de trabajo para cartografía mediante análisis de densidad de variantes y de secuenciación de loci de caracteres cuantitativos (QTL-seq). Easymap v.2 puede analizar datos derivados de diferentes diseños experimentales —como los exocruzamientos o retrocruzamientos— para la obtención de poblaciones cartográficas, que a su vez pueden ser F_2, M_2 o M_3; acepta como datos de partida secuencias de ADN genómico o complementario y lecturas simples o pareadas, y como muestras de control, archivos en formatos FASTQ o VCF. También puede usar como marcadores tanto las mutaciones espontáneas como las inducidas por un mutágeno químico, analizar variantes sin aplicar algoritmo cartográfico alguno y multiprocesar datos.

Las dos versiones de Easymap pueden instalarse en entornos UNIX, en Windows 10 (mediante la aplicación Ubuntu disponible en la Microsoft Store), en algunas distribuciones de Linux, y en máquinas virtuales que ejecuten un subsistema Linux dentro de cualquier otro sistema operativo, como macOS. Incorporan una guía de instalación rápida, una interfaz web amigable y un manual. Cada ejecución de Easymap rinde un informe cartográfico con tablas y diagramas que facilitan la interpretación de los resultados obtenidos.

Esta Tesis también incluye un protocolo pormenorizado de todas las etapas de una cartografía mediante secuenciación, desde la construcción inicial de una población cartográfica hasta la interpretación de los resultados obtenidos por Easymap.