

Development and application of bioinformatic tools for the representation and analysis of genetic diversity. Sònia Casillas Viladerrams. Universitat Autònoma de Barcelona.

RESUMEN DE TESIS:

La variación genética es la piedra angular de la evolución biológica. La descripción y explicación de las fuerzas que controlan la variación genética dentro y entre poblaciones es el principal objetivo de la genética de poblaciones. La obtención de un número explosivo de secuencias nucleotídicas en distintos genes y especies ha cambiado radicalmente las perspectivas de la genética de poblaciones, transformándola desde una ciencia empírica insuficiente a un esfuerzo interdisciplinario de un gran alcance, donde los aparatos de generación de nuevas secuencias a gran escala se integran con herramientas bioinformáticas para la extracción y gestión de datos, junto a avanzados modelos teóricos y estadísticos para su interpretación.

Esta tesis es un proyecto de bioinformática y genética de poblaciones completo, cuyo objetivo es el estudio de la diversidad genética en las poblaciones, que se ha llevado a cabo en tres pasos secuenciales: (i) el desarrollo de herramientas para la extracción, procesado, filtrado y control de calidad de secuencias nucleotídicas, (ii) la generación de bases de datos de conocimiento a partir de los datos obtenidos en la primera parte y (iii) la puesta a prueba de hipótesis que requieren de datos de varias especies y loci. En la primera parte de la tesis hemos desarrollado PDA (Pipeline Diversity Analysis), una aplicación Web de código abierto que permite la exploración del polimorfismo en grandes conjuntos de secuencias de DNA heterogéneas. Esta herramienta se alimenta de los millones de secuencias haplotípicas de estudios individuales que hay almacenados en las principales bases de datos moleculares y genera datos de genética de poblaciones que pueden ser utilizados para describir patrones de variación nucleotídica en cualquier especie o gen. Todos los datos extraídos y analizados en la primera parte de la tesis son utilizados en la segunda parte para crear un recurso vía Web completo que proporciona colecciones de secuencias polimórficas con sus medidas de diversidad asociadas en el género *Drosophila* (DPDB, *Drosophila* Polymorphism Database). Este recurso ha significado un reto ambicioso que ha permitido poner a prueba la eficiencia del sistema creado en la primera parte.

Finalmente, se incluyen dos estudios que utilizan los módulos de extracción y análisis de datos desarrollados en la primera parte. En el primero, hemos estudiado los patrones de variación genética en secuencias conservadas no codificadoras para inferir selección negativa y positiva en *Drosophila*. En este estudio hemos utilizado datos de re-secuenciación en *D. melanogaster* junto con datos genómicos comparativos en otras especies de *Drosophila* para demostrar que las regiones frías de mutación no pueden explicar estos bloques conservados. Los resultados muestran que las secuencias conservadas no codificadoras se mantienen por la acción de la selección purificadora. El segundo estudio se centra en la evolución codificadora de los genes *Hox*, una clase de factores de transcripción esenciales en el desarrollo temprano que están involucrados en la especificación de las regiones a lo largo del eje anteroposterior del cuerpo. Hemos medido las tasas de divergencia nucleotídica y de fijación de inserciones y deleciones en tres genes *Hox*, y las hemos comparado con las de tres genes derivados de *Hox* y un conjunto de genes no *Hox* para probar la hipótesis que los genes *Hox* evolucionan lentamente. Los resultados muestran que tanto el número de sustituciones no sinónimas como el grado de constreñimiento funcional no son significativamente distintos entre los genes *Hox* y los no *Hox*, y que los genes *Hox* y los derivados de *Hox* contienen significativamente más inserciones y deleciones que los genes no *Hox* en sus secuencias codificadoras. Así, los genes *Hox* evolucionan más rápidamente que otros genes esenciales expresados en el desarrollo temprano, con patrones de expresión complejos o con intrones largos ricos en elementos *cis*-reguladores.

En síntesis, los trabajos presentados en esta tesis cierran un ciclo completo de proyecto bioinformático, incluyendo todos los pasos necesarios desde la extracción de datos hasta la generación de nuevo conocimiento científico. Es más, el resultado de cada paso es la semilla para múltiples posibles estudios en el siguiente paso, y por lo tanto esta tesis tiene muchas aplicaciones para la comunidad científica.