Resumen

A lo largo de las últimas dos décadas, los datos generados por las tecnologías de secuenciación de nueva generación han revolucionado nuestro entendimiento de la biología humana. Es más, nos han permitido desarrollar y mejorar nuestro conocimiento sobre cómo los cambios (variaciones) en el ADN pueden estar relacionados con el riesgo de sufrir determinadas enfermedades.

Actualmente, hay una gran cantidad de datos genómicos disponibles de forma pública, que son consultados con frecuencia por la comunidad científica para extraer conclusiones significativas sobre las asociaciones entre los genes de riesgo y los mecanismos que producen las enfermedades. Sin embargo, el manejo de esta cantidad de datos que crece de forma exponencial se ha convertido en un reto. Los investigadores se ven obligados a sumergirse en un lago de datos muy complejos que están dispersos en más de mil repositorios heterogéneos, representados en múltiples formatos y con diferentes niveles de calidad. Además, cuando se trata de resolver una tarea en concreto sólo una pequeña parte de la gran cantidad de datos disponibles es realmente significativa. Estos son los que nosotros denominamos datos *"inteligentes"*.

El principal objetivo de esta tesis es proponer un enfoque sistemático para el manejo eficiente de datos genómicos inteligentes mediante el uso de técnicas de modelado conceptual y evaluación de calidad de los datos. Este enfoque está dirigido a poblar un sistema de información con datos que sean lo suficientemente accesibles, informativos y útiles para la extracción de conocimiento de valor.