SUPERCOMPUTACIÓN CASTILLA Y LEÓN

# **Metodologías** basadas en **Machine Learning** para el **análisis** de **variaciones genómicas**

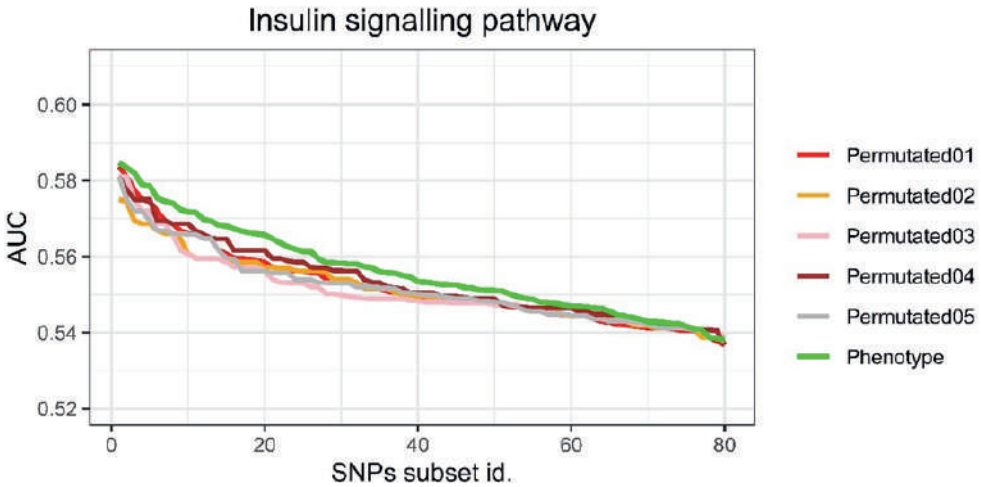
## Objetivo del proyecto

*Un polimorfismo de un solo nucleótido o SNP, es una variación en la secuencia de ADN que afecta a una sola base de la secuencia del genoma. Para que una variación no sea considerada como mutación, debe estar presente en al menos el 1% de la población. Dado que los SNP se heredan de forma muy estable, es posible seguir su evolución de una generación a otra en estudios de poblaciones. Los SNP son de interés para el desarrollo de la medicina personalizada dado que permiten conocer el riesgo de padecimiento de ciertas enfermedades.*

*Los estudios de asociación de genoma completo o GWAS, comparan el ADN de dos grupos de participantes: sujetos con el fenotipo de interés (casos o personas con una enfermedad en particular) frente a sujetos sin el fenotipo (controles). Cada individuo, proporciona una muestra de ADN, de la que se pueden leer millones de variantes genéticas usando los SNP. Así, si un alelo es más frecuente en las personas con la enfermedad, se dice que el SNP se relaciona con dicha enfermedad. El SNP asociado se considera que marca una región del genoma humano que influye en el riesgo del fenotipo. A diferencia de otros métodos, los GWAS analizan el genoma humano por completo. Así, en este tipo de aproximación, no se realiza una búsqueda de relación con un fenotipo sobre un único gen candidato, sino sobre el genoma completo.*

*Los GWAS buscan identificar el alelo de una variante genética que se encuentra de forma más frecuente de lo esperado en los individuos con el fenotipo de interés.*

*El conocimiento de la secuencia de codificación de todos los nucleótidos en un organismo ha permitido a los investigadores estudiar la influencia colectiva de todos los genes de manera simultánea y su papel en las características de los organismos, incluyendo sus enfermedades específicas. El objetivo de esta tesis consiste en la evaluación de los métodos estadísticos y de machine learning que en la actualidad se emplean en los estudios de GWAS, así como el desarrollo de nuevas metodologías, que permitan un análisis automatizado del genoma en los estudios GWAS. Como parte inicial, y con el fin de poseer una herramienta objetiva de evaluación del rendimiento de las metodologías de análisis que se integran en el presente proyecto, se desarrollará un generador de datos sintéticos de genoma. Se propone el uso de técnicas de paralelización en los nuevos algoritmos desarrollados.*




AUC values of the 80 iterations performed for the insulin signaling pathway in the case of cases and controls (phenotype) and five different permutations.

## Periodo de ejecución

Del año **2019** al año **2021.**

## Financiación del proyecto

Tesis Doctoral.

## Participantes del proyecto

Universidad de Oviedo, *www.uniovi.es*

Universidad de León, *www.unileon.es*

SCAYLE, Supercomputación Castilla y León, Spain, *www.scayle.es*

## Funciones de SCAYLE

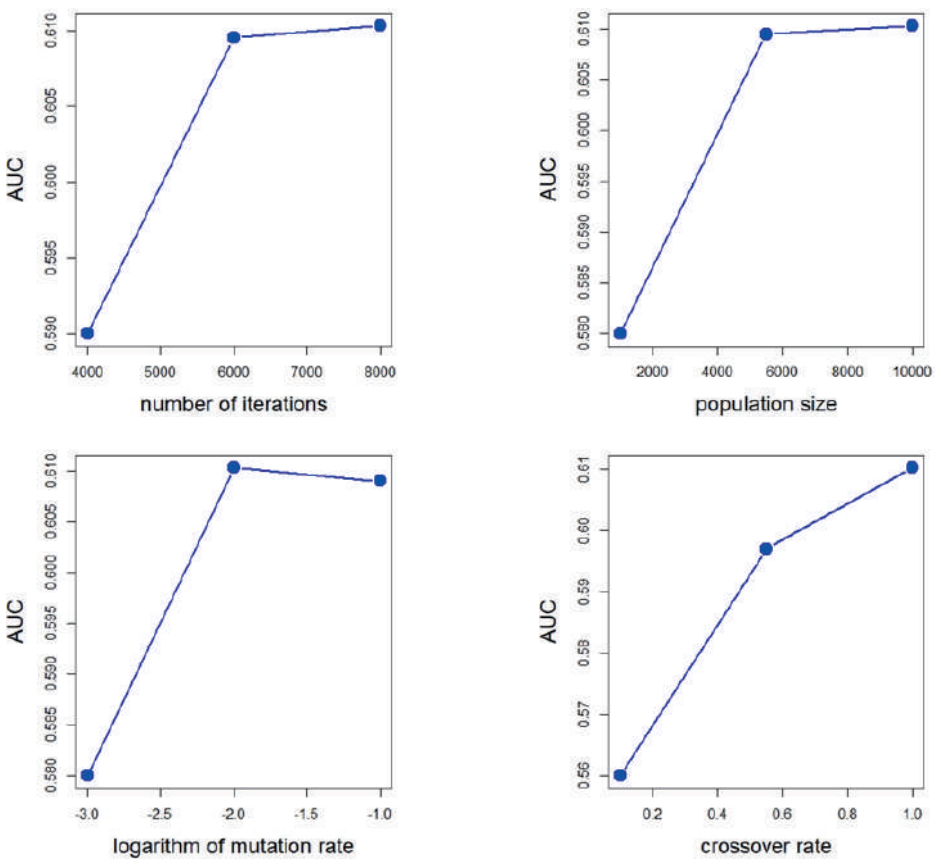
Ejecutar algoritmos de machine learning.

## Líder del proyecto

**La Universidad de León es una universidad pública con sede en la ciudad de León, (España), y con un campus adicional en Ponferrada. Fue fundada en 1979 como escisión de la Universidad de Oviedo, a partir de las diversas Escuelas y Facultades que, dependientes de aquella, existían desde mayor o menor tiempo atrás en la ciudad de León. En la actualidad cuenta con casi 13000 estudiantes**.

universidad de león

RTI2018-093535-B-I00




Main effects plots of: (a) Number of iterations; (b) Population size; (c) Logarithm of mutation rate; (d) Crossover rate.