QTL as a service (QTLaaS), a cloud service for genetic analysis

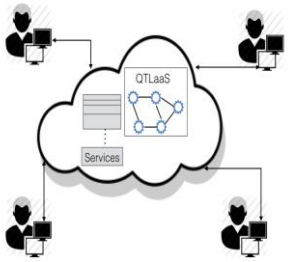
Masters Thesis Project

Understanding the relation between genes and traits is a fundamental problem in genetics. Such knowledge can lead to e.g. the identification of possible drug targets, treatment of heritable diseases, and efficient designs for plant and animal breeding. The aim of quantitative trait loci (QTL) analysis is to locate regions in the genome that can be associated to quantitative traits, i.e. traits where the individuals in a population exhibit continuous distributions.

With a powerful statistical model, efficient numerical algorithms, and a suitable environment for computational experiments, it is then possible to identify the QTL position and the corresponding statistical significance levels. Mathematically, the search for the positions of the QTL corresponds to solving a **multidimensional global optimization** problem where the objective function is the statistical model fit. Distributed Computing Applications (DCA)[1] research group at Department Information Technology, Uppsala University is working on computational and numerical aspects of QTL application.

Standard tools for multiple QTL analysis, using standard computational algorithms, are not able to cope with the massive computations needed. Significant development, both in terms of algorithms and implementations, is needed to provide accurate and efficient tools for simultaneous search of multiple interacting QTL. Geneticists have generally not been main users of high end computing resources. While this is changing to some extent, many groups still do not own or have access to their own computational resources. Furthermore, the need for such resources for QTL analysis is intermittent in nature. Analysis and finding a proper model will only be relevant during a specific phase of the execution of a study. We therefore propose that **cloud computing** is ideal for this user group. Cloud Computing services in our case are:

- **Cloud platform:** We use a cloud infrastructure (OpenStack) for multidimensional QTL scan.
- **R software:** We use R statistical software to interact with the user. R is familiar environment to many geneticists. The power of R compared to more specialized environments is the ability for the end user to freely adapt and extend the methods and workflows.
- **Spark framework:** We use Spark to create the computational infrastructure inside the cloud.



We have developed a framework for analysis of multiple QTL based on a novel, highly efficient search algorithm PruneDIRECT. From the R program, cloud computational resources are accessed using RSpark and used in a transparent way for performing the demanding computations. However, using modern software and services, we now show that this type of tool can be implemented in a portable and easy-to-use way, allowing even inexperienced users of larger-scale computational resources to extend and modify their QTL analysis toolbox.

The goal of this project is to develop **QTL as a service** (QaaS) based on the underlying computational resources for QTL search.

Contact information: Salman Toor(salman.toor@it.uu.se) , Behrang Mahjani(behrang.mahjan@it.uu.se)

[1] https://www.it.uu.se/research/group/dca