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

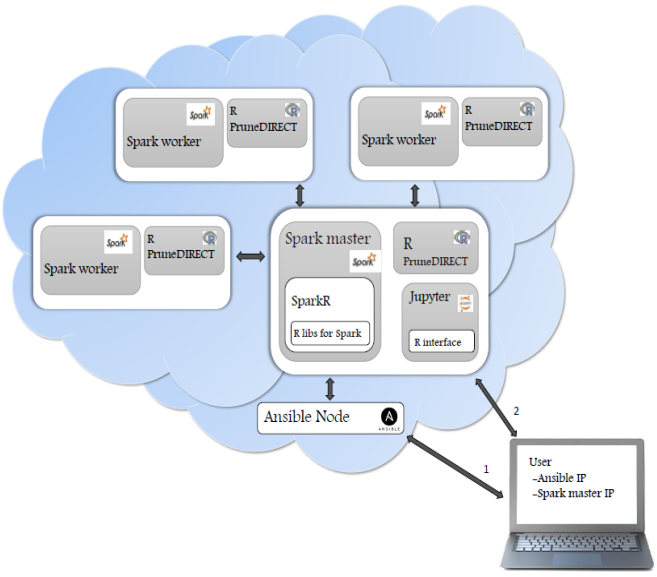
Understanding the relationship between genes and traits is a fundamental problem in genetics. Such knowledge can lead to e.g. the identification of possible drug targets, treatment of heritable diseases, and efficient designs for plant and animal breeding. The aim of quantitative trait loci (QTL) analysis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

With a robust 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.

Standard tools for multiple QTL analysis, using standard computational algorithms, are not able to cope with the massive computations needed. A significant development, both regarding algorithms and implementations, is needed to provide accurate and efficient tools for the simultaneous search of multiple interacting QTL. Geneticists were generally not the 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 a familiar environment to many geneticists. The power of R compared to more specialized environments is the ability for the end user to adapt and extend the methods and workflows freely.
- **Spark framework:** We use Spark to create the computational infrastructure in the cloud.

We have developed a framework for the analysis of multiple QTL based on a novel, highly efficient search algorithm PruneDIRECT. From the Jupyter notebook, cloud resources are accessed using IRKernal supported by the Spark framework. The platform allows transparent access to the underlying computational resources while working in the analysis friendly R software environment.



The available QTL service contextualizes the platform with semi-automatic approach. Application Experts have to start all the VMs manually, inject the IP addresses in the “Ansible-variable” file and then start the installation process. Due to the static configuration settings, the platform lacks essential features like scalability and elasticity. The goal of this project is to extend the already available QTLaaS platform with the following features:

1. Scalable **QTL as a Service** (QTLaaS) based on the underlying computational resources for QTL search (Horizontal scalability on request).
2. REST API/interface for the end-users to create, resize and decommission the QTL platform.
3. Efficient deployment scheme (Current implementation takes more than 20 minutes to a fresh deployment)
4. Flexible interfaces to inject new data files.

Further instructions about how to install QTLaaS can be found here: https://github.com/QTLaaS/QTLaaS