**Project1: Grid-Computing**

The early years of the 21st century have seen the rise of both "data science" and "big data" as terms associated with scientific enquiry. The field of bioinformatics in particular has seen a "data deluge" as dataset sizes move from megabytes to gigabytes to terabytes. As a result, data management forms a key challenge for the modern science lab. One part of data management is securely and efficiently moving data between scientific collaborators at different institutions.

In the past data has been exchanged largely by two mechanisms: network file transfer and physical media. For network file transfer FTP is widely used, with the limitation that this relies on authentication from local databases of users that require remote users to be added on an ad-hoc basis. FTP also does not make optimal use of new high speed data networks, such as the 10Gb SANReN network available to many South African universities. For large datasets, physical transfer is sometimes used, with data being shipped between collaborators on hard disks. This involves high cost and as such is seldom a viable often for datasets that need to be updated regularly.

More recently collaborators have taken to using cloud services such as Google Drive and Dropbox to share data, adding the cost of moving data to the cloud to the cost of moving data.

This project proposes to research alternatives to existing data exchange mechanisms, evaluating Grid based technologies as an alternative to traditional file transfer. Grid technologies rely on cross-institutional authentication frameworks to enable sharing resources between collaborators and do the actual data transfer using GridFTP, a file transfer tool optimised for high throughput on high speed networks. The project will focus on two aspects:

1) The costs and challenges involved in implementing cross-site authentication using a variety of Grid technologies.

2) The performance of the following data transfer options: FTP, SSH, HPN-SSH, GridFTP.

The ultimate goal is to gain a better understand of the options available to solve the data exchange problem facing a real world bioinformatics lab. To this end the project will be undertaken together with SANBI, the South African National Bioinformatics Institute, who will provide sample data and computational and network infrastructure to implement the project.