End to end genomics acquisition and analysis platform

Ben Lynch, Jeff McDonald, and J. Vinals. Minnesota Supercomputing Institute, University of Minnesota.

We have been developing processes and protocols that integrate data capture of all genomics and proteomics data generated on campus (about 1-2 TB/week), a custom QA/QC pipeline, the delivery of the data to principal investigators via defined environments (Galaxy, CLC Genomics Workbench, and PacBio), and the availability of a large complement of analytic tools for genomics, proteomics and metabolomics. This work is currently funded by NSF. The main difficulty is in scaling up the user community while leveraging frameworks that have been mostly developed having small platforms in mind (desktops, small clusters). Therefore they do not integrate well in a shared, scheduled environment as is typical of an HPC installation. Furthermore, more and more of the high end analytic tools require either significant computational resources, or nodes with sufficiently large memories (TBs).

Our current user base in Minnesota comprises approximately 150 principal investigators (plus their students, postdocs and collaborators) in a variety of disciplines: Biology including Plant Biology, our Medical School and affiliated hospitals, Veterinary Medicine, Public Health and other informatics efforts in Science and Engineering departments. We also operate a clinical pipeline for the Pathology department of the University hospital. In the aggregate, we maintain a wide range of analytic tools, and have contributed to the development of Galaxy and PacBio to accommodate workflows that require large scale computing and/or fast I/O. Common challenges include integration with external authentication systems leading to difficulties in resource allocation, file ownership, or interaction with schedulers (e.g. PBS across our supercomputers).

The experiment that we propose is to deploy the end to end pipeline to a virtualized, remote, environment provided by the NSF cloud in order to demonstrate scaling with a large number of users, including the acquisition phase of Next Generation Sequencing (NGS) or Single Molecule Real Time (SMRT) data, the provision of customizable analytic environments, and the availability of genomics and proteomics tools in those environments, all done remotely. It is our experience that commercial cloud service providers can accommodate the smaller jobs. However, it would greatly benefit the informatics community to have access to providers whose primary goal is scientific in nature, with an understanding of the tools and methods that are of interest to a given discipline, and with access to significant resources in large scale computation, memory, and I/O characteristics.