Bioinformatics Workflow and Applications Using the PoPLAR Gateway

Mary Lauren Harris (Baylor University), Catherine Eason (Wofford College), Bhanu Rekepalli (NICS)

Introduction

Thanks to advances made by Next Generation Sequencing (NGS) technology, a large volume of genetic data is now accessible for analysis. From raw data to published results, an efficient and automated pipeline for the analysis of genetic data will revolutionize modern research. Individual programs can be optimized and placed in a science gateway for researchers to customize their pipelines. In particular, the addition of reliable high performance computing (HPC) programs to the PoPLAR Gateway opens the doors to computational abilities even for scientists with little to no programming experience or access to their own HPC resources.

Materials

We are currently using the Beacon super-computer. The 256GB of memory per node allows for tests with large databases.

The PoPLAR Gateway is under development. This is an example screenshot of the login screen.

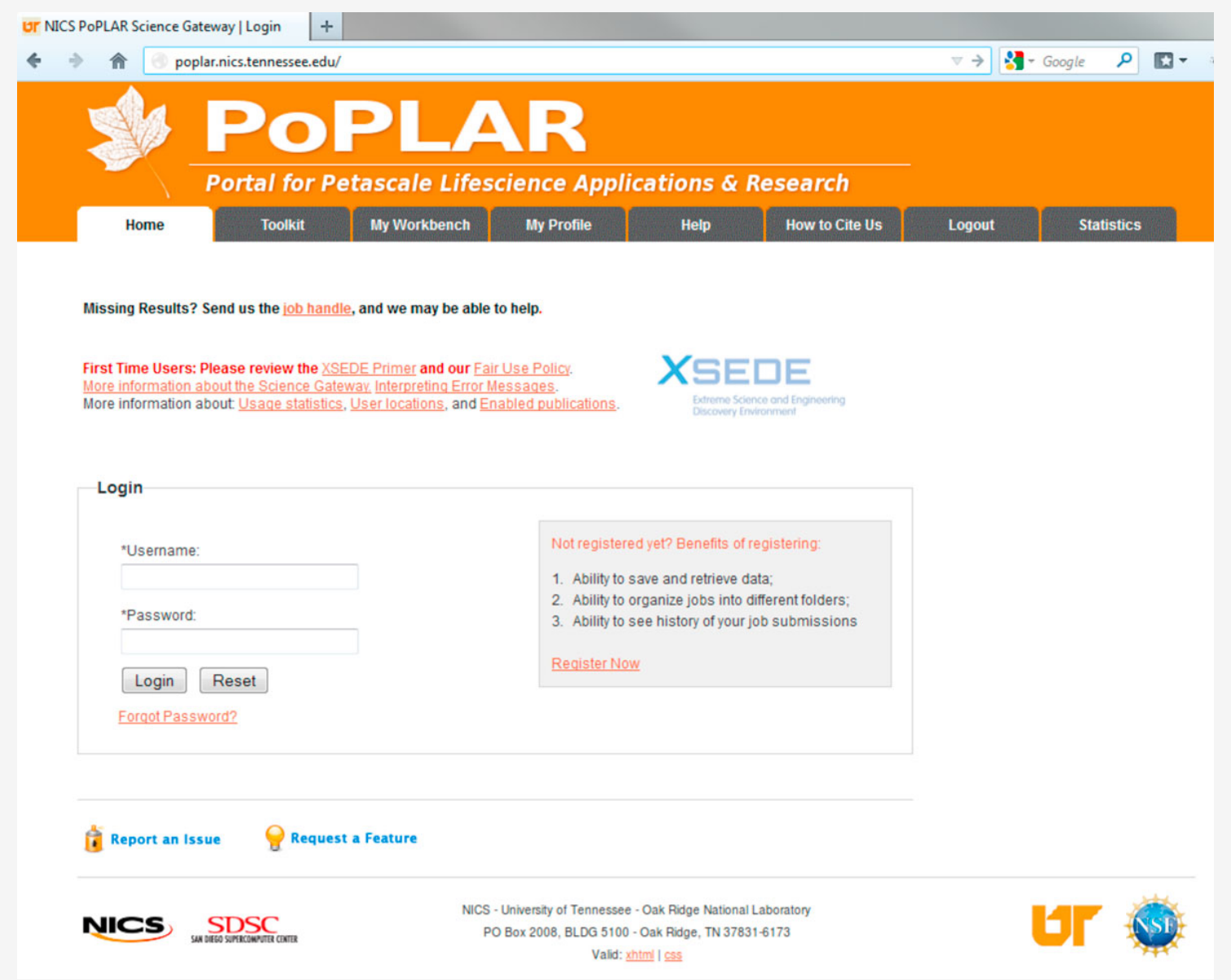

Figure 1. Screenshot of PoPLAR login screen.

Workflow

Experimental Data
Reference Sequences

Sequence Data:

Typically collected in short reads anywhere from 20 to 1000 base pairs. The length and accuracy of the read depends on the method used to obtain the sequence.

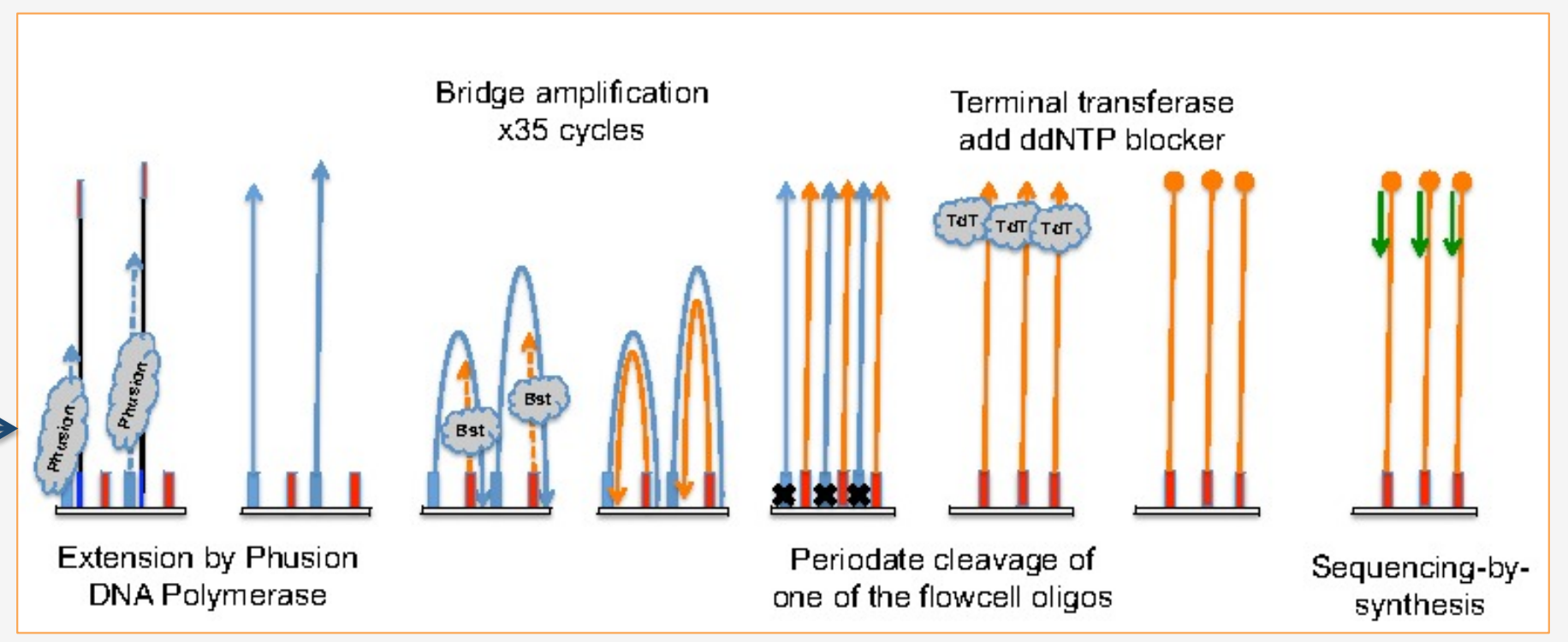

Figure 2. Image describing Illumina sequencing.

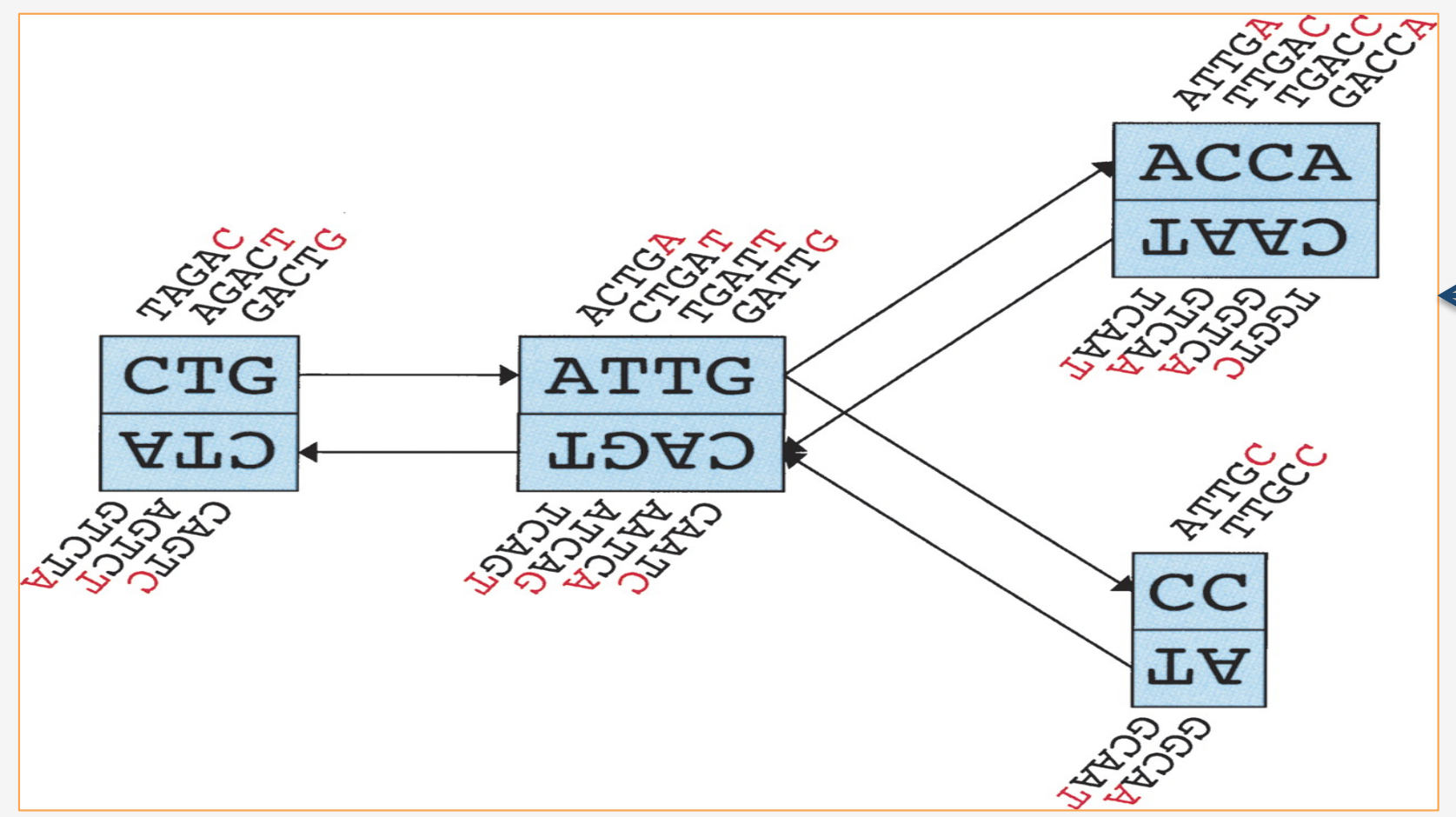

Figure 3. Example of de Bruijn method from the Velvet assembler.

Sequence Assembly:

Joins the short reads together into full sequences to be studied as a whole. Contiguous sequences produced by this process are referred to as contigs. It is difficult to quantify the accuracy of sequence assembly especially with a large number of genomic repeats or within a novel genome.

Analysis:

Sequence Analysis may include information such as how similar sets of data are to one another or identifying specific genetic features of one data set. Analysis is directed toward the desired outcome of the study.

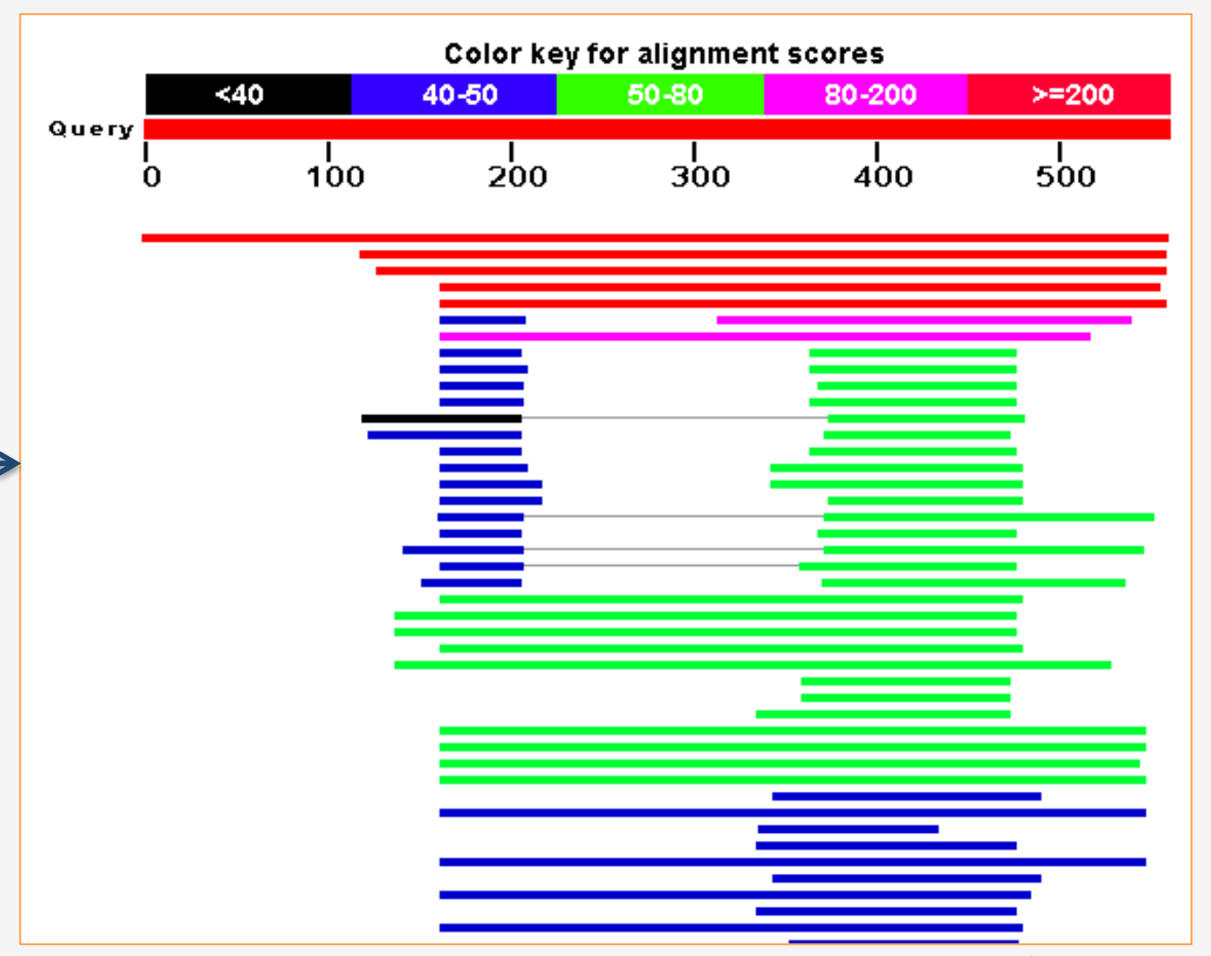

Figure 4. Example alignment for yeast.

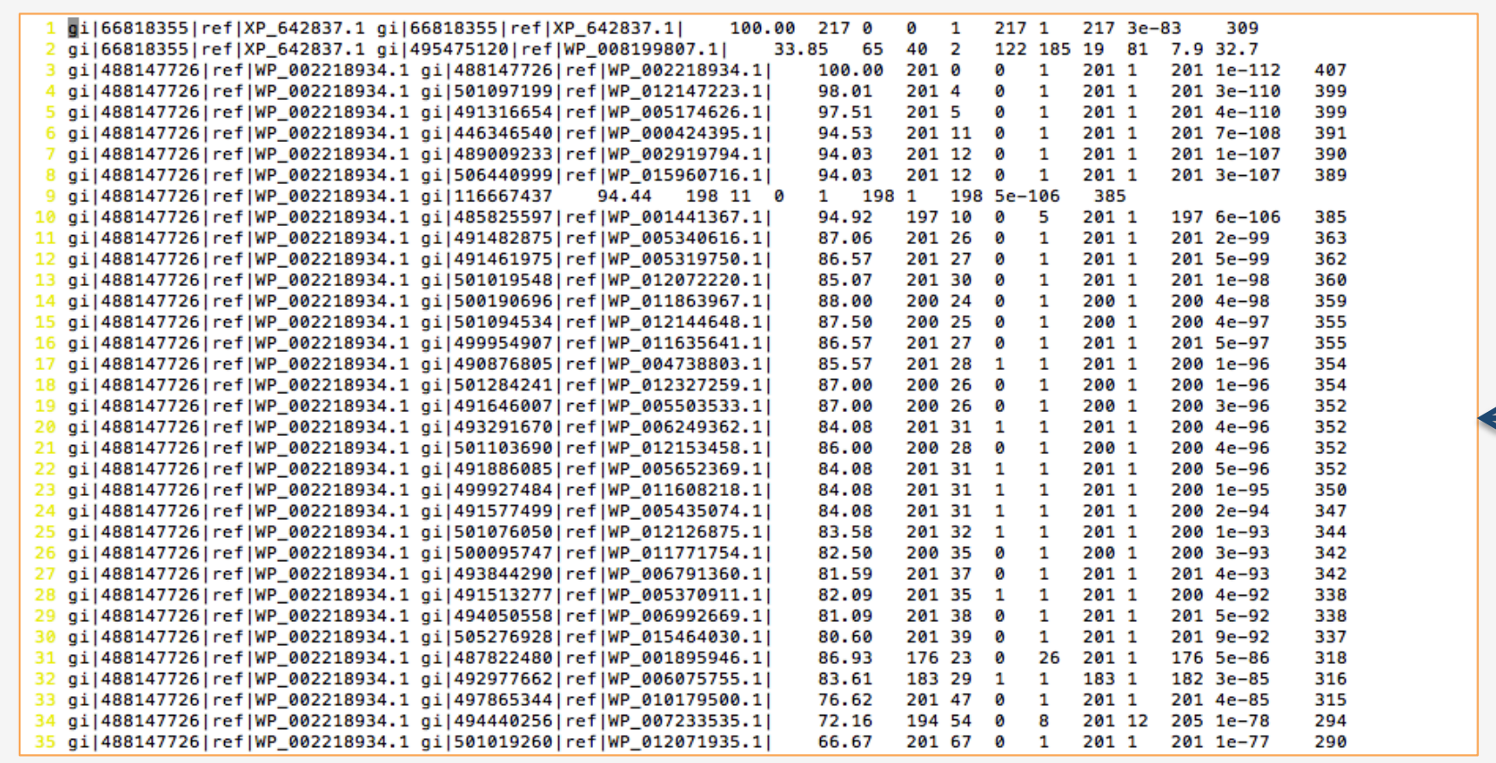

Figure 5. Tabular output from HSP-BLAST.

Output:

There are many possible output formats available for sequence analysis and the chosen output is dependent on which analysis process was utilized and the end goal of the study.

New Models and Discoveries
Future Studies

Future Goals

In the future, the PoPLAR Gateway will be a comprehensive resource for the manipulation and analysis of genetic data. It will allow scientists to compete with the large amount of data generated by today's biological research, and create a user-friendly environment that accommodates rather than hinders those with little computational experience.

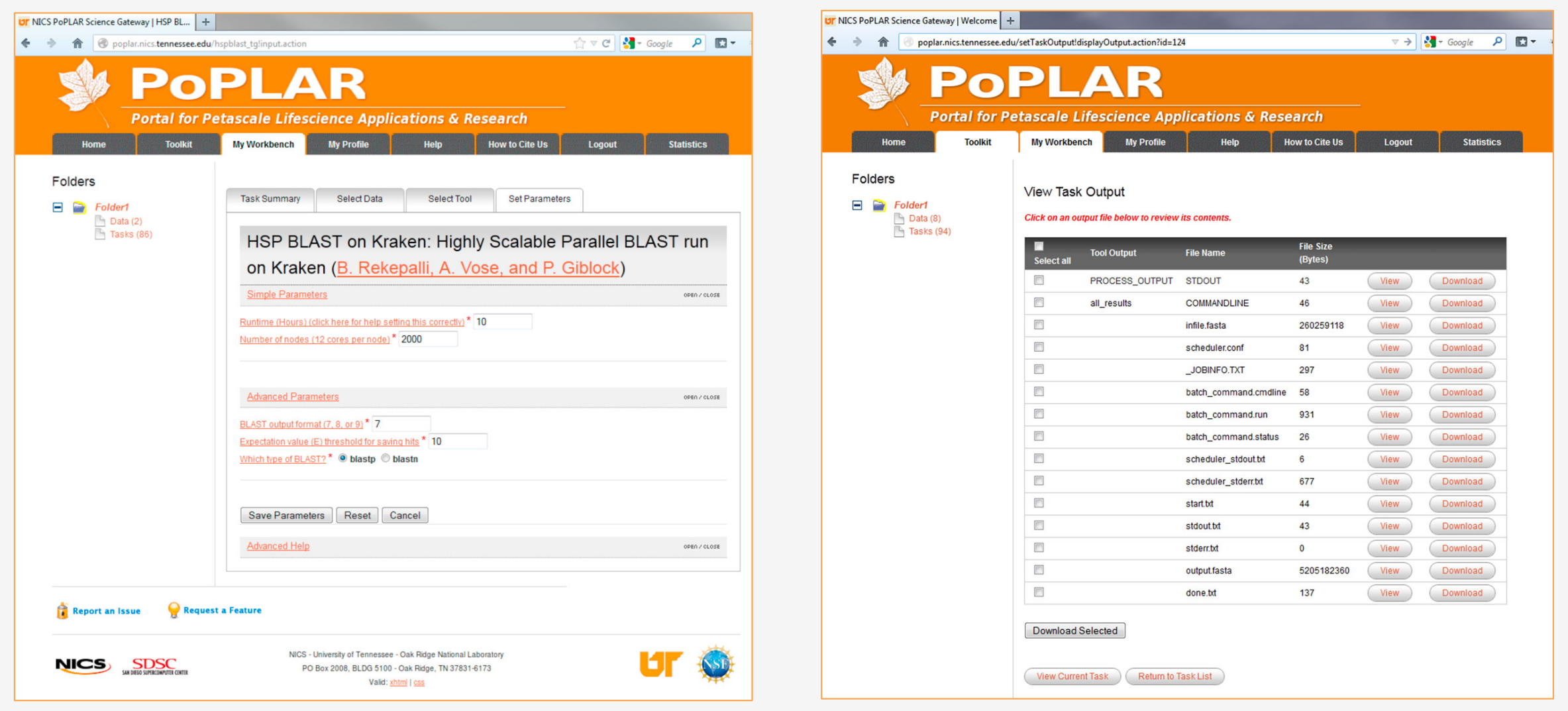

Figure 6 & 7. Screenshots of the PoPLAR Gateway.

References

Dewey, F. E. (2012-02-21). DNA sequencing: clinical applications of new DNA sequencing technologies. Circulation (New York, N.Y.), 125(7), 931.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. Nature Reviews Genetics, 14(3), 157+.

Rekapalli et al.: PoPLAR: Portal for Petascale Lifescience Applications and Research. BMC Bioinformatics 2013 14(Suppl 9):S3.

Zerbino, D., & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 821-829.

http://ted.bti.cornell.edu/cgi-bin/epigenome/

http://www.bio.davidson.edu/courses/

Bhanu Rekepalli: brekepal@utk.edu
Mary Lauren Harris: Mary_L_Harris@baylor.edu
Catherine Eason: catheason09@gmail.com