Bioinformatics Analysis Tools Using the PoPLAR Science Gateway

Mary Lauren Harris (Baylor University), Eduardo Ponce (JICS, UTK-ORNL), Bhanu Rekepalli (JICS, UTK-ORNL)

Introduction

Recent advances in Next Generation Sequencing (NGS) technology, generate a large volume of genetic data which 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. The gateway provides the ability to upload data and an interface for users to select the desired command line parameters through graphical means. Additionally, the programs themselves are wrapped and scaled to large parallel architectures, improving the performance to a level that is out of reach for single machines or small clusters. In particular, the addition of reliable high performance computing (HPC) programs to the science gateways opens the doors to computational ability, even for scientists with little to no programming experience or resources.

Methods

All v. All BLAST on 3 datasets:

Set 1 (Con): ~15k sequences
Set 2 (Non): ~30k sequences
Set 3 (All): ~80k sequences

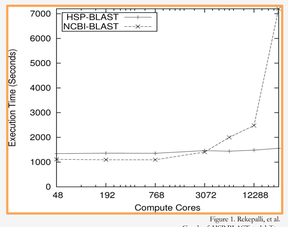

Figure 1. Rekepalli, et al.
Graph of HSP BLAST scalability

Used HSP-BLAST on Darter supercomputer

Compiled/extended documentation for MaCH, an open-source genotype imputation program

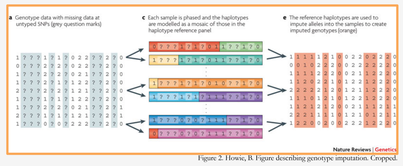

Figure 2. Howie, B. Figure describing genotype imputation. Cropped.

Figure 3. Sample MaCH documentation after modification.

Results

The results shown in both the heat maps and 3D images represent the percent overlap between each file of the dataset and each of the other files in the dataset.

The darker colors represent higher percentages of overlap.

The results show the expected result: that files have a nearly perfect overlap with themselves and that a file comprised of all the sequences of the dataset would have complete overlap with every other file.

The matrix appears to also be symmetric, but slight differences Do occur. Further research is required to determine the significance of this finding.

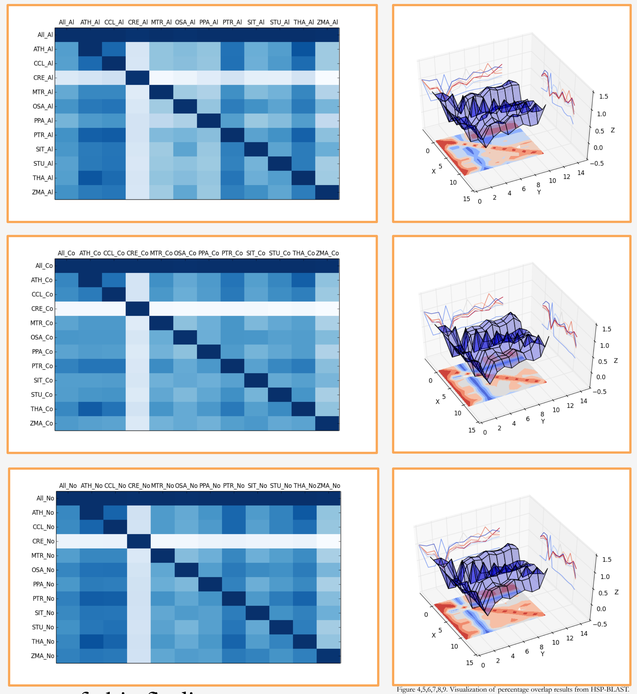

Figure 4,5,6,7,8,9. Visualization of percentage overlap results from HSP-BLAST.

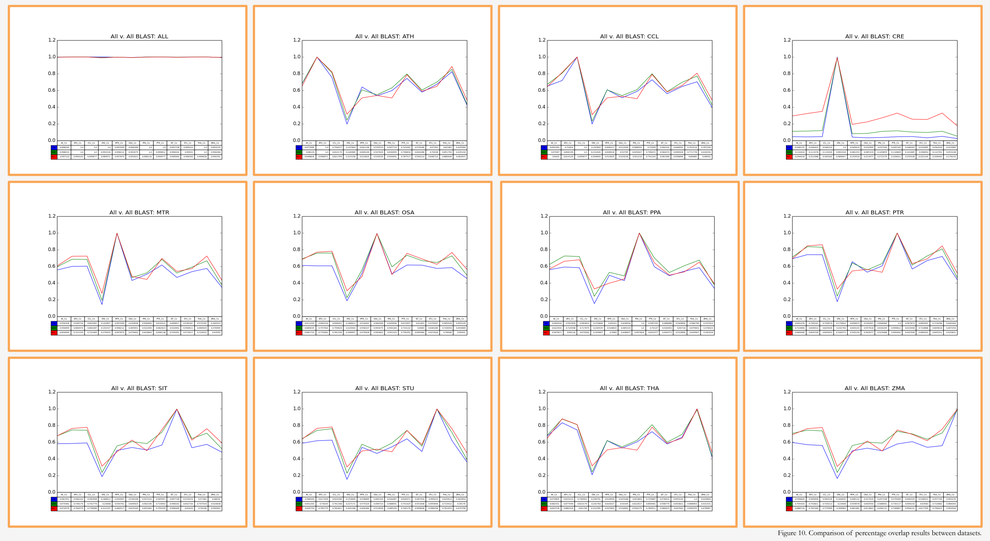

Figure 10. Comparison of percentage overlap results between datasets.

The graphs above show the three datasets in relation to one another. The larger datasets showed higher percentages; however the general shape of each line is similar.

Future Goals

Future goals include:

- Extension of HSP-BLAST all v. all analysis
- Improved documentation for scientific programs
- Automated workflows and HSP implementation of additional tools
- Increased accessibility for computational tools through science gateways

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.

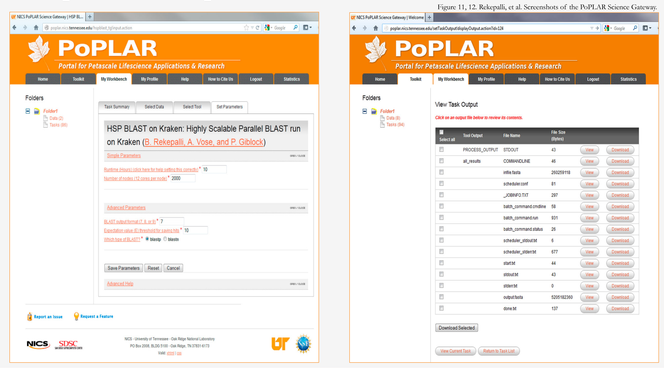

Figure 11, 12. Rekepalli, et al. Screenshots of the PoPLAR Science Gateway.

References

Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 499-511.

Hunter, J. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, *9*, 90-95.

MACH 1.0 - Markov Chain Haplotyping. (n.d.). *MACH 1.0 - Markov Chain Haplotyping*. Retrieved from http://www.sph.umich.edu/csg/abecasis/MaCH/

Moreno-Hagelsieb, G., & Latimer, K. (2007, November 26). Choosing BLAST options for better detection of orthologs as reciprocal best hits.

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

Mary Lauren Harris: Mary_L_Harris@baylor.edu
Bhanu Rekepalli: bhanu@utk.edu
Eduardo Ponce: eponcemo@utk.edu