Principal Investigator:

Dr. Sergei Maslov

Professor of Bioengineering and Bliss Faculty Scholar at the University of Illinois Urbana-Champaign with appointments at the Department of Bioengineering, Carl R. Woese Institute for Genomic Biology, National Center for Supercomputing Applications, and Department of Physics.

Co-principal Investigator:

Dr. Sean R. McCorkle

Senior bioinformatician, Department of Biological, Environmental and Climate Sciences,
Brookhaven National Laboratory, Upton, New York

Title

Quantitative models of bacterial evolution inferred from large number of draft genomes

Computational analysis of large (>1000) collections of draft genomes of bacterial strains.

The current collection of microbial genome sequences includes around 40,000 sequences, and is growing exponentially. Most of these genomes are not in complete form (fully assembled chromosomes), but are left in draft form, consisting of incompletely connected contigs of various number and sizes.

We recently proposed (see Ref. [1-3]) a suite of computational algorithms aimed at extracting the "basic" (or “core”) genome shared by most strains of a given bacterial or archaeal species [1-2]. This basic genome [2] can then be used to detect vertically and horizontally transferred genomic segments in order to quantify their relative contributions to genome evolution. We proposed to expand our computational analysis originally developed and tested for a small number (<100) of high quality "finished" genomes to a large number (>1000) of lower quality "draft" genomes. To accomplish that we leverage scalable computing resources of Blue Waters.

In the course of this project, we constructed the high quality basic genome of the bacterial species *Staphylococcus aureus*. We then used this basic genome to detect pairwise similarities among 2949 S. aureus draft genome sequences. Our original estimate is that it will require 14,000 node hours. We ended up using our allocation to compute pairwise similarities among ~1.2 million pairs of genomes (27% out of 4.3 million). It looks like we underestimated the run time by a factor of 4.

We saved 3 output files from each run from his program, discarding everything else because of file size concerns. These files are small, but there are over a million of them which causes operating systems to struggle doing even the simplest things. We ended up running in super-batches of 100,000 pairs and making sub-batches of 1000 to make them more manageable. The collected and compressed (tarred and gziped) batches total to something like 8 Gbytes or so, so the final size is not a problem. The data were copied into the home directory, and also into the sciteam project directory: /projects/sciteam/bacv/S_aureus_runs_spring_2016/
Both of these are not purged after 30 days like the scratch area. We also rolled off copies onto a local drive for safe keeping.

Description of Code(s)

The reference genome and SNP analysis code is a package or ensemble consisting of (1) a multiple, whole genome sequence aligner, *progressiveMauve*, which is an open source work provided by the Darling Lab (http://darlinglab.org/mauve/user-guide/progressivemauve.html), written in C++, and (2) a custom post-processing script (python) which performs the SNP analysis on aligned genomes.

This ensemble is executed for every pair of draft genomes for a given species. The input to one instance of the ensemble is a pre-selected reference (core) genome sequence, and the two draft genome sequences of the pair under consideration. *progressiveMauve* is invoked to perform a multiple alignment of the three genomes, and creates multiple alignment file that is subdivided into segments. When that completes, the postprocessing script reads the alignment and calculates a number of SNP statistics, which are stored in files of a few kb or less. Our code is "embarrassingly parallel", that is calculations for each of the pairs can be performed independently.

References

1) Studier FW, Daegelen P, Lenski RE, **Maslov S**, Kim JF (2009) Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes. J Mol Biol 394: 653–680. doi:10.1016/j.jmb.2009.09.021. ***(Featured on JMB cover)***
2) Dixit PD, Pang TY, Studier FW, **Maslov S** [corresponding author] (2015) Recombinant transfer in the basic genome of Escherichia coli, PNAS 112:9070–9075, doi: 10.1073/pnas.1510839112, http://arxiv.org/abs/1507.03972; ***(Reviewed by the Faculty of 1000)***
3) Dixit PD, Pang TY, **Maslov S**. [corresponding author] (2016) Recombination-driven genome evolution, population structure, and stability of bacterial species, (Genetics under review) bioRxiv 067942; http://dx.doi.org/10.1101/067942.