Communicating Science in an International Arena: The i5k Initiative

Susan J. Brown, University Distinguished Professor of Biology,
Kansas State University

A few years ago I described to this group an international collaboration that was organized to analyze the genome sequence of the red flour beetle *Tribolium castaneum.* This small beetle, a cosmopolitan pest of stored grain, is a premier model organism for studies in developmental biology and pest management. When the Tribolium genome was sequenced, the cost of sequencing an insect genome of approximately 200 Megabases (10 fold smaller than the human genome) still exceeded 2-3 million dollars. The project received support from the National Human Genome Research Institute (NHGRI) and the USDA. Since then, a new generation of sequencing technology has been introduced.

As technical advances lower the costs of high through-put sequencing, genome sequencing is no longer limited to large sequencing centers and external sources of funding. Individual academic and industrial research groups are getting involved, sequencing genomes to address questions in biology, physiology, biochemistry and phylogenetics. This avalanche of data creates special challenges in how to store, share and analyze huge extremely large datasets with an international cohort of collaborators.

As sequencing costs plummet, sequencing projects are directed toward more ambitious goals. The G10K proposes to sequence 10,000 vertebrate animals to address questions of animal biology and evolution[1]. The 1000 Genomes Project will characterize variation in the human genome to understand the roles such variation has played in our history, evolution and disease[2]. The goal of the 1KITE project (1000 Insect Transcriptome Evolution, http: //www.1kite.org/) is to construct a more complete and accurate phylogenetic tree of insects, based on gene sequences. At K-State, we established a center for genomic studies of arthropods affecting plant, animal and human health. This center established an annual symposium focused on arthropod genomics that has organized and energized the community of arthropod genomic researchers.

In March 2011, the i5k initiative was announced in a letter to the editor of the journal Science[3]. This initiative is based on the understanding that these new sequencing technologies allow us as a research community to sequence not only all the species of interests, as in the G10K, but we can also sequence all the individuals of interest, as in the 1000 Genomes Project. Given the goal of the i5k to sequence the genomes of 5000 insect and related arthropod species, the first challenge is to determine which

species to sequence, and how to prioritize them. Moreover, it is important to justify not only each species, but the entire i5k project. As stated in the Science editorial, the i5k consortium is interested in all insects of importance to agriculture, medicine, energy production, and those that serve as models, as well as those of importance to constructing the Arthropod Tree of Life[4].

Sequencing 5000 insect genomes will provide a wealth of information about the largest group of species on earth. First, the genome sequences will serve as a resource for gene discovery. Identifying genes of interest is the first step in understanding mechanisms of pesticide resistance or disease. Genome sequences will also serve as substrates for analysis, to detect signatures of selection and structural rearrangements associated with their evolutionary history. Genome sequences will also serve as gateways for biological inquiry, providing a "part lists" of genes for investigation.

Arthropods display a breadth of biodiversity unparalleled in other animal groups. In approaching the ambitious goal of sequencing 5000 insect genomes, the i5k initiative began by taking suggestions from insect biologists and prioritizing them based on several criteria. First, the scientific impact of sequencing the genome of each candidate species is considered. A high priority candidate might be relevant to problems in agriculture, human health, evolution or ecology. Some species are used as model systems for basic biological studies, while others are serious pests or vectors of diseases. Still others fulfill beneficial functions, including pollination or biological control of other pests. Second, although access to genome sequences is likely to attract the interest additional researchers, the size of the research group that is already focused on a particular insect or group of insects is considered.

Additional prioritization criteria address the feasibility of a genome sequencing project. Genome sizes vary widely among arthropods and those with relatively small genomes (100-500 Million bases) are usually given higher priority. In general, smaller genomes, containing less repetitive DNA , are easier to sequence and assemble. This explains why genomes of viruses and bacteria were the first genomes to be tackled, and continue to proliferate apace in genome databases (e.g. the genome database at the National Center for Biomedical Information, NCBI). Sample availability is also considered. Model organisms, which have been reared in the lab for several years are often inbred or can be inbred, reducing heterozygousity, which also confounds sequence assembly. In the absence of an inbred strain, a sample from a single large specimen, or haploid organisms, such as hymenopteran males are given higher priority. As technology advances and the costs of sequencing continue to decline, these issues may be of less significance. But regardless of the amount of data that can be afforded, the assembly algorithms must also improve.

The next generation sequencing technologies that have yielded the most dramatic cost reductions produce exceedingly large datasets. Many algorithms have been developed to assembly

a genome from these whole genome shotgun reads, and all are computational expensive. In addition, the resulting assembly must be assessed for accuracy and coverage. Once a satisfactory assembly is a produced, it must be annotated, archived and made available to the research community. After the initial, mostly automated, annotation, a new genome sequencing project is described in a publication which marks the "release" of the genome and the genome project is simultaneously submitted to NCBI; the raw reads are submitted to the "short read archive" and the entire project is submitted as a bioproject (http://www.ncbi.nlm.nih.gov/bioproject).

At K-State, we are using NGS sequence data to improve the original Tribolium castaneum genome assembly. In addition, we have sequence the genomes of three related species. Each project generates terabytes of data to be archived, analyzed, and shared with an international group of collaborators. To update the original T. castaneum genome assembly we used short reads supplied by our colleagues in Germany, which were hand delivered at an international AGS symposium on seven CDs. We worked with the DNA sequencing center at the University of Missouri to generate the sequence data for the related species. Just downloading these datasets took several hours each. Although the initial assembly, annotation and analysis take a team of bioinformaticists and biologist several months, the end product is not static. Each project is dynamic; a web portal for community feedback is an essential component. To be useful, genome annotations must be periodically updated by a combination of community feedback and continued analysis. We currently provide community access to three genome projects at K-State through agripestbase (agripestbase.org). To solve problems in distributed resource sharing we are working with our high performance computing (HPC) group to explore solutions offered by Globus[5].

While the community of scientists associated the i5k continues to expand, it is important to point out that several are right here at Universities that participate in the Merrill retreat. These include Amy Toth at Iowa State University, Chris Elsik at the University of Missouri and Jennifer Brisson at the University of Nebraska, Lincoln as well as a host of colleagues at K-State.

Although each arthropod genome that is sequenced provides a wealth of new data, there is a broader research perspective to consider. Each species is part of a broader community including parasites, predators, prey, competitors, endosymbionts, and conspecifics. Sequencing the genomes of the interacting members of these communities will provide the foundation for arthropod community genomics and even larger datasets to consider.

References Cited:

1. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered *100*, 659-674. 19892720
2. Consortium, G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073. 20981092
3. Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M., and Schneider, D.J. (2011). Creating a buzz about insect genomes. Science *331*, 1386. 21415334
4. Maddison, D.R., Schulz, K.-S., and Maddison, W. (2007). The Tree of Life Web Project. Zootaxa *1668*, 19-40.
5. Foster, I., Kesselman, C., NIck, J., and Tuecke, S. (2003). The Physiology of the Grid. In Grid Computing: Making the Global Infrastructure a Reality, F. Berman, G. Fox and T. Hey, eds. (Chichester, UK: John Wiley and Sons, Ltd).