Genome Analysis News

September 2007 Volume 2 Issue 3

Origins of multicellularity

An NHGRI project is underway at the Broad to use comparative genomics to understand the origins of multicellularity. Eukaryotes range from those with a single cell, such as the amoeba, to complex multicellular animals, including humans. The vast majority of life on Earth has been dominated by unicellular life. Animals, fungi, plants, and other multicellular lineages evolved multicellularity separately, and each lineage has a different common ancestor. As part of this project, the Broad will sequence 10 species located near branch points and other strategic locations of the phylogenetic tree to answer important questions on how multicellularity evolved in each lineage, as well as to gain better insight on a set of well-established animal developmental genes that can be traced to the beginning of the animal lineage but that are completely absent in fungi. The sequence will also better inform the structure of the fungal evolutionary tree, particularly at the basal branches. In addition to several fungi, the organisms to be sequenced include apusomonads and choanoflagellates, which are single-celled, colony forming eukaryotes that are the closest living relatives of multicellular animals.

Choanoflagellate

What has 6 legs, 3 ears, 4 tusks and 2 trunks?

An elephant with spare parts!

High-coverage sequencing to create a parts list for the African savannah elephant (*Loxodonta africana*) is underway. The project commenced at the Broad with our hosting of an elephant community meeting on May 4, 2007, where the research community expressed interest in a variety of aspects related to the genetics of the elephant, particularly elephant welfare and conservation, the ivory trade, the major histocompatibility complex (MHC) and disease, and paleogenomics and the road to an ancient mammoth genome. A SNP discovery project is also planned in which four species of elephant will be sequenced to 0.1X. Identification of markers (e.g., SNPs) will help labs to more efficiently type poached tusks to determine their areas of origin. The elephant is an ideal species for reconstruction of historic genomic events; it represents Afrotheria, the most basal of the eutherian clades. Their long life spans (70 years) may be of interest for longevity research, and they are subject to diseases with homologues in humans or livestock. Finally, the elephant genome may shed light on the evolution of advanced traits; elephants have large brains and cortical volumes that allow them to exhibit relatively high levels of learning and memory, and complex systems of communication and social interaction.

And the EURYI goes to ...

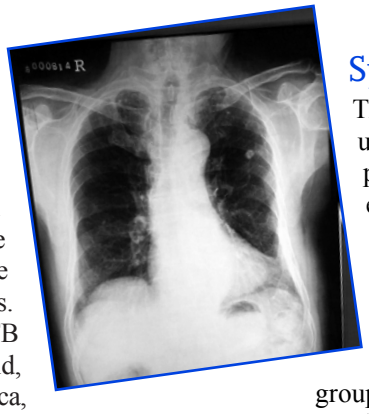
Kerstin Lindblad-Toh has been awarded a 2007 European Young Investigator (EURYI) award for her work on disease gene mapping and functional genomics in the domestic dog. The award will help support Kerstin's research program at Uppsala University, which focuses on the characterization of disease phenotypes in dogs, the generation of tools for disease gene mapping, the identification and functional characterization of canine disease genes, and the application of this knowledge to human disease. Kerstin is currently on 75% sabbatical in Sweden as a guest professor in the Department of Medical Biochemistry and Microbiology at Uppsala University. To see the full story go to http://www.broad.mit.edu/cgi-bin/news/display_news.cgi?id=3642.

Kerstin Lindblad-Toh

Tick bites

Ixodes scapularis, the blacklegged or deer tick, transmits a number of organisms that cause disease, but is best known as the principal vector of *Borrelia burgdorferi*, the causative agent of Lyme disease. Ticks rank second to mosquitoes in importance as vectors of human disease. As part of the NIAID contracts for Microbial Sequencing Centers (MSC), the J. Craig Ventor Institute (JCVI) and the Broad equally shared the sequencing of the *I. scapularis* genome; in addition, the Broad sequenced a library of cDNAs yielding an unprecedented resource for the tick community. JCVI has begun the genome assembly and plans to pioneer the genome annotation effort with the assistance of several Broad scientists. The tick genome sequence is proving extremely difficult to assemble because of the tick's large genome (1.2–1.4 billion base pairs, comparable in size to *Aedes egypti* and nearly half the size of human genome), the high repeat content (~70% repetitive, compared with human, which is 50% repetitive) and unexpectedly high rate of polymorphism, a consequence of differences between individuals of the tick colony targeted for sequencing. No inbred strain of the tick genome was available, so instead an inbred colony was targeted for the DNA extraction.

The A to Z of TB



Tuberculosis (TB) — a disease caused by *Mycobacterium tuberculosis* — has probably killed more than 100 million people in the past 100 years despite the fact that treatment for most forms of the disease has been available for over 50 years. Although there has been a decline in TB incidence in much of the developed world, incidence has increased dramatically in Africa, Eastern Europe and the former Soviet Union so that, overall, the global caseload continues to rise and this disease still kills about 1.5 million people annually. The main reasons for this are co-infection of AIDS patients by TB and the lengthy and complex treatment required to cure TB. Compliance with a treatment regime consisting of at least four antibiotics administered for 6 months is extremely low, leading to relapsing disease and rising drug resistance.

The overarching goal of the TB research projects in our program is to find faster and more effective anti-tubercular drugs by devising better chemical screens, by finding better drug targets, and through a more complete understanding of TB pathogenesis. Hence, we are planning drug screens, undertaking comparative genomic analyses of TB strains with important clinical differences, and applying the best gene expression and systems biology analysis methods available. Each of these approaches has the potential to provide a critical breakthrough for better TB treatment.

James Galagan oversees the Genome Sequencing and Analysis program's TB research and plays a key role in reaching out to collaborators who have a long history working on TB and trying to help them apply new and developing technologies to questions in TB. Collaboration with the TB community is crucial for furthering progress in this field, and has helped us determine the important questions in need of answers. Currently, we can group our projects into three areas: genetics, systems biology, and persistence and pathogenesis (see figure).

Drug resistance

One of the first questions to be answered is the problem of drug resistance in TB. The use of anti-TB drugs has led to the emergence of drug resistant (DR), multi-drug resistant (MDR), and even extensively drug resistant (XDR) strains of TB. Resistance to anti-TB drugs was noted soon after these agents became available in the 1940s. Over the past decade, however, rates of MDR TB have increased, particularly in areas with a high burden of TB and with sub-optimal control programs. Drug therapy has also led to the emergence of XDR TB. A devastating outbreak of XDR TB in Tugela Ferry, in Kwazulu Natal, South Africa, brought XDR to the world's attention, but it was not the first report on the emergence of XDR TB; a report in early 2006 estimated that 7% of samples of MDR TB globally were XDR TB. We've learned, however, that drug resistance comes at a fitness cost, and as a result some strains acquire mutations that compensate for this cost. These poorly understood compensatory mutations may contribute to the transmissibility of drug resistance. For this project we propose to sequence and analyze more than 100 drug sensitive and resistant TB strains to identify primary resistance, compensatory, and predisposing mutations. Megan Murray and Mark Borowsky lead this project, which recently released data on three strains from Kwazulu Natal and in which they will search for drug resistance mutations as well as other mutations.

Systems biology

The program has also undertaken a number of projects to understand *M. tuberculosis* from a systems biological perspective. Using a combination of computation and experiment, Desmond Lun and Brian Weiner are investigating gene regulation in TB from operons to the full regulatory network. As part of this work, Desmond is collaborating with David Sherman (Institute or Systems Biology, Seattle) to apply the newly developed Chip-Seq Solexa process to map transcription factor binding sites in this pathogen. The group is also developing techniques for integrating expression and regulation information with metabolic modeling. Caroline Colijn (a postdoc with both Megan Murray and James Galagan's group) has developed a new approach for combining TB expression data with models of metabolism that can be used to predict the impact of antimicrobial drugs. Desmond Lun and Aaron Brandes have extended this approach to predict nutrient utilization from expression data. All this work draws on a substantial body of genomic data for TB that can be difficult to access and use. To address this issue, a Gates Foundation funded portal of information called the TB Database (www.tbdb.com) has been developed in collaboration with Gary Schoolnik and others at Stanford University and will officially launch in October 2007. TBDB will provide a single site for access to all TB genomics data as well as tools for querying, downloading, visualization and analysis.

Tuberculosis Projects and Collaborators

Genetics: drug resistance and diversity

- XDR TB — Mark Borowsky, Megan Murray (HSPH)
- TB diversity — Sebastian Gagneux (ISB), Peter Small (ISB)

Systems biology: computation, regulation and metabolism

- TB database — Phil Montgomery (Broad), Gary Schoolnick (Stanford)
- Regulation —
  - Operon prediction and analysis: Brian Weiner, Krishnan Eswaran
  - Chip-on-seq to identify TB TFBSs: Desmond Lun, David Sherman (ISB)
  - Computational modeling of TB regulatory networks: Desmond Lun, GNS
  - SVM prediction of TB kinase targets: Mark Borowsky, Bob Husson (CHB)
- Metabolic modeling —
  - Interpreting TB expression data with metabolic models: Caroline Colijn
  - Predicting external nutrients from expression data: Aaron Brandes, Desmond Lun

Persistence and pathogenesis

- Persistence — Mark Borowsky, Kim Lewis (Northeastern)
- Primate models of pathogenesis — Sarah Fortune (HSPH), Joanne Flynn (U. Pitt.)

Persistence and pathogenesis

An NIAID white paper authored by Mark Borowsky in collaboration with Kim Lewis of Northeastern University seeks to answer another important question in the field of tuberculosis: What is the nature of persistent/latent infection? Persistence is believed to underlie clinical latency in TB, a condition thought to affect as much as one third of all people. In the asymptomatic latent state of the disease, *M. tuberculosis* is found in closed-cavity granulomas where it can survive without much replication for up to 30 years. Although persistent tuberculosis has been studied for decades, no molecular mechanism has been identified so far. Using classical bacterial genetics, Kim Lewis will make mutants of TB, and we will sequence the mutant genomes to identify the lesions responsible for the phenotypes of interest. Persistence pathways in TB will be attractive drug targets and will provide molecular markers for persistent *M. tuberculosis* cells that are currently impossible to identify in, or isolate from, latently infected tissues.

New Hires ...

please join us welcoming more new members of the Genome Sequencing and Analysis program.

Mia Champion is one of two new research publication coordinators whose primary role is to coordinate interactions between the community and Broad scientists in order to prepare findings from collaborative projects for publication. Mia received her PhD in genetics from U. California, Davis. She was an editor with Molecular Cell for ~3 years and still likes to read scientific papers. She also enjoys oil painting, biking and cooking.

Edward Freeman is a software engineer with the Annotation Informatics group. Previously, he did research and advanced development in parallel and high-performance scientific computing. He has degrees in physics (UMD), computer science (Harvard), and most recently bioinformatics (Brandeis). He is also vice president of the Boston ACM chapter.

Georgia Giannoukos worked in the Platform's Technology Development Group evaluating new sequencing technologies coming into the Broad. Her new position brings her to the Program doing application development for new sequencing technologies. Georgia likes to travel and take long weekend trips to Europe (1–2x/year). To relax, some people like yoga, but she makes pottery and some of her pieces are floating around the Broad.

Brian Haas is a computational biologist, initially focusing on the analysis of the genome of *Phytophthora infestans*. Brian received a BS in Biology and MS in Molecular Biology from the State University of Albany, near where he grew up, and an MS in Computer Science from Johns Hopkins. He worked at TIGR for 8 years, contributing to the annotation and analysis of numerous diverse eukaryotes.

Theresa Hepburn is a bioinformatics assembly analyst in the Assembly Analysis group. After graduating from RPI in May with a BS in bioinformatics she spent a week hiking the Appalachian trail from the MA–VT border to a few miles south of I-90. She's spent the past month redecorating her recently purchased condo in West Roxbury. Theresa's current project is assembling the KZN strains of *M. tuberculosis*.

Robert Riley is a bioinformatics scientist working on curation of an online TB genomics database. Originally a Boston native, Robert comes to the Broad from UCLA, where he received his PhD in Human Genetics and also worked on the bioinformatics of TB. He spends most of his spare time with his family, and is an enthusiastic guitar player and bicyclist.

Sharadha "Sharda" Sakthikumar is an associate computational biologist and is currently working on the *Batrachochytrium dendrobatidis* genome. She graduated from U. Pennsylvania in May 2007 with an MS in Bioinformatics. In her spare time, Sharda enjoys listening to audio books and loves traveling to "World Heritage" sites (has seen 21 — 830 more to go).

Sophia Zarakhovich is an associate computational biologist working on the comparative genomics of *M. tuberculosis*. Sophia worked at PDL BioPharma in Fremont, CA, where she worked on compiling prior art analyses for potential drug targets and in research project coordination, such as a pipeline for large-scale material production for researchers.

Photo by Maria Nemchuk

And so concludes the fourth successful Undergraduate Summer Research Program in Genomics, which "graduated" seven talented students in 2007. The program is funded by NHGRI as part of a program to increase the presence of underrepresented minorities in the field of genomic research. The program is spearheaded at the Broad by Bruce Birren (far right) and Shawna Young (far left), and was ably aided this summer by Maura Silverstein (second from left). Also pictured are students from the Integrative Cancer Biology Program (ICBP) and HHMI's Exceptional Research Opportunities (EXROP) program.