

i5k

The 5,000 Insect Genome Project

By Richard Levine

It's been called "the Manhattan Project of Entomology," an undertaking that has the potential to revolutionize the way we think about insects. The i5k Initiative aims to sequence the genomes of 5,000 insects and other arthropods over the next five years in order to "improve our lives by contributing to a better understanding of insect biology and transforming our ability to manage arthropods that threaten our health, food supply, and economic security" (Robinson et al. 2011).

The timing is incredible. A recent report shows how the Human Genome Project produced nearly \$800 billion in economic output from an initial investment of just \$3.8 billion (Battelle 2011). The 5,000 Insect Genome Project has the potential for similar yields at a fraction of the cost.

The initiative was first announced in March 2011 in a letter to *Science* from ten signers known as the i5k Ad Hoc Launch Group. According to their Web site (<http://www.arthropodgenomes.org/wiki/i5k>), "This project will be transformative because it aims to sequence the genomes of all insect species known to be important to worldwide agriculture, food safety, medicine, and energy production; all those used as models in biology; the most abundant in world ecosystems; and representatives in every branch of insect phylogeny so as to achieve a deep understanding of arthropod evolution and phylogeny."

Richard Levine, ESA's Communications Program Manager, conducted the following phone interviews with four members of the Launch Group. Gene E. Robinson, a professor at the University of Illinois at Urbana-Champaign, studies how genes influence the social behavior of honey bees and how the honey bee's social life influences its genome; Susan J. Brown, a professor at Kansas State University, is a classical geneticist who has worked on the evolution of gene regulatory networks in the red flour beetle; Kevin J. Hackett is a national program leader at the USDA Agricultural Research Service whose projects have involved biological control and many other insect-related areas in agriculture; and Daniel Lawson, a coordinator at the European Bioinformatics Institute, has worked on genome annotation and genomics databases for nearly 16 years.

Q: How did this project get started?

A: Kevin Hackett: In 2009, Gene Robinson had received a Director Pioneer Award from the National Institutes of Health, which allowed him to sequence about 15 bee genomes. Then, in May 2010, he approached me with the idea of doing something comparable in the greater insect world. It was time to take things to the next level.

A: Gene Robinson: We started looking at the timing of new developments in sequencing technology, which will, within two years or so, make sequencing costs so low that we will get to a thousand-dollar human genome. We realized that when we get to that point, it's really going to change all of biology because genome sequencing will be so affordable that we'll be able to think of projects on a massive scale,



and then we started thinking it was really time to go for all of entomology in one large, integrated vision and put it out there for people to think about and develop into a real plan.

The time seemed right to articulate a vision, so we formed a group of leading experts in insect genomics in order to discuss and flesh this out, and that really culminated in the letter to *Science*, which presented the idea and indicated that the time is now to plan for this. The sequencing costs of today do not yet make the project feasible, but the sequencing costs in a couple of

years from now will. Many insect genomes—not all, but many—are about one-tenth the size of a human genome, so when we get down to the thousand-dollar human genome, we're talking about a hundred-dollar insect genome. It's time to mobilize!

A: Daniel Lawson: Also in 2009, I met Susan Brown at an Arthropod Genome Consortium meeting, and she told me she wanted to make a community-resource Web site where people could build collaborations and document the interests and activities of arthropod researchers. So I produced a new wiki Web site (<http://www.arthropodgenomes.org>) and for about a year, we collected information about various species and people who were interested in those species, and we tried to build a community. Then when Kevin came along with the i5k proposal, it was a nice marriage to integrate their site with the arthropod genomes site, and it was a natural thing for me to take over some of the infrastructure.

Q: This has been called "the Manhattan Project of Entomology." Do you agree?

A: Susan Brown: I love that term—Manhattan, Kansas! And yes, I think it's going to be very transformative for the entire field of entomology. It will allow us to support molecular insect science in many ways that we are finding to be difficult right now.

A: Lawson: Yes, it changes the whole game plan for a lot of communities. Moving into the genetics era revolutionizes what you can do, what you can try to assay in your species, and what you can infer from your experiments.

A: Robinson: That's right, it's grand in scope and will change the playing field in entomology.

Q: Why did you choose the number 5,000?

A: Lawson: 5,000 is an arbitrary number. The number has to be large enough so you feel as if you're covering all arthropods of agricultural and medical importance, but we don't really know what that number should be. That's partly why we want people to sign up, so we can get a better understanding of what that number actually is. The idea of engaging with the genomics community, the bioinformatics community, and the entomological community is to find out which

species are actually important. We're not just choosing randomly, we're looking for species that people actually care about because there are economic or health benefits to better understanding the biology of these animals. We hope that generating this data will lead to better models for insecticide resistance, better models for developing new pesticides, better models for understanding transmission of disease, or for control of agricultural pests.

A: Hackett: We started talking about how many insects we should do if we were to start a major international effort, accounting for all of the agricultural pests, including medical and veterinary, and the human vectors, plus all of the branching-points insects. So 5,000 is actually low-balling it a bit, but we figured that 5,000 was doable if we could get the cost down to \$1,000 per genome, because 5,000 genomes times \$1,000 equals \$5 million. Then if we could get another \$5 million for bioinformatics and another \$5 million for mining the data, that would be a 15-million-dollar project. The honey bee project alone cost about \$9 million, and we're talking about doing 5,000 insects, so there's a financial aspect to it. Secondly, there are about a million described insects, so 5,000 is a relatively small proportion; it's a doable project, and it's a good place to start. In reality, as sequencing costs continue to come down, all of the key insects will be sequenced by somebody. It's more a matter of putting together the communities to make use of that data, to annotate the genomes, and to figure out what the genes are for and to make use of them, and that will require large international teams.

Q: What are the next steps after the genomes are sequenced?

A: Robinson: Sequencing a genome does not deliver biological knowledge instantly; it's just the first step in a transformative process that provides new information that can then lead to major new insights. Sequencing a genome is the start, but then one has to assemble it into its correct order, and then annotate it—that is, identify the genes, their different components, and those parts of the genome that regulate the expression of the genes. That gives the raw information from a genome. So you need sequencing, assembly, and annotation. The driving technological force is ultra-cheap sequencing, but the other two realms are still frontier areas. There are significant computational challenges that will need to be solved in the areas of assembly and annotation. Then, once one has the genome, it's time to extract information from it, to use it in experiments, and to mine information from it.

Q: Can you talk more about the information mining?

A: Robinson: For example, there may be one insect strain that is susceptible to a pesticide and another strain that's resistant to it. With sequence information, bioinformatic tools, and computational tools, it is possible to identify the changes in the genome that are associated with the resistance to a particular pesticide. Having all of the information about the genome will allow for entirely new approaches to control insects of agricultural and medical importance and to empower insect-based research where insects are used as models.

A: Hackett: For example, we could mine data for cytochrome p450 detox genes. Those genes are involved with detoxifying chemicals that are inside insects, so if we know about those genes from one insect

to another, we can use that information to actually kill the insects. Or if you take beneficial insects like honey bees, which do not have as many detoxifying genes and are more susceptible to chemicals, that kind of information could be used to help protect bees.

A: Lawson: We also want to ensure that the data that is produced by i5k ultimately makes it into public-domain archival sites for maximum usage and maximum dissemination to the communities that can use that information, so there will be connectivity with things like the National Center for Biotechnology Information, GenBank, the DNA Data Bank of Japan, and the European Molecular Biology Laboratory.

Q: What tangible uses do you see coming out of this research?

A: Robinson: This will provide information that breeders would need to look for ways of dealing with insect resistance to pesticides. It would also provide geneticists with information on what might be vulnerable points in an insect's makeup, which could be used for novel control strategies.

A: Lawson: I work on vectors, and one of the questions we're most interested in is, "What makes a species a vector?" So if you think about mosquitoes, we currently have three genomes of mosquitoes that are vectors of malaria and other diseases. But there are other very similar, related mosquitoes that do NOT act as vectors, so we want to sequence non-vector species as well so we can determine what makes an arthropod a vector or not a vector. I think many questions will be addressed in this way. We'll have a large corpus of data that everyone can mine and do all these comparatives, and each community will come up with very interesting pieces of biology that they can start to elucidate and tease apart.

A: Brown: Here at Kansas State University, we have an initiative called the Center for Genomic Studies on Arthropods Affecting Human, Animal & Plant Health, and we're basically interested in understanding the biology of insects and their interactions as vectors of pathogens and pests. So I see this as a major step towards understanding insect biology, which will help us understand those interactions, which in turn will help us to control them or eliminate them. I see a larger picture than simply going for the next pesticide target or talking about one gene at a time. Understanding developmental biology helps you treat cancer, for example, because you understand what the cancer-causing gene is supposed to be doing normally. Likewise, if we understand the normal development and interactions of these organisms, then we can hope to control them or fight them better.

A: Hackett: I look at three categories: emerging opportunities, novel opportunities, and transformative ones. Examples of emerging opportunities would be solving malaria, improving sterile insect technology, and biocontrol improvement. There might be ways of improving the production of insects or the insects themselves. Certainly there is also the possibility of finding pathogens by subtracting out the genome, which was sort of the method that was used to find the Israeli acute paralysis virus. Also, biodiversity and conservation areas are important.

Then there are novel opportunities, like using sensory receptors as detectors for biodefense. Insects are really good at detecting

things, so if one could use nanotechnology and other methods to use the information from the sensors, you could develop all kinds of detection devices for biodefense and for other reasons. Another novel opportunity involves interrupting pest immune pathways. Insects have immunity, like humans, and they have pathways that can be interrupted. Or finding RNAi targets within the genome could allow us to use RNAi silencing.

Then there are transformative opportunities: for example, elucidating brain function in bees. I'm also interested in elucidating evolutionary strategies in the cell based on compartmentalization of function—in other words, looking at how pathways are put together in the cell based on the DNA that we can discover from one insect to another through comparative evolutionary work and cellular biology, a systems biology approach. And then there's the possibility of creating GMOs based on insect improvements.

I'd also like to point out the importance of understanding the relationships between insects and other organisms. Since plants, animals, and microbes are already being sequenced, if you add insects to that, then you've got tritrophic relationships—host, insect, and pathogen—and if you can understand those relationships, you can interrupt transmission of plant or animal pathogens.

Q: Your letter in *Science* encourages “collaborative research by computer scientists, bioinformaticians, and biologists to overcome the challenges of handling this unprecedented volume of data and derive meaning from these genomes.” How do you aim to encourage this collaboration?

A: Brown: We plan on working with people at sequencing centers who already have bioinformatics pipelines. In fact, at the Arthropod Genomics Center, we have an annual symposium where we're trying to do just that with bioinformatics workshops at a level where people who are just learning how to do genomics can see what the tools are and how they're being used. So we're trying to bring people together already at an international symposium.

A: Robison: We hope to be able to develop a funding base for the whole initiative that not only includes money for sequencing, assembling, and annotating the genomes, but also for tools that will enhance the ability to understand and analyze this huge, unprecedented amount of genome information to allow scientists to make use of it. These would be in the realms of computational biology and systems biology, so we would want to have a new initiative on opportunities to develop new analytical tools in systems biology and computational biology to be able to derive the maximum amount of information and insight from the genome sequences.

Q: Ideally, what would the project look like in five or ten years?

A: Brown: It would look like a very strong cyber-infrastructure that would help the everyday biologist to access the genomic information for the insects he's interested in and to understand them at a much broader level.

A: Robison: We would have the sequences of 5,000 species, and new databases would make the information available to the scientific community in ways that are easy to use. We would have new forms of computational tools that would enhance our ability to derive information from the genome, and we would have strong, interac-

tive communities of scientists and entomologists working in these various areas, taking advantage of these tools, and interacting with each other to work on common problems and questions.

Q: How can ESA members participate in this project?

A: Hackett: Right now we're trying to get people to sign up and to create wiki pages (<http://arthropodgenomes.org/wiki/i5K>) to recommend which insect genomes should be sequenced while we seek funding. So if ESA members join the i5k effort, we can put all of our cumulative efforts together, which would help us attract the attention we need. We need to contribute and to share data, but we also need to be sharing and developing communities, and one big community could be created through the Arthropod Genomics Consortium and the i5k effort. We really need to expand our team internationally, so we're inviting entomologists across the globe to sign up.

A: Brown: We're trying to find out who's working on what insects, and if they feel that having genomic information about their insects would help. If you look at the ESA membership, quite a few researchers are probably working on transcriptomics, looking at the genes that are transcribed under certain contexts, environmental conditions, or life stages. Looking at the whole genome will help us understand these comparatively and not just in one organism.

A: Robison: They should join the wiki Web site and nominate species to sequence. We want this to be a broad-based, inclusive effort. The word “all” is important—we want all people to be involved, we want all insects of agricultural importance, all insects of medical importance, and so forth. Workshops will be organized and held, and there will be opportunities for further input, interactions, and the ability to shape the project.

A: Lawson: Sign up at the Web site, find out who's working on similar things to yourself, and if you are not doing so already, talk to them. This is an opportunity for entomologists to interact with scientists from wildly different skill sets. The bringing together of entomologists, geneticists, and bioinformaticians that happens during a genome project is engaging for all parties.

Q: What about student member participation?

A: Robison: It would be the same, but I would add that students should start imagining what their research agendas should be like in an era when every insect of interest to anyone, for any reason, has had its genome sequenced. Because during their professional careers—after their Ph.D.s and their postdocs in five or so years—they will be operating in this new landscape, which is totally different from what it is now. So how are they positioning themselves with their training? How are they preparing the way they think about things and the way they conceptualize problems with this in mind? Students should be preparing for a new genome-powered era of entomology.

References Cited

- Battelle 2011.** Economic Impact of the Human Genome Project: How a \$3.8 billion investment drove \$796 billion in economic impact, created 310,000 jobs and launched the genomic revolution. (<http://www.battelle.org/publications/humangenomeproject.pdf>)
- Robison, Gene E., et al. 2011.** Creating a Buzz About Insect Genomes. *Science Magazine* 331: 1386.