Susan Celniker: Foundational Resources To Study a Dynamic Genome

Susan Celniker[1]
Lawrence Berkeley National Laboratory, California 94720

The Genetics Society of America's George W. Beadle Award honors individuals who have made outstanding contributions to the community of genetics researchers and who exemplify the qualities of its namesake. The 2016 recipient, Susan E. Celniker, played a key role in the sequencing, annotation, and characterization of the *Drosophila* genome. She participated in early sequencing efforts at the Lawrence Berkeley National Laboratory and led the modENCODE Fly Transcriptome Consortium. Her efforts were critical to ensuring that the *Drosophila* genome was well-annotated, making it one of the best curated animal genomes available. As the Principal Investigator for the BDGP, Celniker has enabled the study of proteomes by creating a collection of over 13,000 clones that match annotated genes for protein expression in cells or transgenic flies, and she has established the most comprehensive spatial gene expression atlas in any organism, with *in situ* imaging of more than 80% of the *Drosophila* protein-coding transcriptome through embryogenesis. In addition to providing the research community with these invaluable resources and reagents, she continues to develop new tools and datasets for genetics researchers to explore the spatial and temporal control of gene expression.

There's a lot of mystery left in the genome.

—S.C.

It is difficult to imagine where the field would be without her voluminous, dedicated, and expert work.

—L. Perkins, Harvard Medical School

How Did You Become a Biologist?

My interests in biology started as a high school student with classes in comparative anatomy and invertebrate paleontology at the Museum of Natural History in Los Angeles under the direction of Ms. Gretchen Sibley and the curators. I was fortunate to get a summer job there as an assistant teacher for the primary school program teaching herpetology and ichthyology. I attended Pitzer College, which is one of the Claremont Colleges and is a primarily undergraduate institution. I chose a small college because it offers the flexibility to construct an individualized curriculum and provides easy access to faculty members. It turned out to be an excellent choice for me. Pitzer has a philosophy of nurturing independence; as a result, I learned how to find unconventional ways to solve problems. I started out interested in both biology and anthropology, but my emphasis on biology solidified as I gained more exposure to course material and to my professors. I was particularly influenced by one professor, Dr. David Sadava, who taught me Genetics and Cell Biology and encouraged me to apply for a summer internship at City of Hope National Medical Center. That was my first experience doing laboratory research, and it was when I finally understood the excitement of the discovery process itself. I was hooked! We were working on Huntington's disease before they had figured out any of the genetics; after graduation, I continued that work for another 2 years.

How Did You Go from Working on Huntington's Disease to Sequencing the Fly Genome?

I started graduate school working with Dr. Chi-Bom Chae at University of North Carolina at Chapel Hill, but after I passed

doi: 10.1534/genetics.116.196261
Available freely online.
[1]Address for correspondence: Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., MS 977-180, CA 94720. E-mail: celniker@fruitfly.org

my candidacy, my husband's job transfer moved us back to California. Fortunately, I had a very supportive department head who said I could do my dissertation work at a lab in California. I became a visiting student at Caltech and worked in the laboratory of Dr. Judy Campbell, though I defended my thesis at Chapel Hill. It was complex, but the situation gave me independence and I was determined to make it work. I studied DNA replication in yeast, setting up *in vitro* systems and characterizing the first identified origin of replication, ARS 1 (Celniker *et al.* 1984). For my postgraduate work I changed fields, making the logical progression from studies of DNA replication to studies of transcription, with a focus on the control of gene expression in a developmental system. I was eager to explore the molecular mechanisms that controlled temporal and spatial gene expression in a developing embryo, so when the unique opportunity came along to do a postdoctoral fellowship with a giant in the field, the Nobel Laureate Edward B. Lewis, I grabbed it. He was a classical geneticist, a pioneer in the field of evo-devo, who had identified and characterized the *Drosophila* Bithorax Complex (BX-C). My postdoctoral work involved exploring the structure and function of *Abdominal B* (*Abd-B*), the most distal gene in the complex (Celniker *et al.* 1989). I began by performing X-ray screens to identify loss-of-function mutants of *Abd-B*, but when the time came to characterize the mutants molecularly, I had to improvise; Ed's lab was not set up for molecular work. Dr. Norman Davidson, a physical chemist and molecular biologist, gave me space in his laboratory until I was able to build my own molecular lab.

I focused my research on the *Abd-B* mutant, *Transabdominal*, a spectacular dominant gain-of-function mutation that drives islands of sexually dimorphic abdominal cuticle in the dorsal thorax of the adult fly. To characterize the role of *Abd-B* in development, we needed to determine the sequence of the region. Sequencing was a huge task because the gene is one of the largest in *Drosophila*, over 80 kb. At the time, there weren't many other sequencing projects of that scale. Fortunately, when the Sanger sequencing technique came along, I was already very familiar with aspects of the protocol because my work on DNA replication required that I routinely used dideoxynucleotides. So I switched over to the much more efficient, revolutionary Sanger method at a time when other Caltech laboratories were still using Maxam–Gilbert sequencing.

In 1995, in collaboration with Dr. Mike Palazzolo's group at Lawrence Berkeley National Laboratory (LBNL), we published the sequence of the bithorax complex (BX-C) (Martin *et al.* 1995), some 338,234 bp, and that enormous achievement was, in fact, the start of the *Drosophila* genome sequencing project. When Mike and Gerry Rubin initiated the BDGP to sequence the *Drosophila* genome in its entirety, they recruited me to work with them, first as a visiting faculty and then as a Staff Scientist. Once they both moved on to pursue other scientific endeavors, I became the Principal Investigator of the BDGP.

It is fair to say that Sue's laboratory is by far the single largest contributor of genomic data to the *Drosophila* research community, and that the influence (directly and indirectly) of this information pervades the research advances of this community for the last 10–15 years.

—B. Gelbart, Harvard University, George Beadle Awardee

What Are Some Examples of Functional Genomics Resources That Came Out of the Genome Project?

We finished the sequencing project much sooner than expected, as a result of our collaboration with Dr. Craig Venter's team at Celera Genomics, and published the first draft in 2000 (Adams *et al.* 2000). Because we still had 3 years left on our grant, the National Institutes of Health allowed us to reformulate the proposal. That's how the functional genomics projects got started. This was negotiated by Gerry before he moved east to lead the Janelia Research Campus of the Howard Hughes Medical Institute.

One of our first projects was to generate expressed sequence tags and complementary DNAs (cDNAs) (Stapleton *et al.* 2002). The project has been instrumental in the annotation of *Drosophila* gene and transcript models and has produced critical reagents for many research studies over the last 15 years. We also used the cDNAs as probes for basic research, specifically to study the spatial expression patterns of the 13,000 protein-coding genes of *Drosophila* throughout embryonic development (Tomancak *et al.* 2007).

It's very gratifying when I am told by researchers how their work was highly influenced by our studies, in particular, how helpful it was to have access to lists of genes that are coexpressed in the same tissue and at the same time. For instance, investigators who study neural development, the CNS midline, and the germline have used our expression data to identify coexpressed genes, critical data needed to construct their comprehensive maps correlating development and function. The same goes for other tissues.

We also used the cDNAs to create a protein complex map in collaboration with Spyros Artavanis-Tsakonas (Guruharsha *et al.* 2011). We transferred the open reading frames (ORFs) into vectors allowing expression of FLAG-HA-tagged proteins and Spyros's group used them for immunoprecipitation to identify protein complexes. In the process, we generated a large set of vectors that allowed researchers to add a variety of tags and fluorescent markers

onto proteins of interest and to express specific cDNAs in tissue culture and in flies.

The Fly Community Has a Reputation for Sharing. How Important Has This Been for its Success?

There is a long tradition of sharing among *Drosophila* researchers, going back to the inception of the fly community when Bridges and Sturtevant (both students of Thomas Hunt Morgan) set up the first fly stock center. Bridges distributed stocks to anyone who wanted them, and that was a really important part of growing the *Drosophila* research community. We wouldn't know nearly as much about the field of genetics without the *Drosophila* research community's tradition of sharing fly stocks. The same is true today. All of the cDNAs that we generate go to the *Drosophila* Genomics Resource Center and are distributed for a very small fee without restrictions of any kind.

The culture of exchanging resources doesn't only help fly researchers, it also has a direct impact on medicine. There are some large ongoing efforts underway to understand human disease-associated genes. For example, we are involved with Hugo Bellen's efforts to take the mammalian gene collection and move it into vectors for expression in flies. This project, as part of the Undiagnosed Disease Network, will allow us to rapidly study disease-related mutations. Ross Cagan is using *Drosophila* genomics to make cancer models and screen for therapeutics. There are also projects on diabetes and other chronic diseases. It's all pretty spectacular!

What Have Been Some of the Surprises in the Postgenome Era?

I think the public and the broader scientific community have been most surprised by the extent to which genes are conserved in the animal kingdom. One of the first papers from the project published in 2000 showed just how many proteins were conserved across a series of animals, from flies to worms to humans. The genome sequence has also led to a whole revolution in RNA biology, because knowing the sequence allows you to create probes to characterize transcripts. When we sequenced the fly transcriptome with modENCODE (Graveley *et al.* 2011; Brown *et al.* 2014), there were definite surprises, like the number of transcript variants that exist and the correlation of alternative polyadenylation with gene expression in the brain. In the EST and cDNA project, we could not look at very small transcripts or very rare ones because they were undetectable. But now that we have the technology to capture them, a whole new world has become easily accessible, just waiting to be discovered.

The revolutions in our understanding of genetics at the molecular level over the past 40 years have been primarily driven by technological changes; moving from slab gels to capillary gels to next-generation sequencing, the pace of sequencing advances alone is incredible, even faster than Moore's law (the impressive rate at which computer chips get smaller and more powerful over time). Being able to participate in those revolutions has been amazing for me.

How Did modENCODE Relate to the Human ENCODE Project?

Encode stands for Encyclopedia of DNA Elements, and modENCODE was the model organism equivalent. The pilot human project covered only 1% of the genome, so it was necessary to find efficient ways to scale up. It became clear that model organisms with smaller genomes (flies and worms) could be very useful. The goal was to define the sequence-based function of every base in each genome, *i.e.*, does it make an RNA, does it control RNA expression, or is it involved in chromatin formation and so on, and apply that knowledge to the Human genome project.

We pioneered technologies later adopted by the ENCODE Consortium, including the early application of stranded RNA sequencing. Because we already had a more complete genome annotation than was available in human, we were able to construct automated genome annotation pipelines based on sequencing data. So certainly some of the expertise and lessons from modENCODE informed the human project.

However, it's hard to make a direct comparison between the two projects because the human project started out by studying cell lines, whereas the modENCODE project was focused on tissues and whole animals. I think the power of modENCODE to identify new developmentally-expressed genes has demonstrated the importance of studying tissues and organ systems in particular, and model systems in general. Clearly many of the rules of how genes are controlled in time and space — the discovery of enhancers and silencers and how they work — were first figured out in model organisms! Ben Brown, a talented Bioinformatician and LBNL colleague, and I recently completed an extensive (although not entirely comprehensive) review of the many astounding contributions of the modENCODE Consortium, far beyond the work within my group, and these are reviewed in Brown and Celniker (2015).

What's Next for Your Lab?

Because of our affiliation with the Lawrence Berkeley National Labs, it is only natural that we share their interest in the environment. So, we have started a project looking at the effects of environmental perturbagens on the *Drosophila* and mouse host transcriptomes and metabolomes and their associated microbiota. As part of that effort, we sequenced the genomes of a number of *Drosophila* microbes with PacBio and Oxford Nanopore technology. The results demonstrate spectacular changes in the composition of the microbiota (as well as effects on the host) when you treat these flies with atrazine, a controversial herbicide that is heavily used in the United States. We see that germ-free flies die from acute atrazine exposure. These flies can be rescued by the transplantation of certain essential microbes, meaning that their microbiota is essential for metabolizing the herbicide. We are

only in the early stages of this project, but it is already clear from preliminary data that the microbiome plays a major role in metabolizing environmental perturbagens. We have a parallel project in mice using the Collaborative Cross, which is an incredible mouse collection made from a set of eight founder inbred lines that were then crossed to generate 30 different strains capturing much of the variation found in laboratory strains. This will let us test the combined effects of host genetics and the microbiome in response to environmental challenges. Our most important contribution at this early stage of the investigation has been to generate high-quality microbial genome assemblies. We've identified thousands of additional genes, new viruses, and plasmids that had not been described because early sequencing efforts generated assemblies with thousands of contigs using second-, rather than third-generation sequencing technologies.

However, we're not moving entirely away from the nuclear genome. There are still parts of every metazoan genome that have yet to be completed. For instance, there's a large gap in our understanding of the telomeric and centromeric regions; they remain unfinished in both flies and humans. However, Release 6 of the fly genome, published in 2015, contains significant improvements in the heterochromatin regions; we discovered new genes that we never knew existed. Improving reference sequences is extremely important, and it's getting easier to do with new long-read strategies like PacBio.

We are also hoping to study genes encoding small ORFs that are highly conserved. There's a big debate going on about whether regulatory long noncoding RNAs actually encode very small peptides. One of the ways to move the discussion about putative small ORFs forward is to focus on those that are conserved. We found a little over 200 small ORFs encoded in long noncoding RNA that are highly conserved. In collaboration with Norbert Perrimon, we'd like to conduct studies to determine their function. The simple fact that these peptides are conserved throughout the animal kingdom strongly suggests that they play an important role in organismal development. Studying these could lead to the identification of new modes of action for short polypeptides, and enable the annotation of functional ORFs in less conserved regions of the genome. Our environmental studies have revealed that long noncoding RNAs, which of course may ultimately turn out to encode short ORFs, are among the most stress responsive genes in the *Drosophila* genome, and we know nothing about the functions of these RNAs beyond the fact that they respond to particular stressors. Are these incidental enhancer RNAs? Are they functional molecules? We don't know. Broader questions concern the mechanisms of orchestration of complex spatiotemporal expression patterns; it is amazing how little we understand about the specification of individual cell lineages within tissues, or how cellular responses to environmental challenges lead to emergent tissue and organ system activities that manifest, ultimately, in adaption or toxicity. Our view of the genome is still remarkably static. One major transition I expect over the next 5 years or so, and in which my lab will play a role, is the generation of highly detailed molecular time courses enabling truly dynamical models of development and environmental responses.

There's a lot of mystery left in the genome! And *Drosophila* continues to be one our most powerful system vessels for exploration and discovery.

Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science 287: 2185–2195.

Brown, J. B., and S. E. Celniker, 2015 Lessons from modENCODE. Annu. Rev. Genomics Hum. Genet. 16: 31–53.

Brown, J. B., N. Boley, R. Eisman, G. E. May, M. H. Stoiber *et al.*, 2014 Diversity and dynamics of the Drosophila transcriptome. Nature 512: 393–399.

Celniker, S. E., K. Sweder, F. Srienc, J. E. Bailey, and J. L. Campbell, 1984 Deletion mutations affecting autonomously replicating sequence ARS1 of Saccharomyces cerevisiae. Mol. Cell. Biol. 4: 2455–2466.

Celniker, S. E., D. J. Keelan, and E. B. Lewis, 1989 The molecular genetics of the bithorax complex of Drosophila: characterization of the products of the abdominal-B domain. Genes Dev. 3: 1424–1436.

Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin *et al.*, 2011 The developmental transcriptome of Drosophila melanogaster. Nature 471: 473–479.

Guruharsha, K. G., J. F. Rual, B. Zhai, J. Mintseris, P. Vaidya *et al.*, 2011 A protein complex network of Drosophila melanogaster. Cell 147: 690–703.

Martin, C. H., C. A. Mayeda, C. A. Davis, C. L. Ericsson, J. D. Knafels *et al.*, 1995 Complete sequence of the bithorax complex of Drosophila. Proc. Natl. Acad. Sci. USA 92: 8398–8402.

Stapleton, M., J. Carlson, P. Brokstein, C. Yu, M. Champe *et al.*, 2002 A Drosophila full-length cDNA resource. Genome Biol. 3: RESEARCH0080.

Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan *et al.*, 2007 Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol. 8: R145.

Communicating editor: C. Gelling