Meeting report

Recent advances in *Drosophila* genomics

Melissa B Davis and Kevin P White

Address: Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA.

Correspondence: Kevin P White. E-mail: Kevin.white@yale.edu

Published: 28 July 2004

Genome ***Biology*** 2004, **5:**339

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/8/339

A report on the 45th annual *Drosophila* Research Conference, Washington DC, USA, 23-28 March 2004.

The 45th annual *Drosophila* Research Conference convened in Washington DC with over 1,600 attendees. A very broad spectrum of *Drosophila* research was represented at the meeting, and this report highlights some thought-provoking presentations and provides an update on the status of genomic resources for the fly community. Francis Collins (National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, USA) addressed this year's attendees with a challenge to use the *Drosophila melanogaster* genome as a model to develop new technologies and approaches for genomic research that can be extended to the human genome. He pointed out that the *Drosophila* community is positioned to pioneer the technological advances necessary for the study of complex genomes, as is inherent in the new and ongoing 'genomic era' that began with the successes of the sequencing and annotation projects.

As an example, Collins described the effectiveness of using comparative genomics to enhance the functional annotation of genomes. He reported that whole genome sequences for several additional species of *Drosophila* will soon be available through a series of projects sponsored by the NHGRI. At present, sequence has already been generated for *Drosophila pseudoobscura*, *Drosophila yakuba* and *Drosophila ananassae*, and the sequences of *Drosophila simulans*, *Drosophila erecta, Drosophila willistoni, Drosophila grimshawi, Drosophila mojavensis, Drosophila virilis, Drosophila persimilis* and *Drosophila sechellia* are expected to be completed within the next two years. These projects will provide the comparative data necessary for the development of more efficient computational tools to identify regulatory sequences, gene structures and novel functional elements within genomes. Collins also suggested that *Drosophila* will be an ideal testing ground for the methods being used in the NHGRI's encyclopedia of DNA elements (ENCODE) project, which currently focuses on defining the functional elements in the human genome. Accomplishing complete functional annotation of the *Drosophila* genome, an effort that has arguably already begun, will add new dimensions to every area of *Drosophila* research, while providing a critical assessment of computational and experimental approaches that may later be applied to the entire human genome.

Genome annotation

Progress reports were given at the meeting on the annotation efforts for the *D. melanogaster* and *D. pseudoobscura* genomes. Susan Celniker (Lawrence Berkeley National Laboratory, Berkeley, USA) gave an update on the sequencing and experimental annotation of the *D. melanogaster* genome, currently available as release 3.2. The present phase of annotation comprises fine-tuning the details of gene structures, experimentally confirming the computationally identified genes and filling in the few lingering sequence gaps. Celniker reported that "the most important endeavor is an attempt to obtain and sequence full-length cDNA clones for all *D. melanogaster* genes", a project referred to as the *Drosophila* Gene Collection (DGC). The DGC is being coordinated by Mark Stapleton (also at Lawrence Berkeley National Laboratory) and now contains over 10,000 clones with full-length open reading frames (ORFs). This clone set currently has nearly 6,000 clones whose DNA sequence has been confirmed, and these are already being used for a large-scale embryo *in situ* hybridization project, led by Pavel Tomancak and Gerry Rubin (both at University of California, Berkeley, USA). All of the *in situ* hybridization results are available and can be queried and browsed on a website accessible via FlyBase [http://www.fruitfly.org/cgi-bin/ex/insitu.pl], enabling any researcher to retrieve easily not only the sequence and structure of their favorite gene but also the gene's spatial expression

patterns in the embryo. Coupled with the temporal expression patterns throughout the complete life cycle that are also currently available (The *Drosophila* Developmental Gene Expression Timecourse [http://flygenome.yale.edu/Lifecycle/]), the embryonic expression data will help expedite the direct functional analysis of gene expression for the entire gene complement in *Drosophila*. Additionally, the DGC will provide a starting point for research projects in proteomics because all of the ORFs will eventually be inserted into 'high-throughput' cloning vectors such as the GateWay or InFusion systems.

Following the update on the progress of the *D. melanogaster* annotation and clone resources, Stephen Richards (Baylor College of Medicine, Houston, USA) discussed the status of the sequencing and annotation efforts for *D. pseudoobscura,* which are underway in a collaboration between several labs at Baylor College of Medicine, FlyBase, and Pennsylvania State, Cornell, Yale and Harvard Universities. Annotation efforts for this genome are more limited than the intensive *D. melanogaster* annotations, and are largely driven by computational algorithms such as Genescan/Twinscan and Genewise. Considerable efforts are, however, in progress to produce sequence-alignment maps between the *D. pseudoobscura* and *D. melanogaster* genomes. Analysis of these alignments has shown, for example, that known *D. melanogaster* transcription-factor-binding sites are slightly more conserved than other non-coding sequences, despite the inherent flexibility of binding-site sequences. In addition, a comparison of the genomic structure of the two species demonstrated that a large number of paracentric inversions have occurred since their divergence. The remnants of a novel repetitive element were found at a significant number of inversion breakpoints, suggesting that a transposon-mediated mechanism is responsible for recent paracentric inversions. This work will set the stage for comparative genomics studies that are in the pipeline.

Traditional cDNA microarray studies are now being supplemented with genome-scale expression studies that use high-density microarrays composed of sequences representing the entire genome, as opposed to only portions of predicted transcripts. Zareen Gauhar and Chris Mason (both at Yale University, New Haven, USA) presented the results of expression profiling each developmental stage of the *Drosophila* life cycle using high-density genomic 'tiling' arrays. These arrays have probes for all predicted exons and intergenic sequences in the genome, as well as for each potential splice junction for a subset of specific genes, and were constructed through collaboration with Victor Stolc (NASA Ames Research Center, Mountain View, USA). To build the arrays they employed a maskless array technology developed by Michael Sussman (University of Wisconsin, Madison, USA) and Nimblegen, Inc. Among other findings, Gauhar and Mason discovered that more than 50% of the sequences outside predicted exons are expressed during at least one stage in development. They also provided data supporting the existence of approximately 800 transcripts that correspond to ORFs predicted by gene-finding algorithms but not present in the current genome annotation. By assaying splice junctions they found that approximately half of the genes in the genome exhibit alternative splicing. These studies take an important step towards accurately defining the complete set of expressed transcripts in the *Drosophila* genome. Gauhar and Mason also mentioned an analogous study in *D. pseudoobscura*, which should help to address the divergence in expression of gene isoforms throughout the life cycle of the two species.

Advances in systematic gene disruption

Several presentations described efforts to merge large-scale mutagenesis with functional genomics, so as to create a growing catalog of mutant lines. The common goal of these projects is to create at least one mutant allele for each predicted gene. The most advanced systematic mutagenesis project is the *Drosophila* Gene Disruption Project, which continues as a collaborative effort between Rubin, Allan Spradling (Carnegie Institution, Baltimore, USA), Hugo Bellin (Baylor College of Medicine) and Roger Hoskins (Lawrence Berkeley National Laboratory). They are striving to disrupt each annotated gene with a single transposon insertion. Exelixis Inc. has made a large contribution to this endeavor this year, producing over 2,000 mutant lines from over 29,000 inserts created with P and PiggyBac elements. Some of the PiggyBac elements used in the Exelixis screens were engineered to yield mutants with stronger phenotypes. Several adaptations were made to the elements, including the incorporation of 'splice traps' that use splicing acceptor sites to create mis-spliced endogenous transcripts when an element inserts into an intron, and suppressor functional elements to silence the transcription of the targeted genes in the case of promoter insertions. To date, the public effort has yielded over 30,000 P-element insertion lines that have been localized and analyzed, resulting in a grand total of 7,140 mutant lines harboring disruptions in more than 5,300 of the roughly 15,000 predicted genes in the *D. melanogaster* genome. These mutant lines represent 35% of the genome and the number is steadily growing. But Spradling emphasized that the gene disruption project, using P element and PiggyBac mobile elements, will only produce tags or starting points for creating lesions in a maximum of about 70% of the genes because some loci are resistant to transposable element mutagenesis.

An alternative method of large-scale mutagenesis was reported at the meeting: targeting induced local lesions in genomes (TILLING). Bradley Till (Fred Hutchison Cancer Research Center, Seattle, USA) presented the preliminary results from a TILLING study underway in collaboration between Steve Henikoff's group (Fred Hutchinson Research Center) and Barbara Wakimoto (University of Washington,

Seattle, USA). Till discussed their plans to implement a 'special order' mutant service for the *Drosophila* community, modeled on a service already established in the *Arabidopsis* community. Animals from chemically mutagenized lines are pooled and gene-specific PCR screening methods are used to create nicked heteroduplex products that indicate when mutations in a gene of interest have been created. Individuals from positive pools are then scored for the mutation using the same PCR technique, and the mutant line is isolated. This process is expected to be available on a fee-for-service basis in the next year from the TILLING consortium.

Although individual gene disruptions are powerful reagents, for decades *Drosophila* researchers have relied on chromosomal rearrangements and deficiencies for mapping mutations and analyzing gene function. The caveat to most of the deficiency lines that are presently available is the lack of sufficient resolution of their relative deletion endpoints, such that we are often unsure of the presence or absence of particular genes in the deleted regions. This uncertainty of the exact genomic structure becomes problematic when we attempt to attribute a loss-of-function phenotype to a specific gene mutation within a chromosomal aberration. Eric Spana (Duke University, Durham, USA) presented preliminary results of a microarray approach designed to map the deletion endpoints of the commonly used chromosomal deficiency (Df) stocks. His group has incorporated a scheme whereby they isolate and dual-color-label genomic DNA from both a wild-type line and a Df line; the fluorescent probes are then hybridized to an array of long oligos (70mers) that correspond to each cDNA in the genome. On the basis of the fluorescence intensity ratios recovered from the microarray analysis, Spana and colleagues are able to determine whether a transcript or gene is present as a single copy - falls within the deletion region - or in two copies (and is therefore outside the deletion region). Thus far, this approach has been successfully used to map the deletion endpoints of several lines.

Several projects are underway using large-scale RNA interference (RNAi) screens to identify novel genes that function in specific pathways. These projects use a variety of cell-based assays to study the effects of systematically 'knocking down' target gene expression. The screens rely on observing and quantifying phenotypes such as cell morphology, proliferation, immuno-histochemical observations, or on more sophisticated reporter-gene expression analysis. For example, Edan Foley (University of California, San Francisco, USA) presented a screen for regulators of the innate immune response that revealed a novel regulator, Defense repressor 1 (Dnr1), which functions to keep signaling in check in the absence of infection. Sara Cherry (Harvard University, Cambridge, USA) described a genome-wide screen for host factors affecting viral pathogenesis; this showed that approximately 1% of the genome is co-opted by viruses for replication. Another screen, for cell-cycle regulators, was presented by Rita Sinka, (University of Cambridge, UK), and is still in its initial stages but has already established that approximately 4% of the genes examined cause significant changes in cell-cycle progression as measured by a mitotic index. In each case the investigators discovered dozens of novel players in their respective genetic pathways, a feat that would normally have taken multiple researcher-years to accomplish using a traditional mutant-screening method. These presentations provide proof of principle that large-scale RNAi screens are very efficient for discovering gene function. This technology is spreading rapidly thanks to efforts by Norbert Perrimon (Harvard University) who has set up a *Drosophila* RNAi Screening Center [http://flyrnai.org/], and by Pat O'Farrell (University of California, San Francisco) who introduced a resource for generating double-stranded RNAs targeted against genes that are conserved between flies and humans.

User-friendly data

Making lateral comparisons between functional genomics datasets can be a difficult challenge. As datasets grow, they become increasingly arduous to interpret efficiently. Varied formats and methods make the complexity of data-mining a challenge, and in some cases an impossibility, for many researchers. Unfortunately, this reality undermines the usefulness of the enormous data sets being produced. Several efforts have begun that attempt to solve this problem. Gos Micklem and Michael Ashburner (both at University of Cambridge) are spearheading a new endeavor to curate genomic data and to create an integrated database of insect gene-expression, gene ontology, RNAi, proteomics and protein three-dimensional structure data that are being obtained primarily for *Drosophila* and *Anopheles*. A searchable database, called FlyMine [http://www.flymine.org], will be available for queries from either a web interface or a programming interface and is intended to be integrated with the curated data available through FlyBase. FlyMine aims to give researchers from various scientific disciplines access to multiple genomics datasets that are directly linked to annotation databases.

Another effort, FlyExpress [http://www.azbio.org/efg/news/03082001.html], is underway through collaborative efforts at the Center for Evolutionary Functional Genomics at Arizona State University under the direction of Sudhir Kumar (Arizona State University, Tempe, USA). Kumar and collaborators are developing bioinformatic approaches to allow researchers to query *in situ* hybridization images from published data that document wild-type gene-expression patterns. Initially, this effort seeks to curate published embryonic gene-expression patterns and to standardize the images via computational means in order to normalize for image variations that result from illumination and orientation differences. Subsequently, they plan to develop statistical methods to quantify dissimilarity between patterns and

to build image-retrieval systems that will efficiently recognize and categorize matching expression patterns. Such efforts will enable researchers to scan large numbers of images for particular expression patterns and/or expression overlaps that may suggest specific genetic interactions.

Ongoing efforts in the areas of computational and experimental genome annotation, mutational analysis and high-throughput mutant screening, as well as database construction and data integration were the major genomics themes of the meeting. Continuous and rapid progress in these areas is expected to create a foundation that will stimulate new biological discoveries in *Drosophila*, while simultaneously enhancing our basic understanding of complex genomes as we continue to move forward into the 'genomic era'.