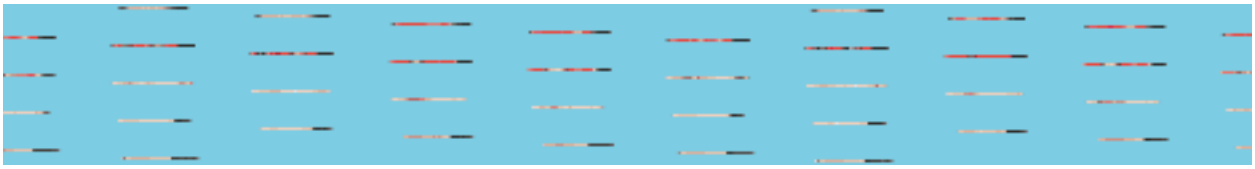
A Biologist's View of the *Drosophila* Genome Annotation Assessment Project

Michael Ashburner

Genome Res. 2000 10: 391-393
Access the most recent version at doi:10.1101/gr.10.4.391



A Biologist's View of the *Drosophila* Genome Annotation Assessment Project

Michael Ashburner[1]

Department of Genetics, University of Cambridge, Cambridge, England; European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Large-scale genomic sequencing projects are acts of faith—the faith that given a sequence of many millions of As, Ts, Gs, and Cs, we have the analytical tools to make sense of it. This faith can only be justified if the results of sequence analysis are tested continually against reality. The Genome Annotation Assessment Project (GASP) experiment, which took place in May, 1999, was one such test (see, in this issue, Birney and Durbin 2000; Gaasterland et al. 2000; Henikoff and Henikoff 2000; Krogh 2000; Ohler 2000; Parra et al. 2000; Reese et al. 2000a,b; Salamov and Solovyev 2000). GASP compared the interpretation of a 2.9-Mb sequence made by a mix of computation and human analysis done over a period of 2 years (Ashburner et al. 1999) with those done by wholly computational procedures carried out over a period of 6 weeks. Interestingly, GASP was an experiment within an experiment. The overt experiment was to assess the performance of a number of different analytical tools; the covert experiment was to assess how such an assessment could be done.

Computation Analysis in Conjunction with Experimental Analysis

The 2.9-Mb test sequence, determined by the Berkeley *Drosophila* Genome Project (BDGP), is known as the *Adh* region of the genome of *Drosophila melanogaster*, named after the gene encoding the enzyme alcohol dehydrogenase about which the region is centered. Ashburner and colleagues (e.g., Woodruff and Ashburner 1979a,b) had begun studying the *Adh* gene of *D. melanogaster* in the late 1970s; at the time it was the only gene in *Drosophila* for which one could select both loss-of-function mutations (by their survival on unsaturated alcohols) and reversions to wild type (by their survival on ethanol). The work on the *Adh* gene gradually recruited genes in its neighborhood and, over a period of some 20 years, produced as detailed a picture of the structure of this region as possible by classical genetics. Although one can never be certain about such matters, the extensive nature of the work suggested that we had identified, and mapped, all of the genes in this region that gave a detectable phenotype when mutant. The number of genes was 73, of which 53 were known to be essential for viability. In addition to >1000 mutations, we had also induced and mapped well over 500 chromosome aberrations. The *Adh* region represents ~1.5% of the genome of *D. melanogaster*: These genetic studies and some hobby sequencing made *Adh* the best known region of the genome. It was for this reason that the BDGP considered that its analysis would be a valuable test of the longer-term strategy of sequencing and annotating the entire genome of this fly.

The objective of this annotation procedure was twofold: to describe the structure of all of the genes (and other sequence features such as transposable elements) correctly in the DNA and to infer their functions. Because many of the genes were known previously, either (or both) from their mutations and sequence, one priority had to be the identification of each on the genomic sequence itself. For those genes that had been sequenced before this project, that was (reasonably) easy; for those that had not, P-element insertions, which were mapped both genetically and on the genomic sequence, proved to be invaluable, although not infallible in identifying genes. P elements preferentially insert in the 5′ noncoding regions of genes. A P element located between a pair of divergently transcribed genes may mutate one or both, and only genetic experiments can distinguish between these possibilities. This identifies one of the major features of the initial *Adh* region annotation, which distinguishes it from all of the GASP experiments and from most other large-scale genome annotation experiments: The original annotation involved, over a period of ~2 years, an intense collaboration among those doing the computational analysis (largely in Berkeley), geneticists in Cambridge, and many other experimentalists in the *Drosophila* community. Following are two examples of this collaboration, to illustrate a very important principle.

To discover how to annotate genomic sequence, the BDGP first analyzed the sequence of one 83-kb P1 clone. This clone (DS02740) is particularly gene rich and included several genes that had at that time already been sequenced by *Drosophila* researchers. However, it also included several genes that were not known but that were predicted by the computational tools then used—`Genefinder` and similarity matches using `BLAST` tools. Were these predictions real? The difficulty is that no computational experiment can validate them. However, there are two classes of validating experiments that would convince a biologist: The first is genetic—mapping a phenotypically characterized mutation to a predicted gene; the second is molecular—isolate and sequence

[1]E-MAIL ma11@gen.cam.ac.uk; FAX 011-44-1223-333-992.

10:391–393 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

a full-length cDNA for the prediction. This would convince most people of the reality of the prediction. The more hard-nosed biologist might demand evidence that the cDNA is actually translated into protein. In fact the BDGP validated their methods by isolating and sequencing cDNAs for all newly predicted genes on this P1 clone. That gave reasonable confidence that the computational methods being used were not returning fantasies.

The second example is rather more parochial, but it illustrates just how difficult annotation can be, even in a very well-studied chromosome region. In P1 clone DS01523, `Genefinder` had predicted the existence of two coding regions just over 5 kb apart; that these were exons of one gene was indicated by both an EST sequence and `BLAST` matches of both sequences to the same targets. These `BLAST` matches indicated that the gene encoded a DYRK-family kinase. In addition to the identification of the presence of this predicted *DYRK*-like gene, we had three genetically characterized P elements, all associated with a subtle phenotype; the mutant flies were impaired in their ability to smell benzaldehyde (the first of these had been discovered and characterized previously (Anolt et al. 1996)). The question was, Which gene (or genes) did these mutations effect? One of these mutations mapped >35 kb, 5′ to the *DYRK*-like gene; two others mapped >65 kb, 5′ to this gene. The latter two mapped closer to the next characterized gene, *wing blister*, encoding a laminin α-chain, than to the *DYRK*-like gene. To show that the *DYRK* gene is responsible for the olfactory phenotype required a combination of genetic analyses of all available P elements (done by R. Anholt) and the isolation and sequencing of a new cDNA (done in Berkeley). The latter showed that one of these P elements is inserted in a computationally undetected 5′ exon, some 35 kb upstream of the predicted coding exons. In the absence of the genetic information, there would have been no reason to suspect that the structure of the *DYRK* gene was any more complex than indicated by the prediction programs.

A View of the GASP Results

Many of the results of the GASP experiment were unsurprising, including the inability of any ab initio method to predict 5′- and 3′-untranslated exons, the tendency of ab initio and even homology-based methods to artificially join or split genes (the former being a particular problem when genes are tandemly duplicated), and the inability of any ab initio method to cope with overlapping genes (which are surprisingly common in *Drosophila*). What was unexpected, at least to me, was the problem that some programs had with EST data. As Reese et al. (2000a) point out in their summary of GASP, the alignment of EST and genomic sequences is not a solved problem. Another lesson—one that we had learned during the original annotation—comes from the work to build the std1 dataset; both EST and cDNA sequences can be misleading. As essential for annotation as they are, such sequences must always be evaluated for errors. They may be incorrect for one of several reasons: they may be genomic DNA contaminants, they may be primed off an internal genomic poly(A) sequence, or they may reflect abnormal or intermediate splice forms of a pre-mRNA.

Several of the ab initio methods used by GASP participants predict genes that were not predicted by Ashburner et al. (1999). It is important for an evaluation of the GASP experiment to know whether or not these predictions are false positives. To this end the BDGP is now systematically screening for cDNAs for each of these newly predicted genes. When these experiments are completed, we will have a much more secure basis for evaluating these GASP submissions. I regard the issue of false positives and false negatives as being the most serious problem for exon and gene prediction methods. The extent of this problem was seen in GASP and, more recently, by a comparison of two methods for gene prediction used by Celera Genomics for the *Drosophila* sequence (Adams et al. 2000): `Genefinder` and `Genie` predicted ~17,500 and ~13,000 genes respectively. The reason why the problem is serious is that false predictions are very time consuming (and, hence, expensive) to refute; furthermore, such refutation may be impossible in the formal sense. Consider a prediction of a gene by an ab initio method with no supporting evidence. Extensive searches for transcripts fail; nothing of similar sequence is found in the public universe of sequences. Now consider that under regular laboratory conditions only one-fourth or so of the genes of *Escherichia coli* are transcribed in log-phase growth, and the severity of the problem is obvious to all—who can refute the hypothesis that a gene is only transcribed in four brain cells (two on each side) during mating?

One obvious conclusion from GASP is that several methods are better than one. Having said that one must immediately guard against the tyranny of the majority, a prediction is not strengthened if it is made by more than one method, unless the methods are independent. Determining the independence of methods may actually become more difficult as prediction programs incorporate more information for their tasks. A second conclusion is that the task of managing predictions from several different programs within an environment that will allow sensible data synthesis is very far from trivial. Standard output methods, such as `GFF` (http://www.sanger.ac.uk/Software/formats/GFF/), and visual tools that can use these as input are essential, and these visual tools must be accessible to biologists.

Chromosomes are, of course, much greater than the sum of their genes. GASP concentrated on the prediction of genes; only three groups looked at other features—transcription start sites (Ohler 2000) and sequence repeats (Benson 1999; Gaasterland et al. 2000). One opportunity offered by the *Drosophila* sequence is to study sequence features responsible for the structure of the interphase chromosome. The 2.9-Mb *Adh* region covers 69 polytene chromosome bands, and methods to analyze long sequences for characteristics that might determine chromosome structure are urgently needed.

The annotation of genomic sequences has often been contentious; see, for example, the rather bad-tempered arguments about the use of the `GlimmerM` program for the analysis of *Plasmodium* sequences (Lawson et al. 2000; Pertea et al. 2000) and the criticisms made of the preliminary annotations of the *Arabidopsis* genome (Pavy et al. 1999). There should be competition between methods, but there needs to be both better collaboration between those developing these methods and between annotators and "wet" experimentalists. This is certainly the major lesson I would take away from this experience.

I hesitate to conclude that I have better insight into the biology of the *Adh* region as a consequence of GASP. GASP has provided a number of predictions that were missed by Ashburner et al. (1999); in retrospect, I think that the analysis by Ashburner et al. may have been too conservative—indicated by the fact that the gene density in the *Adh* region, as estimated by Ashburner et al. (1999), is considerably lower than that estimated for the genome as a whole by Adams et al. (2000), that is, one gene in 13 kb vs. one gene in 9.5 kb. These new predictions and variant models of genes that had been predicted by Ashburner et al. (1999) are being tested experimentally. The most rapid way to do this is by sequencing full-length cDNAs, selected using primers designed against the predicted gene sequences.

Large genomic sequences can only be analyzed computationally. The style of annotation of Ashburner et al. (1999) cannot scale to the whole genome, even of an organism as small as *Drosophila*. For this reason, it is very important that independent assessments of annotation methods be made. It is unlikely that a blind experiment such as that done for GASP can be repeated. This means that assessment of annotation can only be done by third parties and not by the developers of the programs themselves. This may be very difficult to do if source codes are not made available freely. The free availability of well-documented sequence sets are also required, whose features are reliably annotated as being predicted or experimentally determined.

REFERENCES

Adams, M., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S. Scherer, P.W. Li, R.F. Galle, R.A. George et al. 2000. *Science* **287:** 2185–2195.

Anolt, R.R.H., R. Lyman, and T.F.C. MacKay. 1996. *Genetics* **143:** 293-301.

Ashburner, M., S. Misra, J. Roote, S.E. Lewis, R. Blazej, T. Davis, C. Doyle, R. Galle, R. George, N. Harris et al. 1999. *Genetics* **153:** 179–219.

Benson, G. 1999. *Nucleic Acids Res.* **27:** 573–580.

Birney, E. and R. Durbin. 2000. *Genome Res.* (this issue).

Gaasterland, T., A. Sczyrba, E. Thomas, G. Aytekin-Kurban, P. Gordon, and C.W. Sensen. 2000. *Genome Res.* (this issue).

Henikoff, J.G. and S. Henikoff. 2000. *Genome Res.* (this issue).

Krogh, A. 2000. *Genome Res.* (this issue).

Lawson, D., S. Bowman, and B. Barrell. 2000. *Nature* **404:** 34–35.

Ohler, U. 2000. *Genome Res.* (this issue).

Parra, G., E. Blanco, and R. Guigó. 2000. *Genome Res.* (this issue).

Pavy, N., S. Rombauts, P. Dehais, C. Mathe, D.V.V. Ramana, P. Leroy, and P. Rouze. 1999. *Bioinformatics* **15:** 887–899.

Pertea, M., S.L. Salzberg, and M.J. Gardner. 2000. *Nature* **404:** 34.

Reese, M.G., G. Hartzell, N.L. Harris, U. Ohler, and S.E. Lewis. 2000a. *Genome Res.* (this issue).

Reese, M.G., D. Kulp, H. Tammana, and D. Haussler. 2000b. *Genome Res.* (this issue).

Salamov, A.A. and V.V. Solovyev. 2000. *Genome Res.* (this issue).

Woodruff, R.C. and M. Ashburner. 1979a. *Genetics* **143:** 117-132.

Woodruff, R.C. and M. Ashburner. 1979b. *Genetics* **92:** 133-149.