

How many genes in a genome?

Brian Oliver* and Benoit Leblanc[†]

Addresses: *Laboratory of Cellular and Developmental Biology, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Department of Health and Human Services, 50 South Dr. Bethesda, MD 20892-8028, USA. [†]Département de Biologie, Faculté des Sciences, Université de Sherbrooke, 2500 boulevard Université, Sherbrooke, QC J1K 2R1, Canada.

Correspondence: Brian Oliver. E-mail: oliver@helix.nih.gov

Published: 22 December 2003

Genome Biology 2003, **5**:204

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/5/1/204>

© 2003 BioMed Central Ltd

Abstract

Despite the current good level of annotation, the *Drosophila* genome still holds surprises. A recent study has added perhaps 2,000 genes to the predicted total, and raises a number of questions about how genome annotation data should be stored and presented.

As sequenced and assembled whole genomes first began to appear in earnest, there was much discussion about the number of genes in those genomes, usually accompanied by comments about the surprisingly low numbers of genes. Just how fuzzy those numbers are is not generally appreciated. The well-annotated *Drosophila* genome [1,2] is a blessing for *Drosophila* scientists, who make choices every day on the basis of predicted genes - from picking the exons to sequence in the hunt for the genetic lesion in a favorite mutant, to designing elements for a microarray. As good as this annotation is, Hild *et al.* [3] show, in this issue of *Genome Biology*, that we still have no clear idea how many genes there are in *Drosophila*. This should be a little sobering, as the picture for most other sequenced genomes is even less clear.

The goal of annotation is to map features on the genome, initially focusing on developing models for genes that encode proteins. Good annotation requires an assembled sequence and a repository of the evidence for important genome features such as transcripts and sequence homologies to known genes. The annotation itself adds critical and explanatory notes to the genome. Thus, annotation is an executive decision about the relevancy, accuracy, and quality of the evidence, and by definition exposes the curator's point of view. The current *Drosophila* genome annotation (Release 3.1, housed at FlyBase [4]) is conservative. The Hild *et al.* [3] annotation is not.

Hild *et al.* [3] used a more loosely tuned gene-finding algorithm than previous annotations, and in total this generated around 22,000 gene models, including nearly all of the approximately 14,000 Release 3.1 genes. It follows that the price one must pay for exposing more of the genes is the generation of more false gene models, in a classical sensitivity/specificity tradeoff. In order to test the more loosely generated models systematically, Hild *et al.* amplified a genomic region corresponding to each model and used the amplicons as elements on an array to probe for expressed RNAs. They then asked how many of the predicted genes produce transcripts. Microarrays are not sufficiently sensitive to detect every real transcript, and detection of a signal is not always definitive, but detection is very strong evidence in support of RNA synthesis directed by the genome segment in question. Using this metric, around 75% of the predicted genes common to Release 3.1 and to the study by Hild *et al.*, and around 50% of the predicted genes unique to Hild *et al.*, are transcribed at some point in the *Drosophila* life cycle. Spot-checking by reverse-transcriptase-coupled PCR and *in situ* hybridization suggests that there are no systematic problems with the array results. Thus, these data strongly suggest that there are many transcribed regions of the genome that fall outside of the Release 3.1 predictions. The lower detection frequency in the Hild *et al.* unique set than in the set shared with Release 3.1 also indicates that there is more 'chaff' as one loosens the gene calling.

While finding a transcript is good evidence for the presence of a gene, not all transcripts are from genes - depending on what you call a gene [5], the range of transcriptional noise, and a host of other debatable points. While deciding what qualifies as a gene is non-trivial, there are a number of ways to assay for functional importance. A particularly stringent phenotypic test involves asking whether a given transcript is required for cell viability. The amplicons used in the Hild *et al.* microarray form a core set of reagents for genome-wide assays for phenotypes by RNA interference (RNAi) at a newly opened screening center [6]. RNAi is a powerful method for dramatically downregulating the steady-state levels of a given transcript [7]. Systematic RNAi experiments show on tissue-culture cells that transcripts from about 3% of Release 3.1 predicted genes and approximately 1% of the transcripts from the Hild *et al.* predicted genes are required for *Drosophila* cell viability. Thus, there are genes required for the viability of tissue culture cells that evaded annotation in Release 3.1. Clearly, gene models with supporting evidence for transcription, regulated expression in space and time, and genetic function are worth annotating. On the basis of this extensive set of tests, Hild *et al.* [3] make some rough calculations and suggest that there are at least 2,000 new genes to add to the *Drosophila* total.

Finding genes without simultaneously collecting large amounts of useless information is hard. Are more genomes the solution to gene finding? The highly anticipated sequenced genomes of many related *Drosophila* species [8] will certainly be extremely important for informing the annotation of *Drosophila melanogaster* [9]. Sequence similarities and the relative ease of determining sequence quality will make comparative genomics evidence strong. But, as is pointed out by Hild *et al.*, it may not be a panacea: most of the novel predictions of Hild *et al.* do not show good sequence conservation between *Drosophila melanogaster* and other genomes, including those of insects. There are probably several reasons for this. Not all the genes in a genome evolve at the same rate or have the same sequence constraints. One can also imagine situations where the act of transcription carries the genetic function (to promote or block the access of transcription factors to DNA sites, for example). More genomes is not enough.

The biology of the organism drives the annotation of its genome. The work by Hild *et al.* on *Drosophila* and recent work on mammalian genomes clearly points out the value of experimental data in making the distinction between genes and chaff [3,10,11]. We should extend from Hild *et al.* and tackle the genome head-on. We should be using a *Drosophila* tiling-path resource (covering the whole genome with amplicons or oligonucleotides rather than sampling only the gene models) for mapping transcripts and for systematically covering the genome for function via RNAi experiments. We can also use tiling-path arrays to map the 'chromatin code' of DNA-associated proteins and the *in vivo*

occupancy of transcription factors, via procedures such as chromatin immunoprecipitation, as well as to map the replication origins. This need for more data has been recognized by the NIH, which has launched a project called the Encyclopedia of DNA Elements (ENCODE [12]) for the human genome. The main idea behind ENCODE is to develop and validate new computational and experimental means for finding genes and other important features in the human genome. The tremendous effort that goes into sequencing genomes justifies similarly large-scale efforts to map features onto the sequenced genomes.

The hunt for genes in *Drosophila* will go on and the evidence will accumulate - which is a problem in and of itself. The Hild *et al.* annotations are available on a website at Heidelberg [13] and at the Third Party Annotation database [14]. The latter is preferable, as academic and commercial websites with large datasets are not always stable, despite the good intentions of the scientists. There is also a potential problem of too many informed opinions. Multiple versions of the *Drosophila* genome annotation from this and future studies could create confusion in the user community and hinder the cross-referencing of large datasets. Perhaps FlyBase [4] should maintain the gold standard of genome annotation, displaying the high-confidence gene models (some more of which will be generated as a result of the study by Hild *et al.*) in the front window. The equally important need to access the lower confidence information, preferably with associated confidence scores, could be met either by supplying access to evidence from FlyBase, or from the Third Party Annotation database or an equivalent 'boutique' database. Regardless of how this dual requirement for conservative annotation and access to the rawer evidence is handled, the problem of data management will continue to grow, as we slowly approach knowing how many genes there are in a fly. Those facing the more daunting human genome annotation should closely watch how the *Drosophila* community approaches these problems.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.
2. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, *et al.*: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review**. *Genome Biol* 2002, **3**:research0083.1-0083.22.
3. Hild M, Beckman B, Haas SA, Koch B, Solovyev V, Busold C, Fellenberg K, Boutros M, Vingron M, Sauer F, *et al.*: **An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome**. *Genome Biol* 2003, **5**:R3.
4. **FlyBase, a database of the *Drosophila* genome** [<http://flybase.bio.indiana.edu/>]
5. Snyder M, Gerstein M: **Genomics. Defining genes in the genomics era**. *Science* 2003, **300**:258-260.
6. ***Drosophila* RNAi Screening Center** [<http://flyrna.org/>]
7. Weitzman JB: **RNAi and the shape of things to come**. *J Biol* 2003, **2**:23.

8. **NHGRI Genome Sequencing Proposals**
[http://www.genome.gov/10002154]
9. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, et al.: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002, **3**:research0086.1-0086.20.
10. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al.: **The transcriptional activity of human chromosome 22.** *Genes Dev* 2003, **17**:529-540.
11. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al.: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.** *Proc Natl Acad Sci USA* 2003, **100**:1140-1145.
12. **The ENCODE Project: ENCyclopedia Of DNA Elements**
[http://www.genome.gov/10005107]
13. **The Heidelberg Flyarray**
[http://hdflyarray.zmbh.uni-heidelberg.de/]
14. **Third Party Annotation Sequence Database**
[http://www.ncbi.nlm.nih.gov/Genbank/tpa.html]

