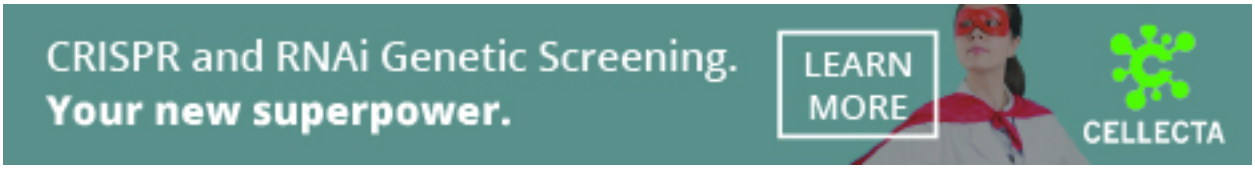
Doubling the Rewards: Testis ESTs for *Drosophila* Gene Discovery and Spermatogenesis Expression Profile Analysis

Barbara T. Wakimoto

Genome Res. 2000 10: 1841-1842
Access the most recent version at doi:10.1101/gr.169400



Doubling the Rewards: Testis ESTs for *Drosophila* Gene Discovery and Spermatogenesis Expression Profile Analysis

Barbara T. Wakimoto[1]

University of Washington, Seattle, Washington 98195, USA

One of the rewards of completing, or essentially completing, the genomic sequence of reference organisms is to get an idea of the number of genes it takes to build an organism. The current consensus is that 6142 protein-encoding genes keep the budding yeast *Saccharomyces cerevisiae* alive and well, while the worm *Caenorhabditis elegans* has on the order of 19,099 genes. It was somewhat gratifying for Drosophilists to learn this year that the fly might be able to make do in life with fewer genes than the worm does; Adams et al. (2000) predicted 13,601 genes from the sequence of the 120-Mb euchromatic portion of the *Drosophila melanogaster* genome. The remaining 60 Mb of the fly genome is heterochromatic, and most of it is unclonable. Of the proportion of heterochromatin that is cloned, only small bits have been sequenced, and these fragments cannot be easily aligned because of interruptions by repetitive sequences. However, genetic studies predict that heterochromatin will contribute at least several dozen genes, and perhaps substantially more, to the total gene count (Gatti and Pimpinelli 1992). Hence, the estimate of 13,601 is a conservative one for the gene number in *D. melanogaster*, but just how conservative it may be is an open question. How many more hundreds or thousands of genes remain to be discovered for *Drosophila*? How can we best go about the business of finding these genes and deciphering their functions?

[1]Corresponding author.
E-MAIL wakimoto@u.washington.edu; FAX (206) 543-3041.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.169400.

Yeast researchers have set the gold standard for addressing such questions in functional genomics, because they can delete each of the predicted open reading frames in the yeast genome and examine consequences in vivo (Winzeler et al. 1999). Unfortunately, such approaches cannot be applied comprehensively to organisms that lack efficient methods for gene disruption or to those that have complex genomes and hundreds of cell types to assay for phenotypes. For *Drosophila*, the gold standard for evaluating gene number and function is provided by the 2.9-Mb Adh region, the most thoroughly understood region of the fly's genome (Ashburner et al. 1999). Drosophilists dream about having as comprehensive a knowledge of the remaining 98.5% of the genome as Ashburner and colleagues have provided for the Adh region. In reality, such in-depth understanding was a hard-won victory. Extensive genetic and molecular analyses carried out over a span of several decades and annotation efforts carried out over a span of two years account for the high confidence level in the gene estimates for the Adh region (Ashburner 2000).

In this issue, Andrews et al. (2000) describe a practical route to gene discovery, one they prove to be useful for *Drosophila* and one that can be applied to other multicellular organisms. The strategy is based on the analysis of expressed sequence tags (ESTs) from a defined tissue. The use of ESTs for gene discovery is not a novel idea (Adams et al. 1991; Rubin et al. 2000); however, the results of Andrews et al. (2000) are particularly timely and satisfying given the current status of the *Drosophila* Genome Project. The key to their success was the application of both computational and microarray approaches to characterize the properties of their new collection of ESTs. These approaches allowed them to assess the complexity of the EST collection, its relationship to in vivo expression profiles, and its redundancy with other available EST banks. Once the potential of this EST collection to provide new information was established, Andrews et al. (2000) demonstrated that the unique ESTs provided biological evidence for the existence of hundreds of predicted genes, newly discovered genes, or transcript forms. The success with this analysis led the authors to propose that the gene identification mission for multicellular organisms could advance considerably by taking advantage of tissue differences in gene expression profiles. Thus, a sampling of a relatively modest number of ESTs (approximately several thousand) from many different tissues could identify novel genes much faster than deeper probing of a few general libraries. In addition, as demonstrated here, an extra reward is gained from generating a collection of tissue ESTs, namely, that the ESTs can be used to learn something about the biology of the tissue of interest.

As the first step in this study, the authors asked if their tissue source, the adult testis, expressed a sufficiently complex RNA population to be useful for whole-scale EST analysis. Given that *Drosophila* males produce sperm with enormously long tails, the expression profile of the testis could have been dominated by a small number of transcript types, for example, those encoding structural components of the tail.

Fortunately, this is not the case. A collection of 3141 testis ESTs were sequenced with an average 5′ read of 449 bp; when compared to the large bank of ESTs from the Berkeley *Drosophila* Genome Project (BDGP), the testis ESTs showed a level of complexity comparable to the brain and ovarian EST collection. In spite of the fact that the testis and ovary share the responsibilities of maintaining a germ line and making a gamete, the proportion of ESTs that overlap in the testis and ovarian EST collections is no greater than the proportion that each shares with the brain EST collection.

With those characteristics of the testis EST collection established, the authors could then assess how useful the ESTs would be for discovering new *Drosophila* genes. The answer is that the testis ESTs proved to be surprisingly informative. Even with a relatively modest number of 1560 nonoverlapping ESTs, 47% failed to align with the ~80,000 ESTs sequenced by the BDGP. So far, the unique testis ESTs provide in vivo evidence for >500 predicted genes and an estimated 200 genes that were not identified by gene finder programs.

It is known that the EST approach can be misleading if, for example, some cDNA libraries contain a large proportion of chimeric molecules, genomic DNA contamination, or unspliced introns. Therefore, the use of any EST collection for gene discovery needs to be validated. A subset of the testis ESTs were mapped onto genomic sequences and examined for artifacts. The authors found that the 5′ EST sequences are consistent with typical gene structure and that at least two-thirds of the candidates subject to this close scrutiny defined new genes. It will take additional experimental studies to determine the precise number of entirely new genes that can be identified by this EST collection. However, it is clear from this analysis that the total gene number for *Drosophila* will exceed the estimate of 13,601 by a significant fraction.

The information gained from the testis EST analysis extends beyond the question of whether it can be useful for gene discovery. The second reward comes from having a collection of cDNAs to study testis biology. Andrews et al. (2000) provide the first microarray analysis of the testis as an isolated tissue. They have catalogued nearly 1700 testis ESTs using microarrays and assayed relative levels of expression in the testis, ovary, and soma. The microarray capabilities are particularly exciting because they can be combined with other large-scale approaches to study gametogenesis. For instance, a collection of ~2000 strains of *Drosophila* carrying recessive male sterile mutations all induced on the same genetic background is now available (B.T. Wakimoto, D. Lindsley, E. Koundakjian, C. Herrera, D. Cowan, R. Hardy, and C. Zuker, pers. comm.). The effects of a single point mutation on testis gene expression can be assayed using the testis EST microarrays. Some of these male sterile mutations arrest spermatogenesis at specific stages (e.g., gonia or primary spermatocyte arrest) or cause overproliferation of certain cell types and can be used with microarrays to characterize stage- or cell-specific profiles of gene expression. These data can be compared to those recently obtained by Reinke et al. (2000), who used microarray analysis of *C. elegans* genes to identify 1416 germ-line-enriched genes, of which 650 were classified as sperm enriched. The studies by Andrews et al. (2000) and Reinke et al. (2000) usher in molecular strategies to characterize gene expression during spermatogenesis on a global scale. In combination with genetic approaches and other types of analyses, such approaches should allow us to define the numbers and types of genes required for spermatogenesis and the extent to which these genes are conserved among organisms.

REFERENCES

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

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. 1991. *Science* **252:** 1651–1656.

Andrews, J., Bouffard, G., Cheadle, C., Lü, J., Becker, K.G, and Oliver, B. 2000. *Genome Res.* **10:** 2030–2043.

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

Ashburner, M. 2000. *Genome Res.* **10:** 391–393.

Gatti, M. and Pimpinelli, S. 1992. *Annu. Rev. Genet.* **26:** 239–275.

Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J.M., Davis, E.B., Scherer, S., Ward, S., et al. 2000. *Mol. Cell* **6:** 605–616.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. 2000. *Science* **287:** 2222–2224.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. *Science* **285:** 901–906.