Gene Discovery Using Computational and Microarray Analysis of Transcription in the *Drosophila melanogaster* Testis

Justen Andrews,1 Gerard G. Bouffard,2 Chris Cheadle3, Jining Lü,1 Kevin G. Becker,3 and Brian Oliver1,4

1Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA; 2Bioinformatics Group, National Institute of Health Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Gaithersburg, Maryland 20877, USA; 3DNA Array Unit, National Institute on Aging, National Institutes of Health, Baltimore, Maryland 21224, USA

Identification and annotation of all the genes in the sequenced *Drosophila* genome is a work in progress. Wild-type testis function requires many genes and is thus of potentially high value for the identification of transcription units. We therefore undertook a survey of the repertoire of genes expressed in the *Drosophila* testis by computational and microarray analysis. We generated 3141 high-quality testis expressed sequence tags (ESTs). Testis ESTs computationally collapsed into 1560 cDNA set used for further analysis. Of those, 11% correspond to named genes, and 33% provide biological evidence for a predicted gene. A surprising 47% fail to align with existing ESTs and 16% with predicted genes in the current genome release. EST frequency and microarray expression profiles indicate that the testis mRNA population is highly complex and shows an extended range of transcript abundance. Furthermore, >80% of the genes expressed in the tests showed onefold overexpression relative to ovaries, or gonadectomized flies. Additionally, >3% showed more than threefold overexpression at p <0.05. Surprisingly, 22% of the genes most highly overexpressed in testis match *Drosophila* genomic sequence, but not predicted genes. These data strongly support the idea that sequencing additional cDNA libraries from defined tissues, such as testis, will be important tools for refined annotation of the *Drosophila* genome. Additionally, these data suggest that the number of genes in *Drosophila* will significantly exceed the conservative estimate of 13,601.

[The sequence data described in this paper have been submitted to the dbEST data library under accession nos. AI944400–AI947263 and BE661985–BE662262.]

[The microarray data described in this paper have been submitted to the GEO data library under accession nos. GPLS, GSM3–GSM10.]

Genome projects are revolutionizing biology. However, extracting biologically relevant information from sequence data will be a protracted process. Clearly, a thorough analysis of even a single gene by traditional molecular genetic techniques is wrought with complexities that are only resolved by multifaceted analysis. Similarly, only by compiling evidence from computational, functional, and evolutionary data will we ultimately arrive at a comprehensive understanding of how an organism is genetically encoded.

The immediate challenge in the post-genomic era is identifying transcribed regions of DNA. Genefinding programs are vitally important tools for identifying transcription units, but are subject to errors. The alcohol dehydrogenase region of *Drosophila* is a case in point. Gene-calling programs failed to identify some known genes in this region (Ashburner et al. 1999; Ashburner 2000; Birney and Durbin 2000; Gaasterland et al. 2000; Henikoff and Henikoff 2000; Krogh 2000; Parra et al. 2000; Reese et al. 2000a;b; Salamov and Solovyev 2000). Expressed sequence tag (EST) analysis is also an important tool for identifying transcription units but is also subject to errors (Adams et al. 1991; Okubo et al. 1992; Weinstock et al. 1994; Adams et al. 1995; Hillier et al. 1996; Audic and Claverie 1997; Wolfsberg and Landsman 1997; Rubin et al. 2000). For instance, it is difficult to know if novel or rare ESTs represent genuine rare transcripts or artifacts arising from splicing intermediates, genomic contamination, or illegitimate transcription from intergenic regions. Consequently, we need multiple sources of evidence for accurate gene discovery and genome annotation. Microarrays derived from EST collections provide an
Drosophila will be quite useful in the ongoing effort to identify genes and gene structure. Therefore, the transcripts in the testis comprise a relatively complete set of instructions for a complex developmental process (for review, see Fuller 1993). Stem cells and dividing cystocytes are located at the apical region. The next third of the testis contains the highly transcriptionally active spermatocytes. These cells derive from, and are approximately 40× the size of, the parent cystocytes. Thus, primary spermatocytes supply the bulk of the mRNA we have analyzed. Small clusters of meiotic cysts and round spermatids occupy the next zone. The remainder of the testis bears transcriptionally silent differentiating spermatids and mature sperm. Whereas the bulk of the testis is of germline origin, there are also small clusters of somatic cells that surround germline cysts, the somatic testis sheath, and the reproductive tract. We constructed a Drosophila testes cDNA library from dissected testis, and some fragments of the male reproductive tract (vas deferens, ejaculatory duct, and fragments of removed accessory glands).

The testis library is of high quality (Table 1). The unamplified library contained 8 × 10⁶ pfu indicating that the library samples a large pool of testis transcripts. Only 6% of inserts sequenced on both strands were chimeric (i.e., having inserts that map to two different regions of the genome that cannot be explained by typical intron/exon structure) and <1% of clones showed no inserts. We employed two measures to estimate the portion of full-length cDNAs in the library. Of the inserts derived from known genes, 67% showed the presence of the predicted AUG initiation codon, and 39% of the testis clones that match Berkeley Drosophila Genome Project (BDGP) ESTs (Rubin et al. 2000) showed more extended 5′ terminal sequences. Both of these indices suggest that high portions of the cDNAs are full length. For the bulk of the cDNAs, we sequenced random clones from the testis library by single pass 5′ sequencing (Table 2). The testis EST data

### RESULTS

The adult Drosophila testis bears all stages of spermatogenesis. Therefore, the transcripts in the testis comprise a relatively complete set of instructions for a complex developmental process (for review, see Fuller 1993). Stem cells and dividing cystocytes are located at the apical region. The next third of the testis contains the highly transcriptionally active spermatocytes. These cells derive from, and are approximately 40× the size of, the parent cystocytes. Thus, primary spermatocytes supply the bulk of the mRNA we have analyzed. Small clusters of meiotic cysts and round spermatids occupy the next zone. The remainder of the testis bears transcriptionally silent differentiating spermatids and mature sperm. Whereas the bulk of the testis is of germline origin, there are also small clusters of somatic cells that surround germline cysts, the somatic testis sheath, and the reproductive tract. We constructed a Drosophila testes cDNA library from dissected testis, and some fragments of the male reproductive tract (vas deferens, ejaculatory duct, and fragments of removed accessory glands).

The testis library is of high quality (Table 1). The unamplified library contained 8 × 10⁶ pfu indicating that the library samples a large pool of testis transcripts. Only 6% of inserts sequenced on both strands were chimeric (i.e., having inserts that map to two different regions of the genome that cannot be explained by typical intron/exon structure) and <1% of clones showed no inserts. We employed two measures to estimate the portion of full-length cDNAs in the library. Of the inserts derived from known genes, 67% showed the presence of the predicted AUG initiation codon, and 39% of the testis clones that match Berkeley Drosophila Genome Project (BDGP) ESTs (Rubin et al. 2000) showed more extended 5′ terminal sequences. Both of these indices suggest that high portions of the cDNAs are full length. For the bulk of the cDNAs, we sequenced random clones from the testis library by single pass 5′ sequencing (Table 2). The testis EST data

### Table 1. Testis cDNA Library Characterization

| Characterization | Value |
|------------------|-------|
| Unamplified      | 8 × 10⁶ pfu |
| Amplified        | 1.8 × 10¹² pfu |
| Mass Excised     | 4.7 × 10¹³ cfu |

#### Insert characterization

| Property | Value |
|----------|-------|
| Size Selection | 1-7 kb |
| Inserts >1 kb | 74% (n = 173) |
| No Insert | <1% (n = 3,408) |
| Chimeric | 5% (n = 86) |
| Initiating AUG Present | 67% (n = 139) |
| Longer (5′) than BDGP EST | 39% (n = 385) |

1 Includes testis, vas deferens, and portions of ejaculatory apparatus and accessory glands.
2 n = number tested for indicated property.
3 Clones where: BLAST E-values versus nr or BDGP EST databases are <1E-20; orientation was plus/plus, and identity = 100%.
consists of 3,141 high-quality sequence reads that passed base calling, length, repeat, vector, and mitochondrial filters. Testis EST sequences can be accessed in the NCBI EST database, dbEST (Boguski et al. 1993).

EST Complexity
If the testis expresses very high levels of only a few genes or was grossly deregulated, it would be unsuitable for an EST project. To determine if the overall form of the transcription profile was unusual, we compared testis EST profiles to those of other tissues. To accomplish this, we did BLASTN alignments of the complete testis collection against itself. Similarly, the ovary and head BDGP EST collections (Rubin et al. 2000) were self-aligned. Because our goal was to determine how many of the ESTs were novel and not to assign sequences to gene families, we set a BLAST cutoff \( E \) value of 1E-20. Empirical tests indicated that this cutoff is stringent enough to avoid calling a related gene a match, but not so stringent as to miss matches due to limited overlap (see Methods). The plot of frequency of sequence matches within all three libraries shows how frequently a given EST species occurs within each of the collections (Fig. 1). In the testis collection, we found that 42% of the EST reads aligned only with themselves, 30% overlapped with only 2–5 reads, and 7% of reads overlapped 20 or more times. These data suggest that most EST species in the library are represented at <0.03% and that only a few ESTs are represented at >0.6%. The maximum testis EST frequency was 1.2% (these ESTs correspond to a novel gene). The EST profiles were similar in all three tissues suggesting that all the mRNA populations from which these libraries were derived are complex with an extended abundance range. The absence of skewing to the right indicates that messages encoding a few terminal differentiation products do not compose the bulk of the mRNA population in any of these libraries. Because all of the libraries show a profile skewed to the left, the testis profile is representative and is unlikely to reflect grossly deregulated transcription in the testis. The testis and head libraries do show a larger proportion of singletons, which is consistent with anecdotal accounts of complex gene-expression patterns in brain and testis. Since the testis EST profile shows no unusual features, these data indicate that testis ESTs are useful tools for gene discovery and genome annotation.

To estimate the total number of different genes represented by the testis ESTs, we assembled the total set into a nonoverlapping set of 1,100 singletons and 460 contigs (Table 3). Singletons plus contigs compose the 1,560 computationally nonoverlapping testis EST collection. This is a computational analysis, so some of the nonoverlapping sequences will ultimately be shown to derive from the same gene. To estimate how many genes are represented by multiple nonoverlapping ESTs, we checked the ESTs that align with 17 known genes expressed in testis, and found two examples (12%) of genes (exuperantia and don juan) represented twice. Conversely, some contigs could be spurious due to overlapping genes on opposite strands or chimeric inserts, for example. Analysis of 89 contigs suggests that 8% of contigs derive from different genes. Because these two types of errors essentially cancel, the nonoverlapping set of 1,560 provides a good approximation of gene representation in the testis EST collection.

| Sequence reads | 3,408 |
| failed base calling QC | 180 |
| short insert sequence | 47 |
| no insert | 1 |
| E. coli | 1 |
| mitochondrial | 38 |
| total high quality | 3,141 |
| average high quality length | 499 nt |
| cumulative sequence | 1.41 million nt |
| NCBI accession numbers | AI944400–AI947263 BE661985–BE662262 |

Table 2. Testis EST Characterization
Alignments of Testis ESTs to Protein, EST, and Predicted Gene Databases

To determine how many previously known or predicted genes occurred in the testis ESTs and how many were novel, we did a series of BLAST alignments (Table 4). We aligned the total EST set and the nonoverlapping testis EST collection reads with the BDGP ESTs and the BDGP/Celera Genomics (CG) Gadfly predicted genes (Adams et al. 2000; Rubin et al. 2000). Additionally, we aligned conceptual translation products with the GenBank nonredundant protein database (Benson et al. 1999).

Identification of ESTs matching genes with known functions is especially important as a method of library validation (Table 4). Of the nonoverlapping ESTs, 11% aligned with known genes based on protein encoding regions (excluding genome project data) and 9% were previously represented by an BDGP EST. Reassuringly, genes expressed specifically in the regions (excluding genome project data) and 9% were aligned with known genes based on protein encoding validation (Table 4). Of the nonoverlapping ESTs, 11% functions is especially important as a method of library et al. 1999).

We performed BLAST alignments between three EST collections from testis, ovary, and head. We found that the high proportion of novel ESTs in the testis collection is not peculiar. In a three-way comparison of

| Sequence match category | nr protein1 | BDGP EST2 | Gadfly predicted3 | Total testis ESTs | Non-overlapping testis ESTs |
|-------------------------|------------|-----------|-------------------|-------------------|----------------------------|
| +                       | +          | +         | 7% (231)          | 9% (139)          |
| +                       | +          |          | 2% (53)           | 2% (30)           |
| +                       | +          |          | <<1% (1)          | <<1% (1)          |
| +                       |          |          | <<1% (1)          | <<1% (1)          |
| +                       | +          |          | 46% (1443)        | 41% (641)         |
|                         | +          |          | 29% (909)         | 33% (511)         |
|                         | +          |          | 4% (139)          | 3% (42)           |
|                         |            |          | 12% (364)         | 13% (195)         |

1BLASTX sequence match (E-value <1E-20 and >90% sequence identity) with GenBank non-redundant protein.  
2BLASTN sequence match (E-value <1E-20) with BDGP EST.  
3BLASTN sequence match (E-value <1E-20) with BDGP/CG GadFly predicted gene.
ESTs in the *Drosophila* testis, ovary, and head EST collections (Fig. 2), we found that a significant proportion of ESTs are represented in only one tissue regardless of source: 55% of testis EST are present only in the testis EST collection, 30% of ovary ESTs are present only in the ovary EST collection, and 60% of the head ESTs are present only in the head and ovary EST collections.

Interestingly, even those EST reads represented in the testis and in another EST collection tend to be highly overrepresented in the testis (Fig. 3). Indeed, only the ESTs represented in all three sets show a distribution pattern that is consistent with so-called "housekeeping" functions. Thus, it is quite likely that the analysis of more *Drosophila* tissues will substantially augment our picture of the transcribed regions of the genome.

**Testis Expression Profile Analyzed by Microarrays**

In terms of achieving full coverage of *Drosophila* transcription units, an important question is how many testis transcripts are so rare in most tissues that they could not be found easily by sequencing other cDNA libraries to a greater depth. The computational analysis of the testis ESTs, outlined above, suggested that many genes are preferentially expressed in testis, but the alarming lack of overlap between the various EST collections might also suggest that many sequences are being lost during library construction (c.f., Wang et al. 2000). Alternatively, the lack of overlap between testis and other EST collections might be due to underlying differences in mRNA profiles. We directly examined how testis gene expression differs from that in related tissues by microarray analysis. Because labeled testis cDNAs should preferentially hybridize to testis cDNAs on the array, we could not normalize to total counts. We therefore normalized against control spots on the filters (see Methods). To bias the results in favor of genes common to the two tissues, and thus against the hypothesis that lack of overlap between collections is biologically based, we selected ovary and gonadectomized flies as test samples. The testis is similar to the ovary because of the presence of a germline, a cell lineage distinguished from the soma by its potential to undergo meiosis and form gametes. The testis shares a common sexual identity with the male soma.

We first asked how many of the testis ESTs were derived from mRNAs enriched in the germline. To investigate the contribution of germline gene expression to the testis EST collection, we compared expression profiles in testes versus gonadectomized males, and ovaries versus gonadectomized females. When labeled cDNA from these tissues was hybridized to the arrays, a remarkable 89% of array elements showed greater than onefold stronger signal when probed with testis versus gonadectomized male cDNA. Fifty percent showed a signal that was greater than 1.6-fold stronger (Fig. 4). Whereas individual data points at greater than onefold overexpression could be due to chance (statistical data are presented in the next section), the population of data points clearly depart from a one-to-one relationship. Additionally, the distribution was not normal, but showed skewing toward high testis expression. These data suggest that most of the cDNAs from these tissues was hybridized to the arrays, a remarkable 89% of array elements showed greater than onefold stronger signal when probed with testis versus gonadectomized male cDNA. Fifty percent showed a signal that was greater than 1.6-fold stronger (Fig. 4). Whereas individual data points at greater than onefold overexpression could be due to chance (statistical data are presented in the next section), the population of data points clearly depart from a one-to-one relationship. Additionally, the distribution was not normal, but showed skewing toward high testis expression. These data suggest that most of the cDNAs in testis EST collection derive from genes overrepresented in the steady-state testis mRNA pool. If this pattern of gene expression revealed by the microarray experiments reflects germline-preferential ex-

---

**Figure 2** Three-way comparison of sequence matches between the testis, ovary, and head EST sets. All figure elements are color coded. 100% of testis (blue), ovary (red), and head (green) ESTs are within each color coded circle. The total number of ESTs in each collection is indicated. The color coded numbers show the percentage of ESTs from any of the three collections represented in the intersecting segments of the Venn diagram. For example, 60% of head ESTs are represented in only the head EST collection, 3% of head ESTs are represented in the head and testis EST collection, 8% of head ESTs are represented in all three EST collections, and 29% of head ESTs are represented in the head and ovary EST collections.

**Figure 3** Frequency scatter plots of testis ESTs also represented in other EST collections. Frequency that a given testis EST is represented in one library is plotted against the frequency that the same testis EST is represented in a second EST collection. (A) Testis frequency versus ovary frequency. (B) Testis frequency versus head frequency. (C) Ovary frequency versus head frequency.
pression, then a similar pattern should be observed when arrays were probed with labeled ovary versus gonadectomized female cDNA. However, comparing ovary to gonadectomized female array intensities revealed essentially equivalent gene expression in the two tissues (and also validates the normalization protocol). The median difference was only 1.1-fold overexpression in ovary (Fig. 4). Thus, although the testis gene expression profile clearly differs radically from gonadectomized male, there is little enrichment for genes expressed non-sex-specifically in germ cells in the testis EST collection.

We then asked how many of the testis ESTs were derived from mRNAs enriched in males. To investigate how the cDNAs species in the testis EST collection might be determined by sexual identity, we compared testes versus ovaries, and gonadectomized males versus females. The microarray experiments indicate that the steady-state testis mRNA pool is as different from the ovary as it is from the bulk of the male soma. When probed with labeled testis cDNA versus ovary cDNA, 85% of the array elements showed a stronger signal and 50% showed greater than 1.5-fold overexpression in testis (Fig. 4). As was the case with testis versus male, these data suggest that the bulk of the cDNAs in the testis EST collection are derived from mRNAs that are highly enriched in the steady-state testis pool. In contrast, comparing gonadectomized males to gonadectomized females revealed intensity differences clustered tightly around a one-to-one relationship, with a median intensity difference of 1.0 (Fig. 4). Briefly, the testis expression profile does not correlate with either ovary expression, as would be expected for germline-specific functions, or with gonadectomized male, as would be expected for simple sexual dimorphisms. When probed with testis, 82% of the array targets were preferentially illuminated relative to both gonadectomized males and ovaries. These data unambiguously indicate that the *Drosophila* testis expresses a complex and surprisingly restricted profile of genes that is much more than the simple intersection of male/female and germline/soma dichotomies. Given that over 80% of testis cDNAs are overexpressed in testis, it is not surprising that 55% of only 3141 testis ESTs are missing from 29,120 ovary and head ESTs. In terms of gene discovery, the microarray data supports the suggestion that sequencing additional tissue-specific libraries is more efficient than sequencing existing high-quality libraries to greater depth.

### Clustering and Computational Analysis of Statistically Significant Microarray Data Points

Microarray data also provides a good source of high-quality annotation information, as genomic regions that are transcribed in a developmentally regulated fashion are quite likely to be real genes. We were therefore interested in determining if genes that showed developmentally regulated expression in microarray experiments were more or less likely to be represented by novel ESTs or predicted genes. For example, if the majority of the 237 nonoverlapping testis EST that fail to align with predicted genes are derived from contaminating genomic DNA, or other artifacts, then none of these would be expected to hybridize well in DNA microarray experiments. We were therefore interested in parsing out only the most highly expressed genes.

To select individual genes for analysis, we performed a statistical test based on replicate microarray experiments to identify elements where \( p < 0.05 \). We also imposed an arbitrary ratio limit of threefold to restrict the analysis to those genes that are likely to be most highly overexpressed in the testis (or other samples). In comparisons of testis to gonadectomized males, 3% of the array elements showed greater than threefold stronger signal at \( p < 0.05 \) when probed with...
testis cDNA versus gonadectomized male cDNA (Fig. 5). Similarly, in comparisons of testis to ovary, 4% of array elements showed threefold greater signal, at p <0.05, when probed with testis versus ovary cDNA (Fig. 5). The reciprocal biases toward either male or ovarian expression were never over 1%. The array elements showing a threefold overexpression at p <0.05 in any experiment were selected for further study (Fig. 6).

If the individual microarray data points are accurate reflections of testis transcription, then it would be expected that clones showing the greatest microarray expression differences would include some genes with a previously known germline and/or testis function. Indeed, 24% of the selected ESTs represent genes identified in the pregenome literature (18% with sequence matches to the nr protein database [Fig. 6] and 7% with sequence matches to the nr nucleotide database [not shown]). Those in the testis-enriched group include: don juan, encoding a sperm tail specific protein of unknown function (Santel et al. 1997); male-specific transcript 87F (Kuhn et al. 1991); exuperantia, encoding a germline restricted RNA-binding protein required for fertility (Hazelrigg et al. 1990); janus-A and janus-B, which are coordinately regulated genes expressed specifically in the male germline (Yanicostas et al. 1989); prominin-like, encoding a protein found at the leading edges of cytoplasmic projections (Weigmann et al. 1997); and Dynein light chain 90F, encoding a molecular motor (Harrison et al. 1998). The don juan, male specific transcript 87F, and janus genes are expressed only in testis, clearly validating the microarray data. The known gene in the male-enriched group is male specific transcript 57D which encodes an accessory gland protein (Simmerl et al. 1995). This transcript is expected in the gonadectomized male samples and the testis samples, as fragments of the accessory glands were included in both. Known genes in the ovary-enriched group are also unsurprising. Two genes encoding ribosomal proteins were found (ribosomal proteins L32, and P1). Certainly, ribosomal protein gene expression is expected in all cells, but large numbers of ribosomal components are synthesized in the developing eggs, and down regulation of Ribosomal protein genes is known to cause defective egg formation (Qian et al. 1988). Briefly, the observed microarray expression profiles are consistent with the known functions of the above genes.

Microarray data is strong biological evidence that a particular cDNA is derived from a bona fide gene. Thus, genes showing clear patterns of differential expression in the microarray experiments are good tools for validating the rest of the testis EST collection. We therefore looked to see if microarray validated (Fig. 6) and total nonoverlapping ESTs (Table 4) matched similar frequencies of known, predicted, and novel genes. Of the set of elements chosen for analysis because of high hybridization differences, only 18% are from known genes (Fig. 6). This is essentially the same figure (11%) obtained when the entire set of nonoverlapping testis EST set was analyzed (Table 4). Similarly, of the cDNAs that show high and differential hybridization in the microarray experiments, 22% of the corresponding ESTs do not match a predicted gene, versus 16% for the entire nonoverlapping testis EST collection. The analysis of the ESTs derived from the most enriched testis mRNAs provides very strong, and thus far unique, evidence for 10 predicted genes. These data also provide strong evidence for 16 novel genes or novel transcription forms. Because of the similar frequencies of named, predicted, and novel genes seen when total nonoverlapping testis ESTs and microarray verified ESTs are analyzed computationally, confidence in the analysis of the entire set is raised. Thus, these computational and microarray data provide strong biological evidence for over 500 predicted, and over 200 new, genes or transcript forms.

Mapping Novel Genes
An abundantly expressed gene could fail to match a
Figure 6  Summary of statistically significant microarray expression profiles ($P < 0.05$, and threefold intensity difference). (A) Spotted cDNA clone name. (B) Microarray intensity differences from replicate comparisons (1, 2, 3, 4, and mean) of hybridizations with labeled cDNA from the indicated tissues. Tissues are color coded. For a given spotted cDNA, high relative hybridization with labeled “red” cDNA is indicated by red boxes, while high relative hybridization with labeled “green” cDNA is indicated by green boxes. Colorimetric scale is shown on the right. The $P$ values for each spotted cDNA in each experimental series are indicated by blue color coded bars, with the scale shown on the right. The microarray cDNA clones are clustered into those showing testis (62), male (3), ovary (8), or female (1) preferential microarray expression profiles as indicated on the right. (C) A summary of sequence matches between ESTs from the respective microarray cDNAs and the indicated sequence databases. The black and white key is shown on the right (cutoffs are as follows: GenBank nr protein: BLASTX $E$-value <1E-20, and >90% sequence identity, BDGP ESTs and BDGP/CG predicted genes from the GadFly database: BLASTN $E$-value <1E-20).
predicted gene for a number of reasons. Given the current state of gene prediction programs, trivial reasons include missed exons due to promoter, splice, and poly-A variants. A nontrivial reason is a completely missed transcription unit. A definitive resolution of why a testis EST fails to align with a predicted gene is a difficult task. We begin this process by examining ESTs identified in the microarray experiments for indication of exon/intron structure, and for arrangement relative to known or predicted genes. To do this, we examined the sequence alignment of testis ESTs, GenBank nonredundant, BDGP ESTs, and BDGP/CG GadFly predicted genes, to the genomic DNA sequence surrounding novel testis EST reads (Fig. 7; Benson et al. 1999; Adams et al. 2000; Rubin et al. 2000). The 15 genes with the strongest biological support were examined (Fig. 7): 67% show no obvious connection to named or predicted genes, 7% are clearly due to additional exons of predicted genes, and 27% are ambiguous.

The organization of 5’ testis EST reads are consistent with typical gene structures. In seven cases where multiple overlapping EST reads are mapped, the sequences show good 5’ alignment (Fig. 7A,C,D,E,F,J,L). This suggests the presence of a nearby promoter in the genomic sequence, and that the EST reads are from full-length, or nearly full-length, cDNAs. Similarly, it does not appear that isolation of incompletely processed pre-mRNAs during library construction can account for the common failure to align testis ESTs with known or predicted genes. Of the mapped testis ESTs, 73% show interrupted alignment with genomic sequence, suggesting that they were derived from spliced mRNAs, not pre-mRNAs (Fig. 7A,D,F,J,L,N,O). Indeed, the gene represented by bs13e06 has 14 matching testis ESTs with multiple aligned exons,
suggesting at least 3 mRNA splice variants (Fig. 7L). We found only one case of a failed alignment that is definitively due to exon structure differences between ESTs and predicted genes. The gene defined by bs13h06 was represented by 20 testis ESTs, which match a predicted gene over a single small region (Fig. 7D). This EST contig failed to match this short predicted gene in the BLAST search since the small length of overlap did not meet the E-value cutoff. In this case, the gene prediction program identified one of at least three exons.

Four transcription units deduced from testis EST sequence are in the same orientation and in the vicinity of either predicted genes or BDGP ESTs. It is difficult to determine if these represent extensions of predicted genes (especially in the absence of full-length testis or BDGP EST sequence), or if the region is particularly gene dense. The transcription units deduced from bs17f06 and bs09e08 are on the same strand and within approximately 150 bp of a predicted gene (Fig. 7E,G). The bs35d06 EST suggests a gene transcribed in the same orientation as, and within an intron of, CG12163, and thus may define a new exon (Fig. 7J). The transcription unit represented by bs35d06 is also in proximity to head and ovary ESTs; however, in the microarray experiments, bs35d06 hybridized two- to threefold better to labeled testis cDNA than to labeled ovary, male, or female cDNAs. This is an argument against joining bs35d06 with the head and ovary ESTs in the region.

Three transcription units within introns are likely. It is known that Drosophila has a surprisingly large number of genes within genes, usually on the opposite strand (Ashburner et al. 1999), and that these are problematic for gene predicting algorithms (Reese et al. 2000a,b). The transcription units defined by bs33g12, and bs03f11 EST reads are within introns of known genes, but in opposite orientation, suggesting the presence of genes within genes (Fig. 7I,O). The transcription unit defined by bs35d06 is also within a predicted intron, but in the same orientation (Fig. 7J). Two more transcription units, represented by bs13g05 and bs13e06, are within approximately 400 bp of predicted genes (with confirming ESTs) but in the opposite orientation, and hence are likely to define different genes (Fig. 7K,L). Lastly, five transcription units, represented by bs28a02, bs03a05, bs29d12, bs29a11, and bs35a07, do not map in close proximity to known genes (Fig. 7A–C,F,M). Although we cannot eliminate the possibility that these particular ESTs are derived from overlapping exons, it seems unlikely that this will be the general case. In summary, of the novel testis cDNAs showing the greatest microarray expression differences, all define previously known exons, one matches and extends a predicted gene, four may possibly extend predicted genes, and ten are probably new genes.

**DISCUSSION**

**Testis Gene Expression**

There are many anecdotal reports of genes expressed in a specific and restricted pattern in a tissue of interest and in the testis. The biological relevance of the testis expression is often questioned. The computational and microarray data presented in this study support the idea that the testis expresses a complex set of transcripts. However, the profiles are not consistent with deregulated or random transcription, as EST frequencies and microarray hybridization intensities showed an extended dynamic range. The opposite concern, that the testis expresses abundant terminal differentiation products at levels high enough to swamp out more moderately expressed genes, was not borne out. There are many more rare EST species than common ones. Importantly, computational comparisons of testis, head, and ovary EST collections suggest that the dynamic range of testis expression is similar to other complex body parts.

Whereas overall profiles of gene expression were similar in testis and other tissues (many rare messages and a few common ones), individual genes showed tissue-specific differences. Most strikingly, the gonads of the sexes are analogous organs, but surprisingly the microarray data indicate that the testis expression profile is nearly as different from ovary as it is from the bulk of the soma. Furthermore, microarray experiments show that the majority of the nonoverlapping testis ESTs represent mRNAs with significantly higher abundance in testis compared to other tissues. Perhaps, the testis requires the expression of this impressive battery of genes for the highly complex process required to generate sperm, and not simply because they do no harm (reviewed by Fuller 1993).

**Sequencing More cDNA Libraries**

One of the clear conclusions from nearly 20 years of studying development at the molecular level is that most genes are utilized in multiple developmental events or pathways. Consequently, one might expect that many or most of the genes expressed in an organism could be sampled by screening only a few complex tissues such as brain or testis. However, it is also the case that relatively few genes are expressed at similar levels in all tissues. This suggests that temporally or spatially restricted samples provide a significant enrichment for genes that are expressed at high level in that tissue at that time, while being rare at most points during the life of the organism.

In this study, we found that only 53% of testis ESTs overlap with the much larger set of 80,000 ESTs derived from other Drosophila sources (Rubin et al. 2000). A priori this could indicate that testis gene expression is particularly divergent from other probed...
tissues (ovary, embryo, head, larval, pupal, tissue culture cells, and head). Certainly, the computational and microarray data argues strongly for a highly specialized gene expression profile in the testis. However, this may, or may not be a unique feature of the testis. The computational analysis of EST frequencies within and between testis, ovary, and head ESTs suggests that each library brings a surprisingly restricted set of cDNAs into the EST stable. Whereas the ovary and head enrichments have not been verified by microarray experiments, the testis EST microarray experiments clearly indicate that many genes are highly expressed in testis and lowly expressed elsewhere. Thus, even the modest sequencing depth we employed in this study was sufficient to isolate multiple copies of many previously unknown cDNAs. If this holds for other Drosophila tissues, then shallow sequencing depth in a large number of tissue-specific libraries might maximize gene discovery in Drosophila. Indeed, low overlap between cDNA libraries may be general. In a 292,878 human EST data set, 42% of the nonoverlapping ESTs derived from a single organ or tissue, and only 3% were ubiquitous (Adams et al. 1995).

How Many Genes?
The total gene number in Drosophila has long been a matter of intense speculation (e.g., Muller 1928; Lefèvre and Watkins 1986), and is still far from settled (Ashburner et al. 1999; Ashburner 2000; Adams et al. 2000). Two approaches to gene-finding using the current Drosophila genome sequence data predicted between 17,464 and 13,189 genes (Adams et al. 2000). Guided by gene density estimates from the comprehensive analysis of the AdhI region (Ashburner et al. 1999), Adams et al. made a conservative estimate of 13,601 genes. However, in light of the gene density in the annotated genome sequence, Ashburner conceded that their analysis of the AdhI region may have been too conservative, which in turn affects the estimate by Adams et al. (Adams et al. 2000; Ashburner et al. 2000). Our study of transcription in the testis clearly indicates the existence of a significant class of undetected genes in the current genome release. If cDNAs from other Drosophila tissues show similar tendencies, then sequencing additional cDNA libraries will significantly augment our understanding of the Drosophila transcriptome and boost the Drosophila gene number well beyond 13,601.

METHODS

Flies and Dissections
All flies were from the y w^{a+/a} strain and were grown at 25°C. Flies used for the library construction were 1–5 d posteclosion. Flies used for microarray experiments were 12–24 h posteclosion. All flies were anesthetized on CO2 and dissected in 1 x PBS. Following each dissection, the tissues were transferred to a microcentrifuge tube on dry ice for snap freezing. Testis samples included testis, vas deferens, about half of the anterior ejaculatory duct, and fragments of the removed male accessory glands. Ovary samples include ovary, lateral oviduct, and most of the common oviduct. Gonadectomized samples include all remaining adult tissues.

Testis Library
Total RNA was extracted from flash frozen tissues (Trizol; Life Technologies) and Poly(A)+ RNA (2 µg) was selected using Oligotex (QIagen). cDNA synthesis was according to the manufacturer’s instructions (Stratagene). Briefly, RNA was poly(dT) primed and cDNA fractions between 1 and 6 kb were pooled for cloning. cDNA was directionally cloned with S’ EcoRI and S’ XhoI linkers in the Uni-Zap XR vector (Stratagene). The unamplified library contained 8 x 10^6 pfu. The library was amplified once (1 x 10^9 pfu yielded 1.75 x 10^{12} pfu) and an aliquot was mass excised (1 x 10^9 pfu yielded 4.7 x 10^9 cfu) to give clones in the pBluescript SK- plasmid vector.

Sequencing and Sequence Analysis
Random clones (3408) were 5’-end sequenced (PE/Applied Biosystems M13RP1 reverse primer and dye terminator). Chromatogram traces were evaluated using PHRED (Ewing and Green 1998; Ewing et al. 1998) yielding 3228 sequences, trimmed of vector using CrossMatch (P. Green, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html), yielding 3180 sequences. Reads were also screened for contaminating Escherichia coli and mitochondrial sequences using BLASTN (Altschul et al. 1997), yielding 3179 and 3141 sequences, respectively. All sequences were checked for repetitive elements using RepeatMasker (A.F.A. Smith and P. Green, unpubl.; c.f. RepeatMasker http://ftp.genome.washington.edu/RM/RepeatMasker.html), but were found to be essentially free of known Drosophila repeats. The 3141 high-quality sequences had an average read of 449 bp and a cumulative length of 1.41 Mbp. Sequences were submitted to NCBI and have accession nos. AI944400–AI947263. Clones will soon be available from the UK Human Genome Mapping Project Resource Centre (http://www.hgmp.mrc.ac.uk/).

Computational Analysis
Batch sequence alignments were performed on a computer workstation (Sun Microsystems) using BLAST (Altschul et al. 1997). For all BLASTN work the parameters were: E-value <1E-20, λ = 1.37, K = 0.711, H = 1.31, 1–3 matrix, gap penalty = 5, gap extension = 2. For all BLASTX work, the parameters were E-value <1E-6, λ = 0.27, K = 0.047, H = 0.23, BLOSUM62 matrix, gap penalty = 11, gap extension = 1. Drosophila sequence matches where the E-value was <1E-20 were considered to be derived from the same gene. The 1E-20 cutoff was chosen to approximate perfect matches between sequences from the same gene for empirical reasons. Ideally, a BLAST output where a query sequence is known to be present in a large database only once should give a single match. In tests where we aligned ESTs to the finished Drosophila genomic scaffolds, there were relatively few ESTs that matched multiple segments. For example, 6990 ESTs were aligned to the genomic scaffold at 1E-20 and returned 1.04 matches per EST. The few false-positives are likely due to gene family members. Similarly, a BLAST alignment of a sequence against itself should always result in a match. For example, in a test where
1526 testis ESTs were aligned against themselves, there were only 10, or 0.6%, that failed to self-align at 1E-20 (failed self-alignment is due to the filtering of low-complexity sequences that is required to generate valid output), suggesting that this cutoff also minimizes false negatives. Nonredundant database sequence matches of <1E-6 for the nonoverlapping testis ESTs are reported in the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/; see the section on Microarray Data Handling, below).

The GenBank nonredundant protein database (Benson et al. 1999) was downloaded from the NCBI web site (http://www.ncbi.nlm.nih.gov/Ftp/index.html). Sequence databases of nontestis Drosophila ESTs (Rubin et al. 2000), and BDGP/CG predicted genes (Adams et al. 2000), were downloaded from the BDGP web site (http://www.fruitfly.org/sequence/download.html). The GM, GH, and HL subsets of the BDGP ESTs were parsed out from the BDGP EST data set. To identify nonoverlapping clones for the microarray, the 3141 high-quality sequences were assembled into contigs using PHRAP and Consed (Gordon et al. 1998; P. Green, http://bozeman.mbt.washington.edu/phrap/docs/phrap.html). This yielded 460 contigs. Average contig length was 735 bp and was composed of 4.4 EST reads.

Array Printing

The array was assembled using 1681 testes ESTs, representing 1527 nonoverlapping EST sequences and 144 anonymous clones. An aliquot of the same DNA preparation was used in the sequencing reactions, diluted 2/50 in TE, provided material for the PCRs. Inserts were PCR amplified (1 × PCR buffer [Life Technologies], 1.5 mM MgCl₂, 0.2 mM dNTPs, 1 µM T7 22-mer primer, 1 µM T3 20-mer primer, 2 µL of Plasmid DNA polymerase [Life Technologies], 1.5 mM MgCl₂, 0.2 mM dNTPs, 1 µM T7 22-mer primer, 1 µM T3 20-mer primer, 2 µL of Plasmid DNApolymerase [Life Technologies], 10% dextran sulphate, 0.2 mg/mL polyA RNA at 50 mL centrifuge tubes (Falcon) in roller bottles (Hybaid). Filters were washed twice for 10 min at 25°C, for 15 min at 65°C (2 x SSC, 0.1% SDS), for 20 min at 65°C (2 x SSC, 0.1% SDS). Hybridized microarrays were exposed and quantitated as previously described (Whitney et al. 1999).

Microarray Data Accession

Microarray data can be found in the NCBI Gene Expression Omnibus, GEO under accession nos. GPL5 and GSM3–GSM10 (A. Lash, R. Edgar, M. Domrachev, V. Soussov, and J. Ostell, h t t p : / / w w w . n c b i . n l m . n i h . g o v / g e o /). Briefly, version 1.0 of the testis platform is organized into 12 subarrays printed in duplicate. Each array element is linked to a clone list, sequence accession numbers, and BLAST output definition lines. Two quality control measures are also given. The first is a binary score for successful PCR. The second, is raw intensity data from a hybridization to the primers used to amplify the inserts. Assuming that all array element hybridize similarly, this is a relative measurement of how many molecules were spotted at each array element. Data from individual experiments is linked to the platform. Raw intensity values and background subtracted and normalized data are presented.

Labeling and Hybridization

Total RNA was extracted from flash frozen tissues (Trizol; Life Technologies) and was reverse transcribed [20 µg total RNA, 0.025 µg/µl oligo d(T)₁₂–₁₄, 1 x 1st strand buffer (Life Technologies), 10 mM DTT, 500 µM dATP, dGTP, dTTP, 0.5 µM [3²P]dCTP [2.5 Ci/µmol; NEN Life Sciences Products], 1 U/µL RNaseOUT (Life Technologies), 10 U/µl SuperScript II reverse transcriptase (Life Technologies), in a total volume of 40 µL for 35 min 42°C), followed by an additional 35 min incubation with the addition of 400 U SuperScript II reverse transcriptase. RNA was hydrolyzed (45 mM EDTA, 18 mM NaOH, 1 h at 65°C), and unincorporated nucleotides removed on Biospin 30 columns (BioRad). Prehybridization was in 1 x Microhyb (Research Genetics), 10% dextran sulphate, 0.25 mg/mL sonicated salmon sperm DNA, 0.2 mg/mL polyA RNA at 65°C for 4 h. Heat denatured probe was added to the prehybridization solution and hybridized for 16–18 h at 65°C. All prehybridization and hybridization steps were performed in 50 mL centrifuge tubes (Falcon) in roller bottles (Hybaid). Filters were washed twice for 10 min at 25°C, for 15 min at 65°C (2 x SSC, 0.1% SDS), for 20 min at 65°C (2 x SSC, 0.1% SDS). Hybridized microarrays were exposed and quantitated as previously described (Whitney et al. 1999).

Microarray Data Correction

Background and normalization correction places the lowest intensity elements at the origin and the highest intensity elements, showing low experiment-to-experiment variance, at one. Background correction was accomplished by taking the average intensity of two blank positions in each of the 24 12 × 12 element subarrays. Normalization was to control spots modeled from the raw intensity values. This normalization protocol is similar to “spiking” type controls in that there are control spots, but uses an empirically defined set of spots. Because spiking RNA is exogenous, one cannot be certain that the control and experimental RNA are of identical quality. Twenty-five elements from each quadrant of the array were selected based on low variance over all the experiments. The average intensity value for these control elements in each quadrant was calculated. The first hybridization with labeled testis cDNA to the first replicate set of elements in the first quadrant of the array was set as the standard against which the remaining data sets were normalized. Thus, where Vx is the corrected value for any given point x, S is the raw hybridization intensity, B is the average local background, Nₓ is the average normalization element value in quadrant one of testis experiment one, Bₓ is the average local background in quadrant one of testis experiment one, Nₓ is the average normalization value for quadrant bearing point x, and Bₓ is the average local background for subarray bearing point x, the corrected value was determined according to the following formula:

\[ V_x = (S - B_x) \times \left( \frac{(N_\Gamma - B_\Gamma)}{(N_x - B_x)} \right) + 39,542 \]

Thus for each data set the average background and the average control element intensities were set equal to those in the testis. The arbitrary value 39,542 was added to minimize the occurrence of negative values.

Statistical Treatment of Microarray Data

All experiments were done in quadruplicate. Sample duplicate subarrays are shown in Figure 8. The variability in the system occurred mostly between replicate experiments (fluctuations

Genome Research 2041

www.genome.org
in dissection, RNA isolation, radiolabeling efficiency, \( r^2 = 0.68 \) in the typical example shown. There was very little variability between replicate spots hybridized with the same labeled cDNA (\( r^2 = 0.97 \) in the typical example shown). We therefore averaged duplicate spot intensity values after normalization, then used a statistical test to distinguish significant array intensity differences between tissues. For each pair-wise tissue comparison, \( P \) values were calculated for each clone using the two-sided \( t \) test for differences in means. To satisfy a 0.05 significance cutoff, the difference in mean intensity values must equal or exceed \( 4.3 \times \) the estimated standard error (Claverie 1999).

ACKNOWLEDGMENTS

We thank members of NISC staff and Jeff Touchman for sequencing; William Wood for help with microarray production; Virginia Boulais for growing flies; and Jay Miller and Carl Baker for Phosphoimager use.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celinski, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of Drosophila melanogaster. 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. Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252: 1651–1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377: 556–561.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Ashburner, M. 2000. A biologist’s view of the Drosophila genome annotation assessment project. Genome Res. 10: 391–393.

Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., et al. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster: The Adh region. Genetics 153: 179–219.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. Genome Res. 7: 986–995.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L. 1999. GenBank. Nucleic Acids Res. 27: 12–17.

Birney, E. and Durbin, R. 2000. Using GeneWise in the Drosophila genome annotation experiment. Genome Res. 10: 547–548.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. 1993. dbEST—Database for “expressed sequence tags”. Nat. Genet. 4: 332–333.

Castrillon, D.H., Gonczy, P., Alexander, S., Rawson, R., Eberhart, C.G., Viswanathan, S., D’Nardo, S., and Wasserman, S.A. 1993. Toward a molecular genetic analysis of spermatogenesis in Drosophila melanogaster: Characterization of male-sterile mutants generated by single P element mutagenesis. Genetics 135: 489–505.

Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. Hum. Mol. Genet. 8: 1821–1832.

DiBenedetto, A.J., Lakich, D.M., Kruger, W.D., Belote, J.M., Baker, B.S., and Wolfnier, M.F. 1987. Sequences expressed sex-specifically in Drosophila melanogaster adults. Dev. Biol. 119: 242–251.

Eddy, E.M. 1998. Regulation of gene expression during spermatogenesis. Semin. Cell Dev. Biol. 9: 451–457.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using PHRED. I. Error probabilities. Genome Res. 8: 186–194.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. 1998. Base-calling of automated sequencer traces using PHRED. II. Accuracy assessment. Genome Res. 8: 175–185.

FlyBase. 1999. The FlyBase database of the Drosophila Genome Projects and community literature. Nucleic Acids Res. 27: 85–88.

Fuller, M.T. 1993. Spermatogenesis. In The development of Drosophila melanogaster (ed. M. Bate and A. Martinez-Arias), pp. 1–70. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Gaasterland, T., Szyrba, A., Thomas, E., Aytekin-Kurban, G., Andrews et al.

Figure 8 Reproducibility of microarray data. Data from duplicate experiments with radiolabeled cDNA from testis and gonadectomized male. (A) Merged, pseudocolored images of the same subarray from microarrays hybridized with radiolabeled cDNA from testis (red channel) and gonadectomized male (green channel). Duplicate spots are shown in the columns headed spot A and spot B. Results from duplicate experiments (independent tissue, RNA, reverse transcription, and hybridization) are shown in the rows labeled RNA sample 1 and RNA sample 2. (B, C) Scatter plots of hybridization intensity ratios. (B) Scatter plot of hybridization intensity ratios for duplicate spots (\( r^2 = 0.97 \)). (C) Scatter plot of hybridization intensity for duplicate RNA samples (values from duplicate spots were averaged, \( r^2 = 0.68 \)).
Gordon, P., and Sensen, C.W. 2000. MAGPIE/EGRET annotation of the 2.9-Mb Drosophila melanogaster Adh region. Genome Res. 10: 502–510.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. Genome Res. 8: 195–202.

Harrison, A., Olds-Clarke, P., and King, S.M. 1998. Identification of the t complex-encoded cytoplasmic dynein light chain tctex1 in inner arm I1 supports the involvement of flagellar dyneins in metiotic drive. J. Cell Biol. 140: 1137–1147.

Hazelrigg, T., Watkins, W.S., Marcey, D., Tu, C., Karow, M., and Lin, X.R. 1990. The exuperantia gene is required for Drosophila meiotic drive.

Henikoff, J.G. and Henikoff, S. 2000. Drosophila genomic sequence annotation using the BLOCKS+ database. Genome Res.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiarello, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 6: 807–828.

Krogh, A. 2000. Using database matches with HMMGene for automated gene detection in Drosophila. Genome Res.

Kuhn, R., Kuhn, C., Borsch, D., Glazer, K.H., Schafer, U., and Schaefer, M. 1991. A cluster of four genes selectively expressed in the male germ line of Drosophila melanogaster. Mech. Dev. 35: 143–151.

Lanfranchi, G., Muraro, T., Caldara, F., Pacchioni, B., Pallavicini, A., Pandolfo, D., Toppo, S., Trevisan, S., Scarso, S., and Valle, G. 1996. Identification of 4370 expressed sequence tags from a Drosophila melanogaster genomic DNA. Nat. Genet. 161: 284.

Lefevre, G., and Watkins, W. 1986. The question of the total gene number in Drosophila melanogaster. Genetics 113: 869–895.

Lifschytz, E. and Yakobovitz, N. 1978. The role of X-linked lethal and viable male-sterile mutations in male gametogenesis of Drosophila melanogaster: genetic analysis. Mol. Genet. Genom. 161: 275–284.

Muller, H.J. 1928. The gene as the basis of life. Proc. 4th Intl. Congr. Plant Sci. [Ithaca, NY, 1926] 1: 897–921.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Koijima, Y., and Matsubara, K. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet. 2: 173–179.

Parra, G., Blanco, E., and Guigo R. 2000. GeneID in Drosophila. Genome Res. 10: 511–515.

Persengiev, S.P., Robert, S., and Kilpatrick, D.L. 1996. Transcription of the TATA binding protein gene is highly up-regulated during spermatogenesis. Mol. Endocrinol. 10: 742–747.

Pocca, D. 1994. Molecular aspects of spermatogenesis. R.G. Landes Compay, Austin, TX.

Qian, S., Hong, S., and Jacobs-Lorena, M. 1988. Antisense ribosomal protein gene expression specifically disrupts oogenesis in Drosophila melanogaster. Proc. Natl. Acad. Sci. 85: 9601–9605.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Ahrli, J.F., and Lewis, S.E. 2000a. Genome annotation assessment in Drosophila melanogaster. Genome Res. 10: 483–501.

Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. 2000b. Gene—Gene finding in Drosophila melanogaster. Genome Res. 10: 529–538.

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

Salamon, A.A. and Solovey, V.V. 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10: 516–522.

Santel, A., Winhauer, T., Blumer, N., and Renkawitz-Pohl, R. 1997. The Drosophila don juan (dj) gene encodes a novel sperm specific protein component characterized by an unusual domain of a repetitive amino acid motif. Mech. Dev. 64: 19–30.

Schmidt, E.E. and Schibler, U. 1995. High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. Development 121: 2373–2383.

Simmerl, E., Schafer, M., and Schafer, U. 1995. Structure and regulation of a gene cluster for male accessory gland transcripts in Drosophila melanogaster. Insect Biochem. Mol. Biol. 25: 127–137.

Wang, S. M., Fears, S. C., Zhang, L., Chen, J. J., and Rowley, J. D. 2000. Screening poly(dA/dT)- cDNAs for gene identification. Proc. Natl. Acad. Sci. 97: 4162–4167.

Wei, C., Corbeil, D., Hellwig, A., and Huttner W.B. 1997. Prominin, a novel microvilli-specific polytopic membrane protein of the apical surface of epithelial cells, is targeted to plasmalemmal protrusions of non-epithelial cells. Proc. Natl. Acad. Sci. 94: 12425–12430.

Weinstock, K.G., Kirkness, E.F., Lee, N.H., Earle-Hughes, J.A., and Venter, J.C. 1994. cDNA sequencing: A means of understanding cellular physiology. Curr. Opin. Biotechnol. 5: 599–603.

Whitney, L.W., Becker, K.G., Tresser, N.J., Caballero-Ramos, C.I., Munson, P.J., Prabhu, V.V., Trent, J.M., McFarland, H.F., and Biddison, W.E. 1999. Analysis of gene expression in multiple sclerosis lesions using cDNA microarrays. Ann. Neurol. 46: 425–428.

Wolfsberg, T.G. and Landsman, D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. 25: 1626–1632.

Yaniceostas, C., Vincent, A., and Lepesant, J.A. 1989. Transcriptional and posttranscriptional regulation contributes to the sex-regulated expression of two sequence-related genes at the janus locus of Drosophila melanogaster. Mol. Cell Biol. 9: 2526–2535.

Received August 10, 2000; accepted in revised form October 12, 2000.