Published in final edited form as:

Methods. 2014 June 15; 68(1): 38–47. doi:10.1016/j.ymeth.2014.03.007.

Navigating and Mining modENCODE Data

Nathan Boley^(a), Kenneth H. Wan^(b), Peter J. Bickel^(c), and Susan E. Celniker^(b)

^(a)Department of Biostatistics, University of California Berkeley, Berkeley, CA

^(b)Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, CA

^(c)Department of Statistics, University of California Berkeley, Berkeley, CA

Abstract

modENCODE was a 5 yr NHGRI funded project (2007–2012) to map the function of every base in the genomes of worms and flies characterizing positions of modified histones and other chromatin marks, origins of DNA replication, RNA transcripts and the transcription factor binding sites that control gene expression. Here we describe the *Drosophila* modENCODE datasets and how best to access and use them for genome wide and individual gene studies.

Introduction

In 1998 and 2000, the complete genome sequences for *C. elegans*¹ and *D. melanogaster*^{2,3,4}, respectively, were published. Thus emerged the modENCODE project⁵, whose goal was to map functional elements to the genomes of these model organisms. The modENCODE consortium was formed as a network of research laboratories, comprised of 11 research projects: four projects for worm, six for fly and one contributing to both organisms. The six fly projects are: domains of gene structure; transcription factor binding sites; histone modifications; chromatin structure; origins of DNA replication and timing; and RNA expression profiling. Here we describe the *Drosophila* modENCODE datasets⁶, and how best to access and use them for genome wide and individual gene studies. The data can be accessed through the modENCODE website (<http://www.modencode.org/>). All data from this project is publicly available and was vetted by a data coordinating center (DCC) to ensure consistency and completeness⁷. All the software of the project is publicly available.

DNA Sequencing and Replication

Sequencing Technologies

Four sequencing technologies - Sanger, 454, Illumina, and SOLiD - were used in modENCODE. We describe these briefly here; for comprehensive reviews see^{8,9,10}.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Standard Sanger sequencing¹¹ produces long reads (up to about 900 bp), has easy reaction setup, and can assign sequence to individual clones. The major disadvantage is the low throughput compared to newer technologies. The *D. melanogaster* genome was primarily sequenced using the Sanger Sequencing technology.

Roche's 454 instruments use pyrosequencing¹², which can produce read lengths up to 1kb, but has difficulties sequencing long homopolymer stretches and requires a time consuming sample preparation step. ModENCODE used Roche's 454 instruments to produce the bulk of the 5' RACE data (Transcriptomics - Transcription Start Sites).

The Illumina sequencing platform, the most common "second generation" sequencing technology, was used to sequence most of the modENCODE DNA and RNA data. It is high throughput, costs relatively little per base, and has a relatively easy sample preparation step. However, it can only produce short reads lengths (currently up to 300 nt, but only 100 nt during the majority of modENCODE).

Applied Biosystem's (Life Technologies) SOLiD platform¹³ is also high throughput and has better accuracy than Illumina⁸, but requires a relatively difficult sample preparation step, under-represents AT- and GC- rich regions with substitutions as the most common error type⁸. However, the SOLiD platform was the first to produce stranded data and was used to sequence the total RNA-seq time course through embryogenesis.

Whole Genome Sequencing

For comparative studies, eight additional *Drosophila* species were sequenced. The read data are available in the Sequence Read Archive (SRA), and the assembled genomes are available from Genbank (Table 1).

Copy Number Variation

Immortalized cell line genomes typically differ substantially from the reference genome. To investigate these differences, whole genome sequencing experiments were conducted in 21 different *Drosophila* cell lines. For the cell lines with two independent cultures, S2-DRSC and Cl.8, expression, gene ontology, and chromatin analyses were also conducted¹⁴. The data is available at data.modencode.org under the "Copy Number Variation" project category.

Epigenetics and Transcription Regulation

Chromatin is subject to a diverse array of post-translational modifications, which are known to play important roles in gene regulation. To study the set of histone modifications and elucidate their function, the modENCODE consortium conducted 535 ChIP-chip and 322 ChIP-Seq experiments which localized 42 distinct histone modifications, 64 chromatin-binding factors and 61 transcription factors. The major results of this study can be found in¹⁵; an online browsing tool for the chromatin marks can be found at <http://compbio.med.harvard.edu/flychromatin/>. The data can be downloaded from <http://data.modencode.org/>.

Assays to Identify Regions Bound by DNA Binding Proteins

Briefly, ChIP-chip and ChIP-seq assays work by first cross-linking the protein to DNA (e.g. by formaldehyde fixation). The cells are lysed, and the DNA is sheared (e.g. by sonication) resulting in DNA fragments, where some are bound to the protein of interest. An anti-body is then introduced that binds to the protein of interest, the antibody bound fragments are separated, the cross-linking is reversed, and the DNA fragments are purified. This process results in DNA fragments enriched for regions that were bound to the protein of interest, in this case DNA bound to a specifically modified histone protein. ChIP-chip and ChIP-Seq assays vary from this stage forward.

In ChIP-chip assays, the DNA fragments are amplified, denatured, labeled with a fluorescent tag, and poured over the surface of a DNA microarray. The DNA microarray is spotted with short single stranded DNA sequences tiling the genome, so the poured fragments can hybridize with complementary sequences on the array. The array is illuminated with a fluorescent light, and the fluorescence signals are captured and associated with their underlying sequence. Identifying bound genomic regions from the fluorescence requires substantial post-experimental analysis, although software packages exist which largely automate this process ¹⁶.

In ChIP-Seq assays, the DNA fragments are amplified, and then sequenced. The sequenced reads are mapped to a reference genome, and then summarized by a file containing the read coverage at each given base. Generally, regions with high read coverage can be identified as bound; however, biases in the assay make naive analysis unreliable. Rather, a negative control experiment is performed, and bound regions are identified by locations in which the read coverage is significantly higher than the negative control experiment. This process is called peak calling, and sophisticated tools have been developed which perform this analysis http://www.ebi.ac.uk/~anshul/public/softwareRepo/spp_package.tar.gz. As the cost of sequencing has dropped, ChIP-Seq has largely replaced ChIP-chip. For a review of ChIP-Seq and analysis challenges and tools, see Park¹⁷ and Wilbanks *et al.*¹⁸. Aleksic, *et al.*, wrote an introduction to ChIP entitled, “modENCODE Data Gentle Introduction Guides”, which is available as a public Google document (<https://docs.google.com/document/d/1-BQfalYIZ58POE29dnzIw903FMhrIwiTIorGOcTCQWg/>).

DNA Accessibility

Genes located in tightly packaged chromatin cannot be transcribed; therefore, identifying regions with low nucleosome density provides important information about the active gene set in a particular biological sample. To identify these regions, the modENCODE consortium performed DNase I hypersensitive site Sequencing (Dnase-Seq)¹⁹ in three cell lines. DNase-Seq works by introducing Dnase I into a population of lysed cells, and sequencing the resulting fragments. Since open chromatin regions are dis-proportionally digested by Dnase I, the resulting sequenced fragments provide information about DNA accessibility. The raw and sequenced reads are available from GEO (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25321>) and a browser track with the average read coverage across the genome is available at (<http://compbio.med.harvard.edu/flychromatin/>).

Mapping biases, sequencing biases, and copy number variation means that some care must be taken in the identification of open chromatin regions. Several statistical methods have been developed^{20,21,22}, but they all work by identifying regions in which the DNase-Seq read counts are significantly higher than counts from a negative control. The modENCODE consortium used standard DNA sequencing data as the negative control; positions were identified as enriched in a 300bp scanning window when compared to the DNA sequencing data¹⁵. The identified regions are available in the supplement of¹⁵. However, because chromatin region boundaries show substantial variation between different tissues and cell lines²³, it is important to ensure that samples are properly matched when using previously identified chromatin states.

Histone Modifications

Chemical modifications to the ends of histones alter chromatin structure, and this altered structure correlates with specific transcriptional processes. These modifications can occur simultaneously, and the combinatorics permit for billions of potential distinct combinations. However, by analyzing histone modification patterns, and comparing them to transcription data, patterns of combinations have emerged. To identify these patterns and elucidate their corresponding biological function, the modENCODE consortium conducted ChIP-chip and ChIP-seq experiments on 51 histone modifications in four cell lines. A select number of marks were also analyzed at several developmental timepoints. The full set of histone modification data can be found at <http://intermine.modencode.org/query/experiment.do?experiment=Genomic+Distributions+of+Histone+Modifications>.

Although identifying regions bound by histones with specific modifications is useful, it is the identification of the combinations of such modifications and their relationship to the regulation of transcriptional processes that is the most important. The modENCODE consortium performed sophisticated analysis on the various histone modification data, identifying important combinations of mark and their association with transcription. For instance, H3K27me3 is broadly anti-correlated with transcription and H3K9 methylation marks are generally associated with repetitive elements^{24,25}. The chromatin data was summarized into nine distinct states; these states and visualizations of the data are available at [<http://intermine.modencode.org/release-32/chromatinStates.do>].

One compelling observation made by the modENCODE Consortium is that combinations of chromatin marks that appear punctate and consistent over only very short linear segments on chromosomes reveal dramatically larger structural domains when visualized as planar, space-filling curves called Hilbert curves²⁴. This space-filling organization is consistent with patterns revealed by genome-wide 3D structural profiling, including the hypothesis that chromosomes are organized into fractal globules²⁶. It may be that chromatin marks are set down in three-dimensional domains within the nucleus, rather than by enzymes that “walk” across chromosomes in a fashion analogous to RNA polymerase. If true, this has important implications for the current approaches to the interpretation of these data types (e.g. via visualization in linear genome browsers). The chromatin browsing at <http://compbio.med.harvard.edu/flychromatin/displays> the chromatin data both in a linear fashion and organized in their Hilbert curve representation.

Transcription Factors

83 ChIP-Seq and 50 ChIP-chip experiments were performed in KC167 and 26 developmental time points for 61 different transcription factor proteins. The major results of these studies can be found in ^{24,27}. The sequenced and mapped read data is available from modencode.org. Identified binding regions – which use control experiments to account for assay biases - are available from the modENCODE project (<http://www.modencode.org/>).

The consortium also identified “High Occupancy Target” (HOT) regions. These are short regions bound by a large number of transcription factors, in some cases nearly every factor assayed ^{24,27}. These elements were associated with highly expressed, essential genes ²⁴, but testing in transgenic assays failed to distinguish these elements as enriched for active enhancer elements ²⁷, indicating that they may be associated primarily with regions of open chromatin rather than functional regulatory elements. Attempts to identify pairs of transcription factors with more proximal binding sites than would be expected at random ²⁵ were confounded by the inclusion of HOT regions, and so were excluded in most analyses ^{24,27}. Thus, conclusions drawn from ChIP-Seq binding sites identified in these regions should be treated with some care.

Transcriptomics

To study the diversity and dynamics of the *Drosophila* transcriptome, we performed RNA-seq, cap analysis of gene expression (CAGE), 5' rapid amplification of cDNA ends (RACE), and poly(A)+Seq in 29 tissues, 21 cell lines, and 21 treatments. An analysis of the development time course can be found in Graveley et. al ²⁸; the major findings of the comparative analysis can be found in Chen et. al. ²⁹ the major findings after analysis of the full RNA data set can be found in Brown et. al ³⁰. The data are available in various stages of processing including: sequenced reads; reads aligned to a reference genome; an inferred transcriptome; discovered splice sites; gene and transcript expression scores; inferred promoters; and inferred poly(A) sites. Most of the data (323/346 RNA-seq samples) were prepared from *D. Melanogaster*, although six other *Drosophila* species were analyzed for the purpose of comparative analysis.

RNA-seq

The bulk of the modENCODE RNA-seq data was generated using Illumina's TruSeq line of mRNA sample preparation kits. The developmental timeline poly(A)+ data was unstranded, but all subsequent RNASeq used a stranded Illumina mRNA kit. These kits take input RNA and reverse transcribe it into DNA. The DNA is optionally fragmented, adapters are ligated onto the ends of the DNA, and the library is loaded onto a flow cell for clustering and sequencing. RNA-Seq provides single-nucleotide resolution, strand specificity, and connectivity, all of which were advantages over the previously used microarray technology. Synthetic spike-in controls, which help estimate absolute RNA quantities, were used in most experiments ³¹.

The RNA-seq data was generated to provide information both about the structure of transcript isoforms and the expression of genes and transcripts in various tissues, treatments,

developmental stages, and cell lines. This data can be split into two main categories: short and long. The short RNA-seq is primarily used to identify non-coding RNA molecules under 50 bases, including micro RNAs (miRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs) and short small nuclear RNAs (snoRNAs). The short RNA libraries were typically selected for fragments under 150 nucleotides. The short RNA data are relatively simple to analyze because the reads cover the entire fragment.

The long RNA-seq data is primarily used to study protein coding transcripts, but also contains information about non-coding RNAs including long non-coding RNAs (lncRNAs) and long snRNAs. The long RNA-seq can be further subdivided into poly(A)+ and total. The poly(A)+ libraries were constructed to only contain polyadenylated transcripts – although the selection process is imperfect so some unprocessed signal does remain. Most of the long RNA-seq libraries are poly(A)+ including all of the tissues, treatments, cell lines; they are best suited for the analysis of mRNAs and functional RNAs. The total RNA-seq libraries are not purified for poly(A) transcripts, and thus contain unprocessed RNA. The primary use of these libraries is the analysis of splicing dynamics and non-polyadenylated ncRNAs, although Brown et. al.³⁰ reports that that nearly all long RNAs are polyadenylated. All of the RNA-seq data are paired end, and all but the poly(A)+ developmental timecourse data is stranded. It can be difficult to identify anti-sense transcripts or quantify overlapping genes in the unstranded data, but integrating the gene boundary data (e.g. CAGE, poly(A)-seq, and RACE) and splice motifs can help.

The raw and aligned reads are available at <http://fruitfly.org/modENCODE-transcription-project/>. Assembling long transcripts from mapped RNA-seq reads is difficult, due to a fundamental limitation in the assay that stems from the read length limitations of second generation sequencing technologies³². However, transcript elements like exons, introns and, to some extent, transcript boundaries can be directly identified from the mapped RNA-seq reads – although RNA-seq is intrinsically noisy and care must be taken to ensure that the discovered elements are not artifacts of the assays. The modENCODE consortium assembled lists of inferred exons, introns, transcript boundaries, genes, and their relative expression levels. The identified elements are available at http://wotan.lbl.gov/modENCODE_transcriptome.html.

Introns were identified directly from “junction reads”, and verified using the filtering mechanism described in Brown et. al.³⁰. Essentially, any junction reads not observed in multiple biological samples, that mapped to another location in an ungapped fashion, or that were not supported by other junction reads were removed. The corresponding introns, and their counts in each sample, are available as a bed file at the modENCODE DCC <http://fruitfly.org/modENCODE-transcription-project/>.

Exons were assembled using GRIT, and are composed of contiguously transcribed regions that are flanked by a junction, a TSS, or a TES. Exon expression estimates were calculated as the number of bases covered per kilobase million reads (bpkm), a standard measure of gene expression that should be proportional to an exon’s concentration in solution. Reads that overlapped multiple exons were assigned proportionally using a linear model, which estimated the proportionality constant under the assumption that read coverage within an

exon is uniform⁴⁴. The inferred exons, and their expression scores, are available at http://wotan.lbl.gov/modENCODE_transcriptome.html.

Genes were identified by sets of overlapping exons connected by junctions reads. Gene expression is reported and BPKM, where the total length of transcribed regions that belong to the gene are used to calculate the length. This estimate is biased for genes with multiple transcripts of different lengths³³, but is still useful for analyzing changes in gene expression.

Transcription Start Sites

The modENCODE consortium generated CAGE³⁴ and RACE³⁵ data in the embryo, and CAGE data in 16 developmental timepoints, nine dissected samples and three cell lines. This data was analyzed in combination with 5' ESTs, full-length cDNAs, and RNA-seq data to produce a comprehensive transcription start site (TSS) annotation.

The CAGE assay is very high throughput and does not require prior information about the transcriptome. However, the assay is noisy: the cap selection is not completely specific; the short read length makes mapping a problem in many loci; and the cap may be reverse-transcribed, adding a spurious glutamine to the ends of about 5% of reads. Therefore, every mapped read does not necessarily correspond to a TSS, and the modENCODE consortium used an alignment tool and statistical model to account for these problems during the identification of TSS regions. Regions that appeared to have the same distribution as reads drawn from an RNA-seq samples in the same biological sample were removed. After this filtering,³⁶ reports good correspondence with RACE and 5' ESTS, and precision within 2 basepairs of RACE and EST identified promoters. In addition to the 21 million CAGE tags developed in mixed embryo samples, the consortium generated 293 million tags in 40 samples. The CAGE protocol used to generate the tissue and cell line data appears to have less noise than the initial protocol, particularly in 3' ends of genes. The read data and bed files containing consortium identified TSS's and their relative expressions are available at http://wotan.lbl.gov/modENCODE_transcriptome.html.

In addition, the consortium generated CAGE data from the testes, ovaries, and carcass of *D. pseudoobscura*; the resulting TSSs and a comparison to the corresponding tissues in *D. melanogaster* can be found in Chen *et al.*, 2014²⁹. The raw and mapped data are available at <http://fruitfly.org/modENCODE-transcription-project/>. 5' RACE selectively amplifies copies of a region between a chosen point in known transcript and the 5' end of the transcript. The data appear to have single base resolution with little to no bias and very little noise³⁶. The primary disadvantage of RACE is that it is lower throughput than CAGE, and the location of the 5' end needs to be known within a few kilobases. As such, it is ideal for locating the 5' ends of known transcripts, but poor at discovery. The modENCODE consortium generated 1.2 million RACE tags in mixed embryo samples and compared them to 50,000 ESTs and 21 million CAGE tags from the sample samples. The major results of these experiments can be found in Hoskins *et al.* 2011³⁶.

Transcription End Sites

Poly(A)-Site-Seq³⁰ data was generated to identify and quantify polyadenylation (poly(A)) sites. In total, 54 million mapped reads were generated in 33 samples. Fastq files containing

the raw and mapped reads are available at <http://fruitfly.org/modENCODE-transcription-project/>. Like the CAGE data, the protocol is not completely specific, so some post analysis is necessary to identify poly(A) regions. The consortium used a similar filtering mechanism to that used for the CAGE data; bed files containing poly(A) regions and expression estimates are available at <http://fruitfly.org/modENCODE-transcription-project/>.

Small RNAs. The consortium generated 56 RNA-seq datasets selected for short read sizes in 28 cell lines, 8 tissues, and 12 developmental time points. Since second generation sequence are able to sequence the entire fragment, the gene boundaries can be observed directly in the RNA-seq data. The major challenge in estimating relative small RNA concentrations is lack of mappability due to the short transcript lengths, but tools that correct for mappability using a statistical model^{14,33} In general, identifying transcribed short RNAs from RNA-seq data is substantially easier than for long RNAs. The raw and mapped data can be found at data.modencode.org by selecting “small RNA” under the genomic target element heading.

MicroRNAs (miRNAs) are an important class of small RNAs (~22nt) that have been studied extensively in modENCODE. Connecting phenotypes with a particular miRNAs in a high throughput setting is the major difficulty in analysis. When target binding sites identified with computational models are analyzed systematic profiling of the transcriptome^{37,38}) and proteome^{39,40}, most targets seem to only be subtly influenced. Furthermore, most published deletions of *Drosophila* miRNAs don't exhibit any obvious phenotypes⁴¹. Despite this, there are many examples of published miRNA phenotypes, once the appropriate biological setting has been identified⁴². Bejerano *et. al.*⁴³ advocate for the study of gain-of-function phenotypes, and have created a genome wide transgenic resource for the conditional expression of *drosophila* miRNAs.

Transcript Models

The modENCODE consortium used GRIT⁴⁴ to generate and quantify transcripts models. GRIT enumerates all possible transcript models from the discovered elements, and then uses a statistical model to estimate expression values and the corresponding confidence bounds. Because of a fundamental limitation in RNA-seq experiments performed with short read sequencing technologies³², in some gene loci it is not possible to uniquely determine the set of long transcripts that are present. In such loci, some of the transcripts may have a lower confidence bound of zero, indicating that its existence can not be established with high confidence. A gtf file containing the modENCODE inferred transcriptome, and a csv file containing expression estimates, is available at the modENCODE DCC (Need the new DCC page). This annotation contains short and long transcripts.

In high complexity gene loci, experimentally identified full length transcripts are needed to validate in silico transcript models. One of the advantages of using *Drosophila Melanogaster* for studying transcriptomics is that a collection of full length transcripts exists. The BDGP's *Drosophila* Gene Collection (DGC)^{45,46} provides a ready to use set of full-length cDNA clones by which models could be validated. The DGC was produced by sampling random ESTs, aligning them to the genome, and filtering out fragments. Due to low sampling depth, the DGC missed many alternative 5' UTRs, cassette exons, and other alternative splice forms. In 2005, the BDGP began using a new method, Self Ligation of Inverse PCR

products (SLIP)⁴⁷, to validate Flybase gene models identified by modENCODE RNA-seq data in a targeted manner⁴⁸. This validation set includes ~2000 cDNA clones, ~200 RT-PCR clones, and ~2000 existing cDNAs which were not previously identified.

Accessing Data

ModENCODE DCC

The main portal to accessing modENCODE data can be found at www.modencode.org. The site is managed by the modENCODE DCC, whose primary objective is to collect all datasets and organize them in a way that is easily accessible to the general research community. To this end, the DCC developed metadata standards to facilitate data organization. The DCC provides five ways to access data: a faceted viewer; categorical browsing; keyword searching; bulk download; and Amazon Cloud services. The faceted viewer allows the user to filter datasets on different criteria, such as organism, project category, technique, principal investigator, strain, tissue, and many others. Browsing by category provides a more general starting point, allowing the user to focus on the Transcriptome, for example. Keyword queries are useful if the user has a specific set of interests. Keyword searching is available through the modMine website (<http://intermine.modencode.org/>)⁴⁹, which is displayed on the front page of modencode.org. Should a user have adequate local computational and storage resources, bulk downloads of datasets can be performed from the modENCODE FTP site (<ftp://data.modencode.org/>). Otherwise, users can elect to perform analyses using the Amazon Cloud (see <http://data.modencode.org/modencode-cloud.html> for help). A quick guide to using modMine is found in the help section of the website.

Reagents and protocols used to generate the modENCODE datasets can be obtained using the modMine query builder or general search feature. Note, modMine search results may provide links to the reagents and protocols sections of the modENCODE Wiki (wiki.modencode.org), which are normally closed to the public. To access the pages linked from modMine, follow the login instructions for Anonymous users.

To view modENCODE data for a specific gene (e.g. *Abd-B*) type “*Abd-B*” in the search bar on the modMine home page. Choose the top hit for the “gene”, which directs the user to a gene report page. The modENCODE expression data is available in the “Expression Scores” section. Data can be visualized using the modENCODE Genome Browser, which runs GBrowse (Figure 1). The Genome Browser allows users to choose how and what to see, under the “Select Tracks”, “Community Tracks”, “Custom Tracks”, and “Preferences” sections.

Sequencing Data Repositories

The **Sequence Read Archive (SRA)**⁵⁰, formerly known as the Short Read Archive, is a database which stores sequences from “next-generation” (2nd generation and beyond) sequencing platforms. The SRA stores both the raw sequence data and associated metadata. The SRA metadata model consists of six XML objects, each constrained by a hierarchical schema. The XML objects are submission, study, experiment, run, sample, and analysis; each XML object is assigned a unique accession prefix. Publications may reference a

particular submission (SRA), and may further reference specific studies (SRP), experiments (SRX), runs (SRR), samples (SRS), or analyses (SRZ) used, which could include the entire dataset or just a portion of it. Data originating from NCBI begins with an “S”. In addition, one may encounter accessions beginning with “E” (from EMBL) or “D” (from DDBJ-SRA). Fastq files can be extracted from the SRA archive using fastq-dump, which is distributed as part of the SRA Toolkit⁵⁰.

The **Gene Expression Omnibus (GEO)**⁵¹ was created by the NCBI in 2000, and was originally designed for gene expression studies, but it studies involving gene regulation, epigenetics, and functional genomics. GEO stores both micro-array data and sequence data. GEO hosts raw, processed, and analyzed sequence data, while the SRA only hosts raw sequence data. Raw sequence data submitted to GEO is funneled to the SRA and assigned an SRA accession number, so it is common to find cross-repository accession numbers in both GEO and the SRA. The GEO data model consists of three user-submitted records, each of which is assigned a unique accession prefix: Platform (GPL), Samples (GSM), and Series (GSE)⁵². These are further refined by GEO into DataSet (GDS) and Profile records⁵³.

FlyBase

FlyBase⁵⁴ is the main portal to the genomic and genetic information of *Drosophila* species. It integrates data from the worldwide community and provides broad information such as: gene boundaries; transcript data; polypeptide data; fly stock availability; clone resources; literature citations; ontologies; cytology information; known interactions and pathways; expression data; and links to other databases and resources. Guides and tutorials for using FlyBase are found in the “Help” section of the website. FlyBase is updated as new data become available; new releases are issued about every month. This means gene model annotations can frequently change, which affects gene naming and transcript isoforms. FlyBase assigns unique identifiers to genes (FBgn), transcripts (FBtr), and polypeptides (FBpp), but the identifiers may be updated between releases. These updates, while necessary, can cause confusion so specifying the particular FlyBase release that was used for a particular analysis is important. When FlyBase identifiers change, the old identifier or gene name is kept in a synonym field. FlyBase displays the modENCODE RNA-Seq data in the “Expression Data:High-Throughput Expression Data” subsection of the FlyBase gene report page (Figure 2). Users can quickly access a specific gene’s report page by using the “Jump to Gene” search bar. Alternatively, the “Quick Search” box provides a more powerful way of locating genomic elements. For instance, searching for “Abd-B” in the “Quick Search” box yields >2000 hits that span various element types. Selecting the “Genes” data class returns the 108 genes related to Abd-B. Elements for species other than *D. melanogaster* are prefixed with “D”, followed by the first three letters of the species name, and terminated with a backslash (i.e. the *D. erecta* ortholog appears as “Dere\Abd-B”).

The modENCODE RNA-Seq data is divided into four groups: anatomy, developmental stage, cell lines, and treatments. The FlyBase gene report page provides a general, high-level view of the expression data (Figure 3). If the user is interested in a more detailed, exon-level view, FlyBase also hosts its own version of GBrowse: choose the “*D. melanogaster* Expression/Regulation” Data Source.

Users can load the complete FlyBase database for the current and previous versions under the “Files” section of the FlyBase website. Users will also find a description of the files and how to use them. In particular, the “Release Notes” for a particular release and species contains information such as: the sequenced strain; known mutations that differ from the sequenced strain; general statistics regarding gene counts, transcript counts, and other annotation elements; a summary of the changes relative to the prior release; and information on the current genome assembly.

For the modENCODE project, RNA-Seq and RT-PCR data, along with multi-species sequence conservation analyses, were used to predict targets for SLIP^{47, 48}. Clone sequences obtained from these screens were used to validate or refine gene models, continuing the iterative process of improving the genome annotation. cDNA sequences were submitted to GenBank and the clones are available from the DGRC. Clone information is in the “Stocks and Reagents” subsection of the gene report page.

Visualization

Data visualization remains a challenge, particularly given the enormous datasets that are generated. The need to have a single, concise, and simple method of seeing data for a particular gene locus, for example, is often hindered by the desire to see all data for that region. For instance, RNA-seq data for analyzing the *Drosophila* transcriptome interrogated 30 different developmental stages alone. Much of that data contains strand information, which essentially doubles the amount of screen real estate needed. Adding data for different tissues and environmental perturbations further increases the complexity. Several genome browsers have been written, such as Gbrowse/Gmod⁵⁵ (Figure 1) – available at the modENCODE website, the UCSC Genome Browser⁵⁶ (Figure 4), JBrowse^{57,58}, and others, but none seem to fully solve the problem. Peter Park’s group at Harvard⁵⁹ (<http://compbio.med.harvard.edu/flychromatin/>) created an alternative method to show ChIP-seq with RNA-seq data.

Stock Centers

There are a number of centers around the world dedicated to the maintenance and distribution of *Drosophila* fly stocks. These range from public to private, for- and not-for-profit, organizations. A partial list can be found at FlyBase, under the Resources section. The most extensive collection of fly stocks can be found at Indiana University’s Bloomington Stock Center (BSC, <http://flystocks.bio.indiana.edu/>). Bloomington’s website is also a good resource for fly food recipes and sources of supplies for fly maintenance. The *D. melanogaster*, y1; cn1 bw1 sp1, strain was used by the BDGP to produce the genome sequence, and is available from the BSC, Kyoto Stock Center (<http://www.dgcr.kit.ac.jp/en/index.html>), and the UCSD Stock Center (<https://stockcenter.ucsd.edu>). Other strains used in the modENCODE project include Oregon-R, Canton-S, *D. biarmipes*, *D. eugracilis*, *D. ficusphila*, *D. takahashii*, *D. elegans*, *D. rhopaloa*, *D. bipectinata*, and *D. kikkawai*. Protocols for maintaining fly stocks and harvesting flies from different developmental stages can be found on the modENCODE Wiki (wiki.modencode.org)

Acknowledgments

We thank the members of the modENCODE transcription consortium and especially Ben Brown for helpful discussions. This work was funded by a contract from the National Human Genome Research Institute modENCODE Project, contract U01 HG004271, to S.E.C. (Principal Investigator) under Department of Energy contract no. DE-AC02-05CH11231. NB and PB (Principle Investigator) were support by 1U01HG007031-01 and the ENCODE DAC 5U01HG004695-04.

References

1. C. e. S Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*. 1998; 282:2012–2018. [PubMed: 9851916]
2. Adams MD, et al. The genome sequence of *Drosophila melanogaster*. *Science*. 2000; 287:2185–2195. [PubMed: 10731132]
3. Celniker SE, et al. Finishing a whole genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biology*. 2002; 3:research0079. [PubMed: 12537568]
4. Hoskins RA, et al. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science*. 2007; 316:1625–1628. [PubMed: 17569867]
5. Celniker SE, et al. Unlocking the secrets of the genome. *Nature*. 2009; 459:927–930. [PubMed: 19536255]
6. modENCODE et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*. 2010; 330:1787–1797. [PubMed: 21177974]
7. Washington NL, et al. The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details. *Database (Oxford)*. 2011; 2011:bar023. [PubMed: 21856757]
8. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2009; 11:31–46. [PubMed: 19997069]
9. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008; 26:1135–1145. [PubMed: 18846087]
10. Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*. 2013; 6:287–303.10.1146/annurev-anchem-062012-092628 [PubMed: 23560931]
11. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*. 1975; 94:441–448. [PubMed: 1100841]
12. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science*. 1998;281, 363–365. [PubMed: 9765151]
13. Kaczorowski T, Szybalski W. Automated four-color DNA sequencing using primers assembled by hexamer ligation. *Gene*. 1996; 179:195–198. S0378-1119(96)00325-3 [pii]. [PubMed: 8955647]
14. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci U S A*. 2011; 108:19867–19872. [PubMed: 22135461]
15. Kharchenko PV, et al. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*. 2011; 471:480–485. [PubMed: 21179089]
16. Toedling J, Huber W. Analyzing ChIP-chip data using bioconductor. *PLoS Comput Biol*. 2008; 4:e1000227. [PubMed: 19043553]
17. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009; 10:669–680. [PubMed: 19736561]
18. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*. 2010; 5:e11471.10.1371/journal.pone.0011471 [PubMed: 20628599]
19. Crawford GE, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*. 2006; 16:123–131. [PubMed: 16344561]
20. Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to 11 controls. *Nat Biotechnol*. 2009; 27:66–75. [PubMed: 19122651]
21. Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. 2011; 12:139. [PubMed: 21554709]

22. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008; 9:R137. [PubMed: 18798982]
23. Riddle NC, et al. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.* 2011; 21:147–163. [PubMed: 21177972]
24. ModENCODE Consortium et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science.* 2010; 330:1787–1797. [PubMed: 21177974]
25. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
26. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009; 326:289–293. [PubMed: 19815776]
27. Negre N, et al. A cis-regulatory map of the *Drosophila* genome. *Nature.* 2011; 471:527–531. [PubMed: 21430782]
28. Graveley BR, et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature.* 2011; 471:473–479. [PubMed: 21179090]
29. Chen Z-X, et al. Comparative Validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res.* 2014
30. Brown JB. Diversity and dynamics of the *Drosophila* transcriptome. *Nature.* 2013
31. Jiang L, et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 2011; 21:1543–1551. [PubMed: 21816910]
32. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics.* 2009; 25:1026–1032. [PubMed: 19244387]
33. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010; 28:511–515. [PubMed: 20436464]
34. Takahashi H, Lassmann T, Murata M, Carninci P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc.* 2012; 7:542–561. [PubMed: 22362160]
35. Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A.* 1988; 85:8998–9002. [PubMed: 2461560]
36. Hoskins RA, et al. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* 2011; 21:182–192. [PubMed: 21177961]
37. Lim LP, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature.* 2005; 433:769–773. [PubMed: 15685193]
38. Giraldez AJ, et al. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science.* 2006; 312:75–79. [PubMed: 16484454]
39. Baek D, et al. The impact of microRNAs on protein output. *Nature.* 2008; 455:64–71. [PubMed: 18668037]
40. Selbach M, et al. Widespread changes in protein synthesis induced by microRNAs. *Nature.* 2008; 455:58–63. [PubMed: 18668040]
41. Smibert P, Lai EC. Lessons from microRNA mutants in worms, flies and mice. *Cell Cycle.* 2008; 7:2500–2508. [PubMed: 18719388]
42. Dai Q, Smibert P, Lai EC. Exploiting *Drosophila* genetics to understand microRNA function and regulation. *Curr Top Dev Biol.* 2012; 99:201–235. [PubMed: 22365740]
43. Bejarano F, et al. A genome-wide transgenic resource for conditional expression of *Drosophila* microRNAs. *Development.* 2012; 139:2821–2831. [PubMed: 22745315]
44. Boley N, et al. Genome guided transcript construction from integrative analysis of RNA 12 sequence data. *Nature Biotechnology.* 2014
45. Stapleton M, et al. The *Drosophila* Gene Collection: Identification of Putative Full-Length cDNAs for 70% of *D. melanogaster* Genes. *Genome Res.* 2002; 12:1294–1300. [PubMed: 12176937]
46. Stapleton M, et al. A *Drosophila* full-length cDNA resource. *Genome Biology.* 2002; 3:research0080. [PubMed: 12537569]

47. Hoskins RA, et al. Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res.* 2005; 33:e185. [PubMed: 16326860]
48. Wan KH, et al. High-throughput plasmid cDNA library screening. *Nat Protoc.* 2006; 1:624–632. [PubMed: 17406289]
49. Contrino S, et al. modMine: flexible access to modENCODE data. *Nucleic Acids Res.* 2012; 40:D1082–1088. [PubMed: 22080565]
50. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011; 39:D19–21. [PubMed: 21062823]
51. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res.* 2011; 39:D1005–1010. [PubMed: 21097893]
52. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30:207–210. [PubMed: 11752295]
53. Barrett T, et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res.* 2005; 33:D562–566. [PubMed: 15608262]
54. McQuilton P, St Pierre SE, Thurmond J. FlyBase 101--the basics of navigating FlyBase. *Nucleic Acids Res.* 2011; 40:D706–714. [PubMed: 22127867]
55. Stein LD. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform.* 2013; 14:162–171. [PubMed: 23376193]
56. Goldman M, et al. The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res.* 2013; 41:D949–954. [PubMed: 23109555]
57. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. *Genome Res.* 2009; 19:1630–1638. [PubMed: 19570905]
58. Westesson O, Skinner M, Holmes I. Visualizing next-generation sequencing data with JBrowse. *Brief Bioinform.* 2012; 14:172–177. [PubMed: 22411711]
59. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.* 2008; 26:1351–1359. [PubMed: 19029915]

Figure 1. modMine Genome Browser view of the *Abdominal-B* (*Adb-B*) gene region. 1) Sections for browser customization. 2) Data displayed in tracks. 3) Individual track features.

The image shows a screenshot of the FlyBase website for the gene *Abd-B* in *D. melanogaster*. The page is annotated with red circles and numbers 1 through 6. Annotation 1 points to the top header area showing the FlyBase logo and the gene name. Annotation 2 points to the navigation bar with links like Home, Tools, Files, Species, Documents, Resources, News, Help, and Archives. Annotation 3 points to the 'Jump to Gene' search bar. Annotation 4 points to the 'View in GBrowse' link under the 'Genomic Maps' section. Annotation 5 points to the 'Get genome region' and 'Get FastA' links in the 'Genomic Maps' section. Annotation 6 points to the 'Gene Ontology (GO)' section, which lists various biological processes and molecular functions.

General Information

Symbol	Dmel\Abd-B	Species	<i>D. melanogaster</i>
Name	Abdominal B	Annotation symbol	CG11648
Feature type	protein_coding_gene	FlyBase ID	FBgn0000015
Gene Model Status		Stock availability	41 publicly available
Also Known As	AbdB, Mcp, iab-7, Fab-7, iab-5, DROABDB, iab-6, iab-8, tuh-3, Fab-6		

Genomic Location

Cytogenetic map	89E4-89E5	Sequence location	3R:12,752,932..12,797,958 [-]
-----------------	-----------	-------------------	-------------------------------

Genomic Maps

Select View: [Gene Models/E](#) [View in GBrowse](#) [GBrowse 2](#)

Decorated FastA [Get genome region](#) [Gene region](#) [Get FastA](#)

Families, Domains and Molecular Function

Protein Family (UniProt, Sequence Similarities)
Belongs to the Abd-B homeobox family. (P09087)

Protein Domains/Motifs
[UniProt \(Sequence Similarities\)](#)
Contains 1 homeobox DNA-binding domain. (P09087)
[InterPro](#)
Homeobox domain; Homeodomain-like; Homeobox, conserved site; Homeodomain, metazoa

Molecular Function (see GO section for details)
Experimental Evidence
[RNA polymerase II distal enhancer sequence-specific DNA binding](#); [sequence-specific DNA binding transcription factor activity](#); [transcription regulatory region sequence-specific DNA binding](#)
Predictions/Assertions
[sequence-specific DNA binding transcription factor activity](#)

Gene Ontology (GO): (39 terms)

- ☒ Molecular Function (3 terms)
- ☒ Biological Process (35 terms)
- ☒ Cellular Component (1 term)
- ☒ Summaries
- ☒ Recent Updates
- ☒ Interactions and Pathways
- ☒ Expression Data
- ☒ Transcript Expression
- ☒ Polypeptide Expression
- ☒ Expression Deduced from Reporters
- ☒ High-Throughput Expression Data

Associated Tools
[View Dmel\Abd-B in GBrowse 2](#) [RNA-Seq search](#)

Reference
See [Gelbart and Emmert, 2013](#) for analysis details and data files for all genes.

FlyAtlas Anatomy Microarray
modENCODE Anatomy RNA-Seq

Figure 2.

FlyBase gene report for *Abdominal-B* (*Abd-B*). 1) Current FlyBase version and Gene name. 2) Navigation bar for different sections: “Tools” section is useful for custom queries, batch downloads, and identifier conversion; “Files” section is useful for downloads; “Documents” section contains the Release Notes and Reference Manual; “Archives” section contains prior versions. 3) “Jump to Gene” bar, enter exact gene name here. 4) Link to GBrowse, choose the “Expression/Regulation” View to see modENCODE RNA-Seq data. 5) Links to download sequence files. 6) Other gene information, including high-level view of modENCODE Expression Data”.

Figure 3. Continuation of FlyBase gene report page for Abdominal-B (Adb-B), showing expanded “Expression Data” and “High-throughput Expression Data” subsections.

Figure 4.

UCSC Genome Browser, displaying a custom track hub for the *Abdominal-B* (*Abd-B*) gene. The “My Data” menu allows users add custom sessions, tracks, and track hubs, Custom tracks are the simplest way to add custom tracks, but are removed automatically after a few weeks if they have not been accessed. Track hubs can be hosted on the user’s local servers and thus the data will not be automatically removed.

Figure 5.
Heat map visualization of modENCODE ChIP-Seq (left) and RNA-Seq (right) data.

Table 1

Whole Genome Sequencing Experiments

Species	SRA Accession	Genbank ID
D. biarmipes	SRP007984	Dbia_1.0
D. eugracilis	SRP007991	Deug_1.0
D. ficusphila	SRP008002	Dfic_1.0
D. takahashii	SRP008019	Dtak_1.0
D. elegans	SRP008020	Dele_1.0
D. rhopaloa	SRP008021	Drho_1.0
D. bipectinata	SRP008024	Dbip_1.0
D. kikkawai	SRP008029	Dkik_1.0