Lessons from modENCODE

James B. Brown[1,2] and Susan E. Celniker[2]

[1] Department of Statistics, University of California, Berkeley, California 94720; email: benbrownofberkeley@gmail.com

[2] Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, California 94720; email: celniker@fruitfly.org

Annu. Rev. Genomics Hum. Genet. 2015. 16:31–53

First published online as a Review in Advance on June 26, 2015

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi:
10.1146/annurev-genom-090413-025448

Keywords

Drosophila melanogaster, *Caenorhabditis elegans*, transcription, replication, epigenetics, regulation of gene expression

Abstract

The modENCODE (Model Organism Encyclopedia of DNA Elements) Consortium aimed to map functional elements—including transcripts, chromatin marks, regulatory factor binding sites, and origins of DNA replication—in the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*. During its five-year span, the consortium conducted more than 2,000 genome-wide assays in developmentally staged animals, dissected tissues, and homogeneous cell lines. Analysis of these data sets provided foundational insights into genome, epigenome, and transcriptome structure and the evolutionary turnover of regulatory pathways. These studies facilitated a comparative analysis with similar data types produced by the ENCODE Consortium for human cells. Genome organization differs drastically in these distant species, and yet quantitative relationships among chromatin state, transcription, and cotranscriptional RNA processing are deeply conserved. Of the many biological discoveries of the modENCODE Consortium, we highlight insights that emerged from integrative studies. We focus on operational and scientific lessons that may aid future projects of similar scale or aims in other, emerging model systems.

INTRODUCTION

The modENCODE (Model Organism Encyclopedia of DNA Elements) Consortium collected substantial amounts of data to interrogate the genomes of the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*. Chromatin states, transcription factor (TF) binding sites, origins of replication (ORs), and RNA transcripts were mapped and quantified genome-wide in a wide variety of cell types, tissues, developmental time points, and environmental conditions (**Figure 1**). As with the ENCODE Consortium, the primary aim was to discover and provide biologically informative characterizations of as many genomic elements as possible.

The use of model organisms enabled a variety of experimental approaches not tenable in humans: conducting developmental time courses in isogenic populations, performing environmental perturbations, and validating findings in the most powerful metazoan genetic systems yet developed. These studies facilitated a number of biological insights, including the following:

- Quantitative relationships between chromatin state and transcript abundance are conserved among flies, worms, and humans.
- Di- and trimethylation of histone H3 at lysine 36 (H3K36me2/me3) couple chromatin state and splicing and transmit epigenetic memory of maternal transcription to the embryo.
- In neural tissue, some genes encode thousands of transcript isoforms, including extended 3′ untranslated regions (UTRs) as long as 18 kb in *Drosophila* and mammals.
- Male gonad-specific antisense transcripts exist at orthologous loci in *Drosophila* and mammals, suggesting that some aspects of transcriptional organization may be basal to spermatogenesis in metazoans.

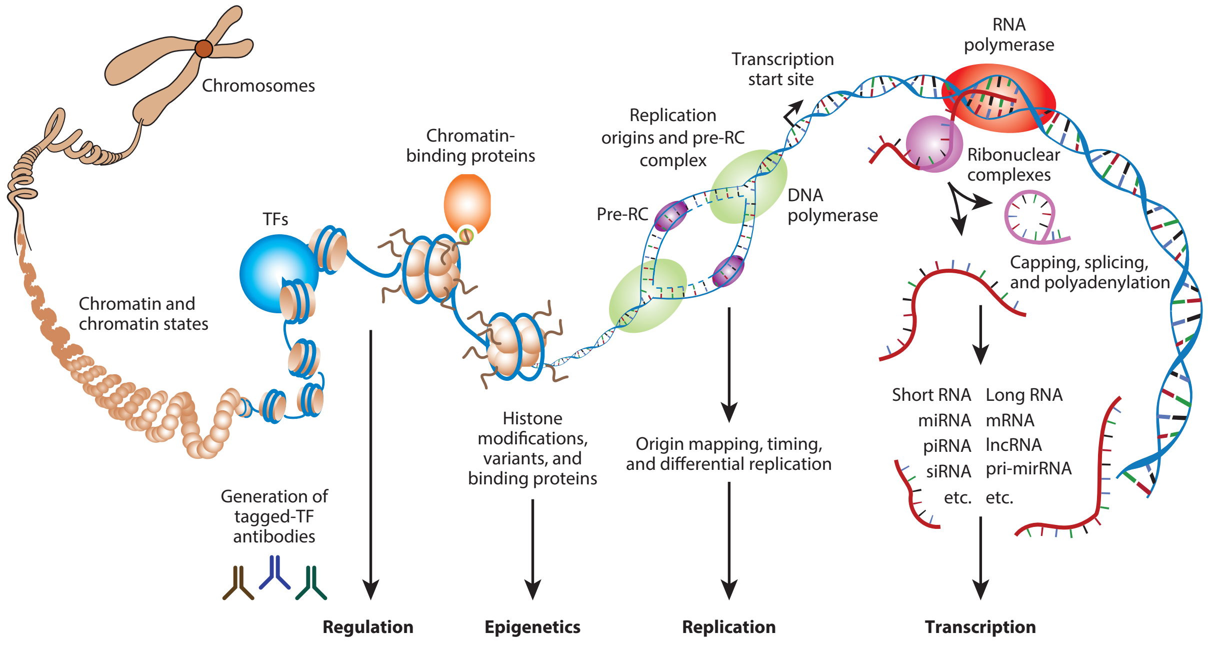
Chromosomes
Chromatin and chromatin states
TFs
Chromatin-binding proteins
Generation of tagged-TF antibodies
Histone modifications, variants, and binding proteins
Replication origins and pre-RC complex
Pre-RC
DNA polymerase
Transcription start site
RNA polymerase
Ribonuclear complexes
Capping, splicing, and polyadenylation
Origin mapping, timing, and differential replication
Short RNA
miRNA
piRNA
siRNA
etc.
Long RNA
mRNA
lncRNA
pri-mirRNA
etc.
Regulation
Epigenetics
Replication
Transcription

Figure 1

A graphical overview of the transcript, chromatin, replication, and regulatory factor data sets for *Drosophila melanogaster* and *Caenorhabditis elegans*. Abbreviations: lncRNA, long noncoding RNA; miRNA, microRNA; piRNA, *piwi*-interacting RNA; pri-miRNA, primary microRNA; RC, replication complex; siRNA, small interfering RNA; TF, transcription factor. Adapted from Reference 19; original figure created by Darryl Leja, National Human Genome Research Institute.

In addition to a necessarily incomplete review of the science—the consortium published more than 200 papers—here we outline the foundational lessons that can now inform future research directions.

THE modENCODE DATA RESOURCE: MAPS OF CHROMATIN MARKS, DNA-BINDING PROTEINS, AND TRANSCRIPTION IN TWO METAZOAN PHYLA

The modENCODE Project generated genomic and epigenomic data sets on a scale unprecedented in the study of model organisms, including more than 2,000 assays and more than 50 billion sequenced reads (**Figure 2**). The genomes of eight drosophilids and one worm species were assembled to provide evolutionary context for mapped genomic elements. Similarly, the sequences of 19 cell lines provide an atlas of copy-number and structural variation in many distinct cellular lineages (83).

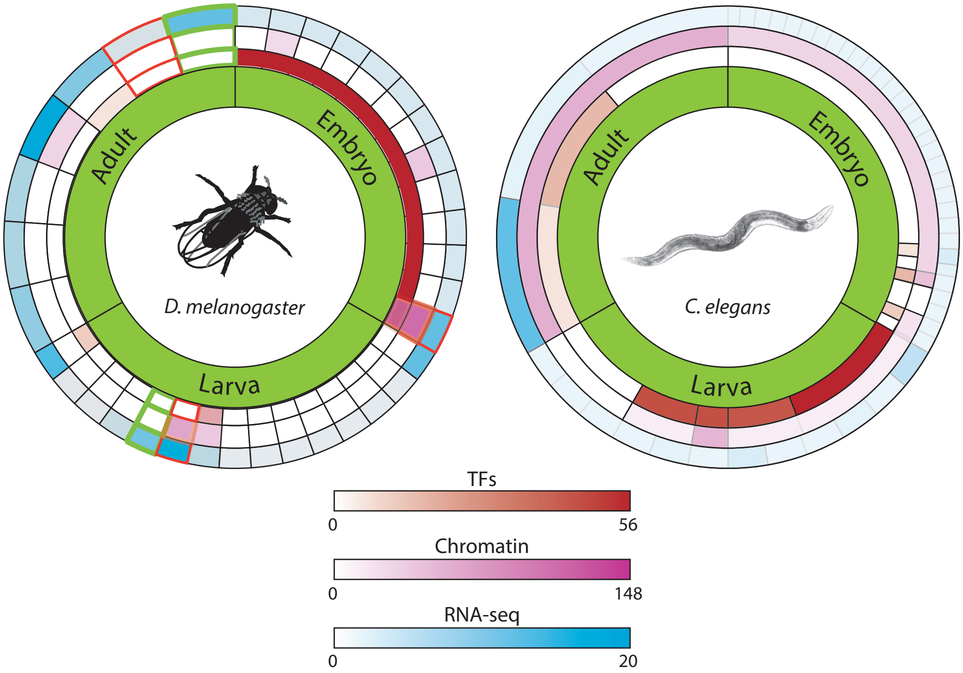

Figure 2

The landscape of modENCODE assays conducted in *Drosophila melanogaster* and *Caenorhabditis elegans*. Colors correspond to heat map intensities illustrating the type and number of biological samples assayed. Individual biological and technical replicates have been collapsed and counted as a single sample. When microarray and sequencing readouts existed for the same experiment, only the sequencing-based assay was counted. In total, 680 distinct assays were conducted in *D. melanogaster*, and 514 were conducted in *C. elegans*. Individual boxes coradial with the developmental stage labels correspond to time points, cell lines (*outlined in red*), or environmental perturbations (*outlined in green*). The entire modENCODE data set is available for analysis in the Amazon compute cloud (**http://www.modencode.org**), and individual data sets can be found at the Sequence Read Archive (**http://www.ncbi.nlm.nih.gov/sra**). Note that additional assays were conducted in other species of both flies and worms. Abbreviations: RNA-seq, RNA sequencing; TF, transcription factor.

COMMON PATTERNS OF CHROMATIN MODIFICATIONS ACROSS METAZOANS

Chromatin immunoprecipitation–on–chip (ChIP-chip) and chromatin immunoprecipitation sequencing (ChIP-seq) experiments were conducted for 34 chromatin modifications in flies and 35 in worms, in many cases at multiple developmental stages and in flies in three cell lines. Combinatorial patterns of histone marks were recovered in both species, many with a striking resemblance to those previously reported in humans; marks appear to function in globally similar fashions, e.g., H3K27me3 is broadly anticorrelated with transcription, and H3K9me marks are generally associated with repetitive elements (9, 53, 74, 87, 96). Patterns of histone modifications at promoters were highly similar and deeply conserved in all three species and more divergent at enhancer elements. Notably, human cell lines also show substantial variation in average chromatin state around enhancer elements, suggesting that these marks may themselves be subject to additional epigenetic context, possibly including higher-order interactions between marks at the same nucleosome. A few such interactions have been examined in detail—for example, histone H4R3me2 is weakly repressive via the recruitment of DMNT3A and subsequent DNA methylation (139) but can become strongly activating via the recruitment of acetyltransferase to H3 and H4 lysines (68).

Additionally, modENCODE demonstrated several strong cell-type-specific associations with various histone deacetylase proteins (53, 96). Genome-wide averages may be expected to show cell-type-specific behavior as a consequence of the expressed complement of other chromatin-modifying proteins and complexes. The global similarity of average patterns of chromatin modifications at promoters and enhancers across 600 million years of evolutionary divergence—comparing human cell lines with, in the case of *C. elegans*, whole-animal ChIP experiments—is striking, and indicates that although considerable combinatorial complexities may remain to be discovered in metazoan chromatin regulation, the main effects of individual marks are sufficiently strong to inform our understanding of genome organization.

A particularly compelling example of this comes from the observation that combinations of chromatin marks that appear to be punctate and consistent over only very short, linear chromosomal segments reveal dramatically larger structural domains when visualized as planar, space-filling curves called Hilbert curves (74). This space-filling organization is consistent with patterns revealed by genome-wide three-dimensional structural profiling, including the hypothesis that chromosomes are organized into fractal globules (a broader class of curves that can be modeled as Hamiltonian walks) (86). Chromatin marks are likely established in three-dimensional domains within the nucleus rather than by enzymes that traverse chromosomes, such as RNA polymerase.

Heterochromatic boundaries are surprisingly plastic across various cell types (113). In *C. elegans*, heterochromatic domains are maintained in close association with the nuclear lamina via the membrane-bound protein LEM-2 (69), and the relationship between laminal proximity and gene silencing appears to be a general property of metazoan cells (6). Genes embedded in euchromatic islands within otherwise silenced, usually repeat-rich domains are often highly expressed (69, 112), and in flies, on chromosome 4, they may be regulated via a novel mechanism specific to repeat-rich regions (112).

Chromatin signatures of ORs were also identified (42). Early-firing ORs were strongly associated with activating chromatin marks [e.g., H3K4me, H3K18 acetylation (H3K18ac), and H3K27ac] and late-firing ORs with repressive marks (e.g., H3K9me2 and H3K9me3). No single mark was predictive of replication timing, but joint analysis of many marks enabled significant prediction accuracy, consistent with the view that combinatorial epigenetic states are necessary to specify and coordinate replication timing (42). Analysis across three fly cell lines—S2, KC157 (embryonic), and BG3 (neuronal)—confirmed that these patterns are broadly consistent across disparate cell states. Some cell-type-specific quantitative variation was observed, indicating that the local chromatin environment may function as a rheostat to tune replication timing (42). Furthermore, in flies, RNA interference (RNAi) depletion of *geminin* results in preferential rereplication of heterochromatic but not euchromatic sequences owing to the formation of pre-replication complexes (pre-RCs) outside of G_1 phase (35), suggesting that pre-RC assembly is also subject to context-specific controls. These results are broadly similar to those of the ENCODE Consortium for human cell lines, and may point toward general metazoan mechanisms of coordination for DNA replication timing and initiation (12).

ORGANIZATIONAL FEATURES OF REGULATORY FACTOR BINDING SITES

The modENCODE Consortium surveyed TFs (2, 14, 53, 98) and identified several high-occupancy target (HOT) regions. These genomic elements are bound by a large number of TFs—in some cases nearly every factor assayed (53, 96)—and are associated with highly expressed, essential genes (53, 96). In both *C. elegans* and mammals, the most extreme HOT regions correspond to nonmethylated CpG-dense promoters, where CXXC finger protein 1 recruits SETD1A methyltransferase for H3K4me3 (23). Attempts to identify pairs of TFs with more proximal binding sites than would be expected at random (coassociation analysis, e.g., as in Reference 9) were confounded by the inclusion of HOT regions, and these were excluded in most analyses (53, 96). HOT regions had been previously identified in both flies (97) and humans (e.g., 9), but their status as functional elements remains unclear. In the *Drosophila* blastoderm, many canonical enhancer elements, such as the *eve* stripe 2 enhancer, are clearly HOTs by the modENCODE definition (90). Analysis of the quantitative binding profiles of 21 TFs in the early embryo via a transgenic enhancer assay revealed that the majority of the HOT elements tested are functional (49).

Are regions of apparently universally (or at least broadly) open chromatin merely incidentally bound? At these regions, are ChIP experiments merely providing a readout of the chromatin state, not unlike a DNase experiment? Because these TFs are present in tens of thousands to millions of molecules per cell and most TFs are bound to chromatin most of the time (reviewed in 10), both seem plausible. This is also consistent with the ENCODE Consortium's observation that transcription levels can be predicted with surprising accuracy from ChIP-seq data using small TF cohorts (9, 40), and this predictive power is comparable to that obtained from DNase sequencing (DNase-seq) and histone modification data (9). In all three species, randomly selected cohorts of factors provide powerful predictive accuracy, presumably ruling out a strong causal linkage between any given small set and the accurately predicted genes (14). This observation underscores the challenges ahead if we are to identify functional, quantitative gene-regulatory networks from genome-wide data sets. The masking of HOT regions made it possible to identify several compelling potential regulatory actions [e.g., *CREBA* as a cell-type-specific activator in flies (53, 96) and PHA-4 as a master regulator of starvation response in worms (140)], but validation assays are needed to confirm functional elements and relationships [e.g., cobinding regions of EN and GRO to the DMX enhancer of *distalless* (53, 96)] and to elucidate their dynamics.

HOT regions may correspond to artifacts in the ChIP protocol, and this possibility has been assessed in yeast (106, 125) and more recently in flies (73). A modified ChIP approach called ORGANIC (occupied regions of genomes from affinity-purified naturally isolated chromatin) involves affinity purification of micrococcal nuclease-digested non-cross-linked chromatin. From such studies, much of the correlation between imputed TF binding sites and open chromatin disappears. Developing parallel assays to detect TF binding sites genome-wide is of considerable interest. If ORGANIC or a similar assay is widely adopted, it will be interesting to revisit the concept of high-occupancy genomic regions.

Coassociation analyses were performed to identify TFs with similar or overlapping binding profiles. Statistically significant overlap is often taken to suggest related or interdependent regulatory functions or pathways. Negre et al. (98) analyzed ChIP data for 38 site-specific TFs and identified dozens of significant pairwise interactions. It is known that TFs tend to bind overlapping genomic regions more often than would be expected under random binding models (12). However, Negre et al. (98) found that for some pairs of factors, the degree of coassociation is negative: They overlap much less than would be expected at random. Positive coassociation may often be attributable to correlation dependent on or induced by chromatin structure, but formulating similarly incidental explanations for negative associations seems less likely. The most strongly negatively associated pairs included factors with antagonistic roles in early development.

Challenges associated with interpreting ChIP-seq data in whole-organism experiments are further exacerbated by the inhomogeneity of cell populations. Even early embryos contain many cell types with distinctive TF and cofactor binding patterns. In *C. elegans*, two studies explored the tissue specificity of detected TF binding sites (2, 129). Araya et al. (2) analyzed 241 ChIP-seq experiments corresponding to 92 target regulatory factors [TFs, RNA polymerase II (Pol II), and chromatin-binding proteins]. For 36 of their 92 factors, they used imaging data (13 embryo, 23 larval) to generate a single-cell-resolution time-resolved gene expression map. They found that many pairs of factors with statistically significant coassociations were not expressed in the same cells or even the same tissues as their putative interaction partners, indicating that chromatin structure is similar across broad ranges of cells and tissues. This supports the idea that relatively few TFs serve as “pioneer” chromatin remodelers.

Insulator proteins bind to thousands of genomic sites and are believed to block enhancers. However, when tested in a transgenic enhancer-blocking reporter assay, the majority of elements do not appear to show significant insulator activity (98, 101). Furthermore, binding sites do not preferentially fall between active and silent genes (99, 119). Groups of insulator sites are enriched at the boundaries of H3K27me3 regions (98, 101), although only a small subset of Polycomb targets are included in the bound regions. RNAi knockdown validation assays in cell lines failed to reveal global alterations in gene expression (99, 119); only a few dozen genes (∼0.2%) were differentially expressed in these perturbed backgrounds. Additionally, the ultraconserved DNA-binding protein CTCF, present in drosophilids and mammals, shows surprisingly rapid evolution of individual binding sites, with a divergence of genome-wide binding profiles in drosophilids of ∼2.22% per million years. Most binding sites of insulator-associated proteins apparently do not function as classical insulators (99, 119). It may be that insulators, like TFs (10), have strong effects at only a small set of target genomic loci, despite being bound prolifically to thousands of sites in genomes (99, 119). The idea that insulator proteins act as master actuators of chromatin organization may be outmoded—an insight that could eventually help to elucidate the mechanism of CTCF loss in some of the nematode lineage (61).

DIVERSITY AND DYNAMICS OF METAZOAN TRANSCRIPTOMES

Extensive poly(A)$^+$ RNA sequencing (RNA-seq) was conducted in both worms and flies. Data production began with microarrays, but by the end of the first project year it had migrated to next-generation sequencing because this technology provides richer information than microarrays alone. In worms, a developmental time course was sequenced with an unstranded protocol (64), including 15-min resolution across embryogenesis. Analogous data were produced in flies (27, 57), which also included matched poly(A)$^+$ and total RNA sequencing in an embryonic time course with 2-h resolution (17).

Transcript Structure

The modENCODE Consortium identified hundreds of new genes and more than 100,000 new transcripts. RNA-seq in tissue samples and developmentally staged animals revealed that the vast majority of newly discovered transcribed elements are conserved and expressed in other drosophilids, suggesting evolutionary conservation (25). Unprecedented diversity in polyadenylation site selection was discovered for hundreds of genes (121). Several distal polyadenylation sites resulted in 3′ UTRs longer than 10 kb, and these were uniquely expressed in the nervous system. The transcription of these isoforms requires the bypassing of canonical polyadenylation signals. Parallel efforts found that the RNA-binding factor ELAV, encoded by the gene *embryonic lethal abnormal vision* (*elav*), is required for at least several neural-specific 3′ UTR extensions (63). Long 3′ UTRs have not yet been detected in *C. elegans*, but they have been found extensively in mammalian neural tissue, particularly brain tissue (94). Mammalian 3′ UTR extensions are of similar length to those in *Drosophila*, ranging in size from 10 to 20 kb. It has been suggested that these extensions may serve to incorporate additional microRNA (miRNA) binding sites to modulate posttranscriptional regulation of the parent gene. HITS-CLIP (high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation) data for Argonaut (AGO) shows some enrichment around miRNA seed matches near distal polyadenylation sites (94). Although these findings are suggestive, distinct biological roles have yet to be demonstrated. Work in worms has focused on whole-animal sequencing, and has found diverse patterns of alternative polyadenylation site selection throughout development and, remarkably, that 3′ UTRs systematically shorten with animal age, particularly toward the end of the life span (91).

Full-length transcripts were modeled using RNA sequencing data. In worms, this was facilitated by the integrative analysis of reads that cross splice-leader sequences, reads that cross polyadenylation sites, and all the reads that fall in between (64). In *Drosophila*, in addition to RNA-seq, cap analysis of gene expression (CAGE) data were collected to demarcate 5′ ends (13, 67). Gerstein et al. (54) identified a surprising transcriptional diversity in both organisms, with some genes expressing dozens of transcripts. Flies have dramatically more single-isoform genes than worms (42% versus 36%) but also more genes with exceptionally complex transcript structures (17, 54). In mammals, an estimated 95% of genes produce multiple transcript isoforms (37, 105, 130), but no ultracomplex genes [such as *Down syndrome cell adhesion molecule 1* (*Dscam1*) (e.g., 95)] have yet been identified. The computational procedure applied to the inference of the fly transcriptome, the General RNA Integration Tool (GRIT) (13; **http://grit-bio.org**), was amenable to the detection of ultracomplex splicing patterns. The set of genes with evidence of ultracomplex splicing was small: 47 genes have the capacity to encode >1,000 transcript isoforms, and these account for 50% of all transcripts detected (17). The majority of isoforms for these complex genes were detected in RNA samples from the developing embryo. The *Drosophila* in situ image collection (**http://fruitfly.org**) was used to study the localization of transcripts from these genes. The majority were detectably expressed exclusively in neural tissue. Only 10% of all fly genes are exclusively expressed in the embryonic nervous system, constituting a highly statistically significant enrichment (17). Furthermore, approximately half have long, neural-specific 3′ UTR extensions (94). Many also undergo RNA editing, and a remarkable ten ultracomplex genes are neural specific during early embryogenesis, have long 3′ UTR extensions, and undergo RNA editing. Ultracomplex RNA processing therefore seems to be largely a property of the developing nervous system. Understanding the origins of this complexity—which is likely associated with neural-specific RNA-binding proteins—constitutes an intriguing avenue for future work.

More than 1,200 novel genes were identified in *Drosophila*. We discuss the new genes in detail below, as part of our larger discussion on long noncoding RNAs (lncRNAs), as most were of this class. New protein-coding genes were limited largely to those encoding male-specific short polypeptides specific to the accessory gland, testes, or both, and likely correspond to rapidly evolving seminal fluid genes (48). The fly data set also included 21 environmental perturbations, including heavy metal poisoning, heat and cold shock, ethanol consumption, and herbicide exposure (17). These data sets revealed new lncRNAs and nearly 1,000 splice variants of known genes, indicating that alternative splicing may play a larger role in stress response and environmental adaptation than was previously appreciated.

Transcription and Chromatin

RNA and chromatin data enabled the interrogation of several long-standing phylum-specific phenomena, such as genome-wide profiling of *trans*-splicing in worms (1), and revealed a catalog of apparently pan-metazoan aspects of transcriptome organization and dynamics. Chromatin marks near promoters are predictive of steady-state levels of mRNA (as measured by RNA-seq), and most marks have qualitatively similar average profiles around promoters. A machine learning model fit to any one species predicts reasonably well in each of the others (average ∼15% drop in r^2, from ∼0.80 to ∼0.68) (54), suggesting quantitative preservation of the biological impact of chromatin modifications on transcription across phyla. Performance tends to be worse in worms, but this is likely because *trans*-splicing of the splice-leader sequence obscures the locations of promoters, and for the 70% of genes for which transcripts are *trans*-spliced, the mapping of chromatin data to presumed promoters is less accurate. Nonetheless, this is particularly notable because the ENCODE and modENCODE consortia collected these data in humans from a few transformed cell lines, in worms predominantly from whole animals, and in flies from both. Thus, not only do chromatin marks affect the transcriptional landscape of various phyla in quantitatively similar ways, but this is also likely to be true across cell types within each species. This observation has interesting implications for the potential complexity and function of a combinatorial chromatin code (31, 46); human cells, for instance, have a large complement of histone modification enzymes (∼150) and more than 100 types of distinct sites of posttranslational modification (76), but the joint contribution of the marks assayed so far can explain more than 80% of the variance in steady-state transcript levels. We note, however, that even with 80% of the variance explained, the standard error of prediction is still on the order of a log (40). Rather than acting as binary switches, much of the chromatin code may exist for the fine-tuning of rates and accessibility (reviewed in 109).

One of the most striking relationships between transcription and chromatin was found in *C. elegans*, where Kolasinska-Zwierz et al. (78) first showed that H3K36me3 selectively demarcates exons and not introns. This phenomenon is so striking that a ChIP signal displayed in a genome browser instance can be, at first glance, mistaken for RNA-seq. The exon painting structure of this mark is conserved across all metazoans surveyed, including flies, humans, and mice. This initial study was remarkable because it constituted a clear instance of feedback in transcript processing. This chromatin mark was used to accurately predict alternative splicing events (120, 141)—it is not just a static mark over constitutively transcribed exons. The complete cohort of enzymes involved in reading, writing, and erasing this mark, as well as the cofactors that recruit these enzymes, remains incomplete (e.g., 30; for a review, see 15) and continues to be an exciting area of inquiry opened by non-hypothesis-based genomics in model organisms.

Studies in worms revealed that H3K36me3 is associated with exons of genes in the embryo that were highly transcribed in the female germline, indicating a possible role for this chromatin mark in epigenetic inheritance (110). In humans, methyltransferases that set down this mark are linked to oncogenesis and developmental disorders (100, 118), but the biological roles and targets are unknown. In *C. elegans*, one orthologous enzyme, MES-4, is maternally deposited and is required in germ cells and for normal embryogenesis. It is one of at least two enzymes responsible for this mark in this species (the other being MET-1). Remarkably, Rechtsteiner et al. (110) showed that, unlike MET-1, MES-4 associates with maternally transcribed genes even in the absence of Pol II. MES-4 association is maintained at genes transcribed in the germline that are zygotically silent, and many zygotically transcribed genes are not bound by MES-4. In the embryo, the enzyme appears to associate only with genes that have preexisting, maternal H3K36 marks. Altogether, this suggests that MES-4 specifically transmits the memory of maternal transcription, rather than the incidence of Pol II association as is true with other H3K36 methyltransferases. Embryos from *mes-4* knockout mothers die rapidly, and somatic cells express germline-specific genes. Given the known role of H3K36me3 in splicing regulation and determination, this lethality may be due to genome-wide splicing dysregulation and may have important somatic as well as germline effects.

Chromatin marks acting as mediators of epigenetic inheritance is an intriguing concept, particularly in light of the extremely rapid turnover of individual histones compared with the timescale of cell division (33). Dodd et al. (39) have pointed out that the turnover of individual nucleosomes may be slow compared with the spreading or copying of marks. Multiple marks have now been assayed on single histones (21), and highly consistent patterns have been detected. Not only are individual modifications sustained despite rapid nucleosome turnover, but divalent marks are maintained (21)—a condition that requires the coordination of multiple enzymes. The complexity of chromatin structure revealed by surveying multiple marks indicates that likely higher-order valence is also consistently maintained (74). Rapid turnover of individual histones apparently is no barrier to the transmission of epigenetic information across cell divisions, and by extension, potentially, across generations.

Trans-Splicing

Chen et al. (22) made progress on mapping genome-wide transcription start sites in *C. elegans*, where ubiquitous *trans*-splicing of the SL1 leader obscures the 5′ end of most primary transcripts (1). Capped RNAs were purified in two populations [short (20–100 nucleotides) and long (>200 nucleotides)] and sequenced, and these resolved more than 70,000 clusters demarcating frequent Pol II initiation sites. Because the authors used a strand-preserving sequencing protocol, they were able to identify the direction of transcription. Ubiquitous bidirectional transcription initiation originated from the transcription start sites of protein-coding genes in the short-RNA fraction, as has been observed in *Drosophila*, mammals, yeast, and other eukaryotes (70). The strength of the antisense signal was ∼50% that of the sense signal, comparable to what has been observed in human cell lines, whereas in *Drosophila* bidirectional transcription is at least tenfold weaker (65). A remarkable finding was the representation of enhancer RNAs among regions exhibiting bidirectional transcription. The majority of clusters correspond to gene-proximal enhancers. Chen et al. (22) analyzed the overlap of short-RNA transcription start sites with 22 ChIP-derived TF maps (53) and found that a majority of binding sites (excluding HOT regions) exhibited bidirectional transcription. The long-RNA fraction revealed a fundamentally different story: As with canonical promoters, transcription elongation was primarily unidirectional. For 90% of transcribed enhancers, elongation was in the same orientation as the downstream gene. To the best of our knowledge, this finding has not been reported in another metazoan system. The authors suggested that these long transcribed enhancers may serve as alternative promoters for downstream genes or may provide Pol II loading sites on canonical promoters. Additional studies are needed to understand the three-dimensional structure of chromatin at these loci and to determine whether long enhancer transcripts have a direct or indirect biological function.

Whereas *trans*-splicing is the norm in worms, it is comparatively rare in *Drosophila*. Two well-studied *trans*-spliced genes are *longitudinals lacking* (*lola*) and *modifier of mdg4* [*mod*(*mdg4*)]. The constitutive exons of *mod*(*mdg4*) are encoded on one strand, and nine alternatively spliced exons are encoded on the opposite strand (41). The exons of *lola* are encoded on the same strand, but the 3′ last exon is frequently assembled in the mature mRNA by *trans*-splicing, sometimes between sister chromatids (66). *Trans*-splicing may be more frequent in metazoans, including vertebrates, than was previously appreciated (56). McManus et al. (92) found 80 genes in *Drosophila* that undergo interallelic *trans*-splicing. If most *trans*-splicing is within rather than between alleles, it may remain an underappreciated source of complexity in metazoan transcriptomes.

Revisiting Phylotypic Transcriptional Patterns

Because the modENCODE Consortium conducted RNA-seq in developmental time courses, it was possible to study transcriptional regulation dynamically, during embryogenesis (72). Gerstein et al. (54) identified clusters of orthologous genes with similar developmental expression patterns between the two phyla. Analysis of genes highly specific to particular developmental periods in flies and worms led to the identification of several groups of orthologs expressed in correlated developmental patterns between the species (85). These sets of genes obey the phylotypic pattern, with canalized expression during a portion of embryogenesis (54). Additionally, these clusters can be used to align the developmental time courses of flies and worms, revealing broad similarity in the patterns of embryonic gene regulation across the development of these two distant phyla. Which tissues or cell types within the animals drive the clustering remains unclear.

Small RNAs

The organization of small RNAs differs dramatically among mammals, worms, and flies. In worms, for instance, transcripts of *piwi*-interacting RNAs (piRNAs) initiate from more than 15,000 independent promoters that each give rise to a short transcript (11, 116), the so-called U21 RNAs; in mammals and flies, by contrast, most piRNAs are processed from transposable elements (TEs), lncRNAs, and the transcripts of protein-coding genes, particularly 3′ UTRs and introns. In flies, a large (5 kb) TE of the Gypsy family (DM412) exists as a recent insertion in an intron of the 3′ UTR of *Rho GTPase activating protein at 18B* (*RhoGAP18B*). This TE gives rise to thousands of distinct piRNAs distributed throughout the length of the insertion (82). Indeed, the modENCODE Consortium compared piRNAs encoded outside of TEs, lncRNAs, or mRNAs in humans, flies, and worms and found 88 loci in humans, 27 in flies, and 35,329 in worms (54). Hence, the biogenesis of these RNAs is fundamentally different in nematodes than in other deuterostomes.

Like piRNAs, endogenous small interfering RNAs (endo-siRNAs) have been implicated in AGO-dependent RNA silencing, principally in the soma (55). In flies, more than 300 regions of endogenous antisense transcription that give rise to endo-siRNAs (non-TE derived) were identified, and these are expressed at more than 100-fold-higher average levels in the male gonad compared with other tissues (17, 132). In *C. elegans*, these observations were extended to a class of endo-siRNAs called 26G RNAs (they are 26 nucleotides long, and the first base is a guanine), which were characterized as germline-specific small RNAs (60) (see also sidebar, Functional Characterization of a New Class of Small RNAs). The levels of many 26G RNAs are highly

FUNCTIONAL CHARACTERIZATION OF A NEW CLASS OF SMALL RNAs

26G RNAs fall into two expression classes: male germline (class I) and female germline (class II) (60). Biogenesis requires the ERI-1 exonuclease as well as the RNA-dependent RNA polymerase RRF-3. The targets of class I 26G RNAs are silenced during spermatogenesis; class II targets are silenced throughout development. The AGO protein encoded by *ergo-1* is required for the expression of class II RNAs. A double knockout of the AGO paralogs T22B3.2 and ZK757.3 was required to silence class I expression. Hence, the sorting of these RNAs into RNA-induced silencing complex (RISC) machinery is cell-type dependent, as are their regulatory targets. Phenotypes associated with the loss of these AGO proteins include sterility and temperature-sensitive sterility. The role of siRNAs in spermatogenesis is clearly substantial (52), but do paternal small RNAs have other functions? Do male germline RNAs transmit epigenetic memory? In mice, paternally inherited stress-induced phenotypes have been identified, and small RNAs isolated from sperm can phenocopy these when injected into an egg fertilized by a healthy male (51). There appears to be a direct interaction between chromatin state and endo-siRNAs (58). Could it be that paternally deposited small RNAs serve to transmit memory of chromatin state (135)?

dynamic in the early embryo, suggesting functional importance for the initiation of zygotic transcription (122).

TE-derived endo-siRNAs are produced at substantial levels in the ovaries. For instance, the coding sequence of *AGO2* includes a short tandem repeat that produces siRNAs at more than 50-fold-higher levels in ovaries than in testes and at 100-fold-higher levels than in heads. The core siRNA-loading factor, *r2d2*, which sorts DCR-1 and DCR-2 short-RNA products to the appropriate AGO effector (103), generates substantial levels of siRNA on both strands in most cell lines and tissues surveyed, and expression is at similar levels across the tissues (132). Several components of the RNA-induced silencing complex (RISC) may be targets of autoregulation, particularly in the ovaries. This is not surprising, as a functioning piRNA pathway is essential for the division of both the somatic and germline stem cells of the fly ovary (77). A natural hypothesis is that autoregulation in the ovaries is required to maintain tight control of TE expression that could otherwise threaten genome integrity. Indeed, TE-derived small RNAs are expressed at higher levels in the ovaries than in any other tissue. Cell lines express these too: Whereas gene-derived endo-siRNAs are absent in many immortalized fly cell lines, TE-siRNAs are present in larger quantities than in most fly tissues (132), with only the ovaries at similar levels. Most immortalized cell lines are derived from mid- to late-embryonic tissues, which do not exhibit endogenous expression of TE-siRNAs at high levels. It is intriguing to speculate that immortalized cell lines and germline stem cells may face similar challenges with respect to genome integrity—with similar adaptive solutions.

Cell lines derived from somatic ovarian stem cells in *Drosophila* include a functioning primary piRNA pathway but lack a secondary "ping-pong" pathway, where piRNAs are reciprocally amplified from both sense and antisense transcripts (82). A signature of ping-pong activity is the presence of piRNAs that overlap by exactly 10 nucleotides. The precision of this overlap is due to the mechanics of the interaction between AGO3 and Slicer that gives rise to these piRNAs. Short-RNA sequencing of cocultures of germline and somatic ovarian stem cells revealed this signature of an active ping-pong pathway—the first time this has been observed in a cultured *Drosophila* cell line (82). The modENCODE Consortium did not set out to identify new in vitro models for the female germline; rather, this discovery was an opportunistic dividend of the search for transcribed elements in short-RNA fractions.

It is also notable that TE-siRNAs are almost completely absent in samples enriched for the central nervous system. Given recent findings of widespread somatic retrotransposition in human neural tissue (5), the lack of a TE-siRNA signal in fly neural tissue suggests that a similar mechanism may be in play, or, rather, that TE transposition may be a conserved property of neural tissues. Whether this phenomenon has a biological impact or is simply tolerable in neurons remains to be seen (111).

The consortium also studied miRNAs. The *Drosophila* genome encodes ∼234 miRNAs, a limited number of which were discovered during the course of the modENCODE Project (117, 132). Many miRNAs are processed from UTRs and coding sequence of mRNAs (8). Other miRNAs originate from long (∼8 kb), unspliced, nonpolyadenylated transcripts called primary miRNAs (pri-miRNAs) (57). Pre-miRNAs are found in humans and nematodes but have very different structures: They are substantially shorter in worms (∼0.37 kb) than in flies (7.4 kb), and those of humans are even longer (>20 kb on average) (54). Other miRNAs are spliced and processed from the introns of primary transcripts. These are called mirtrons, and they bypass Drosha processing (28, 102, 115). Mirtrons first described in modENCODE have now been identified in mammals, including humans (80). The processing of miRNAs from mirtrons includes trimming and 3′ untemplated monouridylation (133). Although mirtrons exist in nematodes, insects, and vertebrates, most of these are specific to a particular lineage and are poorly conserved even between related species (7). This is a stark contrast to the slow evolution and deep conservation of Drosha-dependent pri-miRNAs, and could indicate that mirtrons tend to serve fundamentally distinct regulatory roles. Evidence suggests that mirtron-encoded miRNAs may have important and distinct functions in postmitotic neurons in mammals (4).

The structure and function of short RNAs are among the most highly conserved properties of metazoan genomes (54). However, the primary transcripts from which they are processed differ profoundly, from the unusual structure of piRNA transcription in worms to the long pri-miRNAs in flies and humans.

Long Noncoding RNAs

The eukaryotic genome is dynamically transcribed into different RNA families, most of which do not code for proteins. Long-RNA species lacking open reading frames (ORFs) longer than 100 amino acids are generally defined as lncRNAs (47, 114). A complete lncRNA map in highly tractable invertebrate models provides the means to rapidly elucidate the diverse functional classes of these molecules. Expression analyses have demonstrated that lncRNAs are differentially expressed during differentiation and development (37) and between cell types and cellular systems (26, 36, 93). Specific spatiotemporal patterns of lncRNA expression indicate that they may fine-tune early cell fate choices (20, 59). Some lncRNAs exhibit sequence complementarity and, by virtue of their ability to base pair with other RNAs, act as highly specific sensors of mRNA, miRNA, and other lncRNA expression. Upon target recognition, they can influence mRNA stability (79) and regulate translation of complementary mRNAs (e.g., a repressive effect for the targets of lncRNA-p21) (136).

Young et al. (137) analyzed modENCODE data on *Drosophila* to generate an initial map of lncRNAs expressed during development. We worked in parallel to build gene models that included, in addition to the developmental time course, our body map of fly tissues and the environmental perturbations. In addition to using the standard definition of a noncoding RNA (no open reading frame longer than 100 amino acids), we required that the RNA include no open reading frames with known conserved protein domains and no conserved open reading frames longer than 20 amino acids. We discovered 1,875 *Drosophila* lncRNAs, most with highly localized expression patterns. The majority of lncRNAs are expressed at one or a few developmental time points, and many are expressed in only a few tissues. More than 30% of lncRNAs have peak expression in the testes, and nearly 7% are expressed exclusively in the adult male gonad. Remarkably, every environmental perturbation experiment revealed lncRNAs as among the genome-wide most significant responders to stress, with some being induced or repressed hundreds of fold. What these transcripts do remains to be determined. None of the stress-responsive lncRNAs have a previously reported phenotype, but interestingly, 36 newly discovered lncRNAs do overlap molecularly defined mutations with associated phenotypes (17). However, most of these are likely regulatory alleles of nearby protein-coding genes. Is the RNA only an incidental by-product of a highly active enhancer? Would an RNAi knockdown of these transcripts reproduce the phenotype? Foundational *cis/trans* assays and some well-designed CRISPR/RNAi experiments are needed to assess these possibilities and uncover the biological functions of the many lncRNAs in *Drosophila*.

One particularly interesting class of lncRNAs is those that are antisense to protein-coding genes. We identified 402 antisense lncRNAs in *Drosophila*. The majority of these have peak expression in the male gonad. To our surprise, we found statistically significant enrichment for testes-specific antisense lncRNAs in humans (using the unpublished Illumina Body Map 2.0 data set). Genes with antisense transcripts in the fly and human male gonads include *monocarboxylate transporter 1* [*Mct1*, human ortholog *solute carrier family 16*, *monocarboxylic acid transporter 12* (*SLC16A12*)] (**Figure 3**), *even skipped* (*eve*, human ortholog *EVX1*), and *deoxyuridine triphosphatase* (*dUTPase*, human ortholog *DUT*). The term positional equivalence has been used to describe lncRNAs that overlap orthologous protein-coding genes but lack conserved sequence outside the antisense region (45). To the best of our knowledge, this was the first observation of positional equivalence between species as distantly related as flies and humans. Seven genes in *Drosophila* are reciprocally transcribed, like *dUTPase* and *modENCODE gene 9994* (*Mgn9994*), where the sense and antisense transcripts share boundaries (5′-3′ and 3′-5′). Generally, both strands are transcribed at approximately equal levels. In worms, a natural explanation for such loci would be RNA-dependent reverse transcription, but no such enzymes exist endogenously in flies. These also do not appear to be a library preparation artifact, because they were captured in several libraries and in both biological replicates within each library. Expression in tissue samples does not ensure expression in the same cells, nor does expression in the same cell ensure that transcripts are present together at the same time. The majority of these antisense loci do not give rise to significant levels of endo-siRNAs, and hence these sense and antisense transcripts may not co-occur. Additional single-cell and time-resolved studies will be needed to elucidate the roles of these transcripts in the male gonad. Unfortunately, we were not able to conduct these studies in worms during the consortium project period owing to a lack of stranded RNA data—we could not clearly distinguish which protein-coding genes were transcribed on both strands. The field of lncRNA biology is rapidly growing, but the coding capacity of lncRNAs remains in question. Ribosome profiling and subsequent tandem mass spectrometry proteomics studies are needed to distinguish noncoding RNAs from those that encode short peptides.

DNA Replication and Transcriptional Coupling

In fly cell lines, origins of DNA replication were mapped genome-wide using ChIP-seq (42, 89). In yeast, they are marked by a simple sequence motif, but in flies no such motif has been identified in vivo or in vitro. However, the application of a support vector machine to *k*-mer counts in intervals around OR complexes (ORCs) consistently utilized in three cell lines (two of embryonic origin and one from the central nervous system) revealed that sequence content

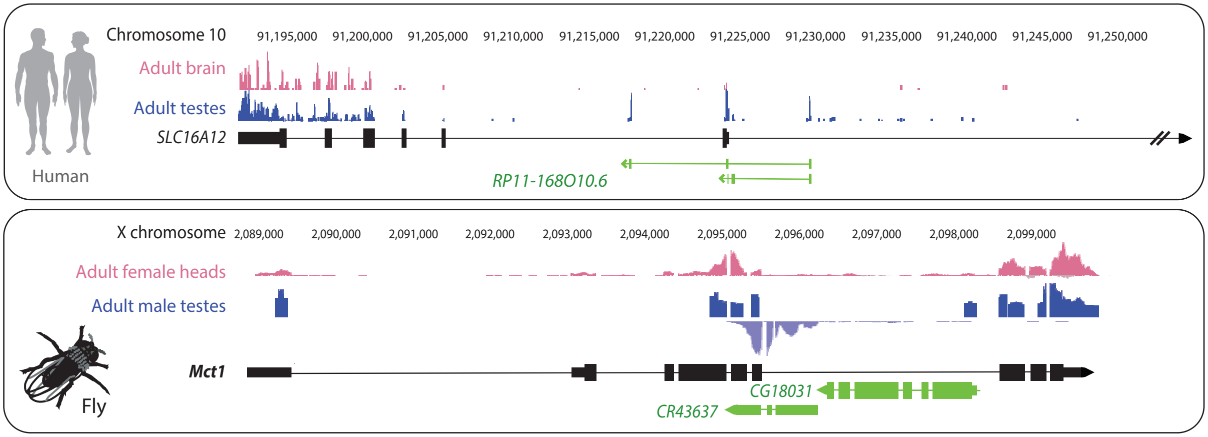

Figure 3

Positional equivalence antisense transcription in human and fly testes. The top panel shows human RNA-seq data from Illumina Body Map 2.0 (**http://www.illumina.com/informatics/research/sequencing-data-analysis-management/sequencing-data-library.html**) along with transcript models from GENCODE version 19 (**http://www.gencodegenes.org**). Transcripts on the minus strand are in green, and arrows indicate the direction of transcription. Illumina Body Map 2.0 RNA-seq data was unstranded, and here we impute the strand from two lines of evidence: First, reads that cross canonical splice junctions are assigned strandedness based on acceptor donor sequences, and second, reads that unambiguously overlap a stranded transcript model are assigned the strand of the putative parent transcript. The bottom panel shows modENCODE RNA-seq data from the *Drosophila* tissue atlas. These data are stranded; reads on the minus strand are shown inverted and in a lighter color. In humans, the lncRNA RP11-168010.6 is antisense to *SLC16A12*, an ortholog of the *Drosophila* gene *Mct1*. In *Drosophila*, the lncRNA *CR43637* is antisense to *Mct1*. Abbreviations: lncRNA, long noncoding RNA; *Mct1*, *monocarboxylate transporter 1*; RNA-seq, RNA sequencing; *SLC16A12*, *solute carrier family 16*, *monocarboxylic acid transporter 12*.

is sufficient to accomplish robust distinction of ORC binding sites from background genomic sequence. There is a sequence component to the establishment of ORCs in metazoans, but it may be a complex function of the binding motifs of many factors and cofactors. Examination of the genomic context of ORC binding sites and early-firing ORs revealed that many early-replicating sites occur in the promoter-proximal regions of highly transcribed genes (42). The chromatin context of these ORs tells a similar story: They are enriched for activating chromatin marks and regions of accessible chromatin. Heterochromatic regions tend to be replicated later. Eaton et al. (42) also found that ChIP data on the ORC were quantitative: Regions with more, higher-occupancy ORC binding sites tended to be replicated earlier than regions with fewer sites. Support vector machines trained on primary DNA sequence (local *k*-mers), TF binding sites, and chromatin marks provided exceptionally accurate and quantitative predictions of replication timing. Intriguingly, many predictors from each class appeared to be necessary to achieve high accuracy, suggesting that a variety of chromatin marks and DNA-binding factors and cofactors (known and unknown) interact in the determination of ORs and timing of DNA replication.

For the majority of the genome, replication timing is consistent across the modENCODE cell lines: Early- and late-replicated regions tend to be early and late (respectively) in both embryo- and nervous-system-derived cells. The small subset of dynamic ORs that are not consistently either late or early are marked by low ORC binding density in all cell types surveyed (88). This suggests that regions with static replication timing are determined by the density of ORC binding, which is likely driven by chromatin accessibility. The timing of dynamic regions may be determined by the regulation of origin activation rather than origin selection.

METHODOLOGICAL ADVANCES, BOTTLENECKS, AND ONGOING CHALLENGES

Integrating RNA Sequence Data Types

The first RNA-seq data sets produced were poly(A)$^+$ RNA-seq. Using only these data, we struggled to model transcripts and quantify their expression. At the time, the numerical optimization procedure in Cufflinks (128) was unstable, so two runs of the algorithm produced different results and many genes with zero expression values despite thousands of mapped reads. Running Trinity or Cufflinks to build gene models resulted mostly in fragments and false gene merges—something that was obvious because we could compare imputed models to the *Drosophila* cDNA collection (**http://fruitfly.org**), which includes full-length cDNAs for the majority of fly genes. We obtained CAGE data, which demarcated the 5′ boundaries of transcripts. We used 3′ expressed sequence tags to obtain polyadenylation site information [and later poly(A) site sequencing (PAS-seq)], and bringing these three data types together simplified the annotation of inferred transcript structures: RNA-seq data, particularly intron-spanning (splice-junction) reads, were used to connect 5′-3′ boundaries. The consortium generated a new integration package for the analysis of multiple RNA sequencing data types—GRIT (13)—that returns transcript models that begin with a cap and end with a poly(A) tail. This was an important finding of the consortium: Without 5′ and 3′ transcript end information, the connectivity of transcript models is uncertain. Even in the extensively curated human genome, hundreds, possibly thousands, of annotated lncRNAs were recently revealed as fragments of 3′ UTRs from upstream protein-coding genes (94). Using transcript boundary information removed most fragmented and merged models. The resulting annotation was not perfect—it still required some manual curation to correct missed poly(A) sites, poorly mapped junctions, and other errors—but it was a dramatic improvement over the products of other tools.

However, a serious challenge remains: GRIT requires splice junctions, which are derived from reads spanning spliced introns and crossing exon-exon boundaries. Many algorithms have been developed to perform the gapped alignments required to identify and map these reads, including TopHat (127), STAR (Spliced Transcripts Alignment to a Reference) (38), and MAQ (Mapping and Assembly with Quality) (84). Although these algorithms are remarkable in many ways, they all generate false positives and false negatives. We used auxiliary statistics based on the density of 5′ read mapping locations relative to the splice junction to filter false positives. In modENCODE, two tools were constructed to compute a variety of quality control metrics for mapped splice junctions (16, 123) and conduct differential expression analysis at the level of splicing events. Both tools are useful for the analysis of splice-junction reliability and the quantification of splicing events, but additional work remains: False negatives have not yet been sufficiently addressed, and more rigorous statistical and machine learning methods may improve overall accuracy.

The Design and Analysis of ChIP-Seq Experiments and the Mapping of Chromatin Structure

The modENCODE Consortium constructed experimental designs and methods for the analysis of the ChIP-seq assay (71, 75, 81, 107). In particular, the SPP peak caller was built for the identification of narrow peaks (75), and sequencing depth standards were developed based on statistical analyses of ultradeep sequencing experiments (71). Several novel assays were developed to interrogate chromatin structure, map the epigenome, sequence nascent RNA transcripts, and target assays to particular cell populations isolated from animals or complex tissues. These have recently been reviewed in detail (138), so here we add only that the modENCODE Consortium, particularly the Henikoff laboratory, opened the door to single-base-resolution mapping of the epigenome (62), and many of these techniques have now been applied to studies in humans, especially for the elucidation of noncoding variation in human populations (131). Because these assays were developed in highly tractable model systems, validation by genetic methods was readily accessible and data interpretation was rapidly refined.

Some Additional Challenges

Each group within the consortium confronted and overcame a variety of challenges. All groups transitioned from microarrays to sequencing starting in the second year of the project. The computational groups developed methods for integrating and jointly analyzing data types on different scales and resolutions. For the chromatin and regulation groups, antibody quality was a serious challenge. Many of the commercial antibodies raised to specific histone marks cross-reacted with one another and did not have the expected specificity (43). For TF ChIP, high-quality antibody production became rate limiting, and a tagged-TF strategy was adopted. Extensive work was needed to develop generalizable quality control and analysis procedures (24, 81). In flies, several ChIP-seq libraries were constructed using a telomerase-based approach to simultaneously fragment and amplify DNA, enabling much lower concentrations of starting chromatin and therefore fewer cells and improved automation. Although this pipeline works extremely well for interrogating chromatin structure (18), in ChIP-seq it frequently failed to produce high-quality data sets (14). It is important to separate research and development from large-scale data production and to maintain rapid communication between data producers and data analysts.

For the transcription groups, the lead technologies changed twice: first from arrays to sequencing, and then from unstranded (57) to stranded RNA-seq. We were an alpha-test group for the first Illumina stranded RNA-seq kit, which was used to produce nearly half of the fly data (17). Strand information allowed us to map patterns of antisense transcription, disambiguate nested genes from retained introns, and more reliably quantify gene expression. Although changing platforms introduced some computational challenges, the additional data were revolutionary for our understanding of the structure of the fly transcriptome.

LOOKING FORWARD: EMERGING APPLICATIONS AND NEW MODEL SYSTEMS

If a modENCODE-like project were initiated today in, for instance, *Daphnia*, zebrafish, *Xenopus*, or ecologically important nonmodel organisms (124), an ideal data set would include (*a*) detailed RNA profiling data in multiple size fractions and from 5′ and 3′ transcript boundaries, (*b*) chromatin accessibility and modification data to demarcate active and poised enhancers and repressed regions that are not revealed by transcriptional profiling, (*c*) DNA and RNA methylation data, and (*d*) binding data for both RNA- and DNA-binding factors. New assays, like the assay for transposase-accessible chromatin sequencing (ATAC-seq) (18), can be used to profile chromatin structure at an exceptionally low cost, requiring only three hours of lab time and far fewer cells than previous technologies. In small genomes, self-transcribing active regulatory region sequencing (STARR-seq) (3) can be used to comprehensively profile enhancer activity in selected cell types. BruChase-seq can be used to profile dynamic rates of transcription and transcript degradation and complements the steady-state measurements of standard RNA-seq (108). In modENCODE, we focused on the regulation of transcriptional initiation; now we would also include assays targeted to co- or posttranscriptional processing and translation.

Immunoprecipitation-based strategies are available to profile RNA-protein (29, 126), RNA-RNA (44), and RNA-DNA interactions (34). Even the proteome is increasingly accessible, with ribosome profiling strategies and falling mass spectrometry costs. Assays of three-dimensional chromatin structure and large-scale genome organization are also becoming more practical (32), and new consortium-scale projects will push the development of these technologies (e.g., 4D Nucleome; **http://commonfund.nih.gov/4Dnucleome/index**).

Today, sequencing is 50-fold cheaper than it was at the beginning of the modENCODE Project. For nonmodel organisms, it is possible to de novo sequence and assemble a metazoan genome in a matter of weeks and for less than $15,000 (104). Resequencing the *Drosophila* genome yielded a nearly error-free, 27-Mb, completely assembled chromosome arm (3L). New technologies such as CRISPR may soon make it possible to study gene function in systems not amenable to genetic analysis. Continuous technological innovations have made a broader spectrum of data types accessible to projects smaller in scale than modENCODE.

Understanding where and when genes are expressed is a prerequisite to organism-level systems biology. Projects in new model organisms should strive to supplement sequence-level profiling with imaging. With high-throughput microscope automation on the horizon, it may soon be feasible to compile complete spatial atlases on timescales similar to those of sequencing technologies (50).

In the future, genome sequencing will likely become a standard part of health care. Some have predicted that we will soon observe every possible base substitution at every location in the (reference) human genome. If epistasis is the rule for gene function, rather the exception, then conclusions based on alleles identified in one genetic background may translate poorly to others. Generalization of genetic components of disease and disease susceptibility will require a deep understanding of underlying mechanisms. Early strides toward high-throughput functional elucidation have been made by screening rare variants from individuals with unsolved Mendelian disease in model organisms (134). Additionally, it may soon be possible to build realistic in vitro models of human organs. However, effects on organismal biology, development, and ecology will, by definition, continue to require nonhuman model systems.

The broad survey conducted by the modENCODE Consortium has opened new areas of inquiry and has provided a foundation for future in-depth functional studies that will elucidate the genetic and epigenetic bases of development, disease, and resilience. The bright future of translational medicine and bioscience is, as it always has been, illuminated by worms, flies, mice, and many other model organisms.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We apologize to the investigators within the modENCODE Consortium and the larger model organism communities whose research could not be appropriately cited owing to space limitations. We thank the members of the modENCODE Consortium for many thoughtful discussions. We thank Ken Wan for help with **Figure 2**. We thank National Human Genome Research Institute (NHGRI) program officers Dr. Elise Feingold, Dr. Peter Good, and Dr. Mike Pazin for the coordination of the modENCODE Consortium. In addition, we thank Dr. Pazin for the curation of the modENCODE publications and the >160 community publications using modENCODE data. S.E.C. was funded by a contract from the NHGRI modENCODE Project, contracts U01 HG004271 (principal investigator) and R01 GM076655, both under Department of Energy contract DE-AC02-05CH11231. J.B.B. was supported by NHGRI contract R00 HG006698.

LITERATURE CITED

1. Allen MA, Hillier LW, Waterston RH, Blumenthal T. 2011. A global analysis of *C. elegans trans*-splicing. *Genome Res.* 21:255–64
2. Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, et al. 2014. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature* 512:400–5
3. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339:1074–77
4. Babiarz JE, Hsu R, Melton C, Thomas M, Ullian EM, Blelloch R. 2011. A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in *Dgcr8* versus *Dicer1* knockouts and small RNA sequencing. *RNA* 17:1489–501
5. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479:534–37
6. Bank EM, Gruenbaum Y. 2011. The nuclear lamina and heterochromatin: a complex relationship. *Biochem. Soc. Trans.* 39:1705–9
7. Berezikov E, Liu N, Flynt AS, Hodges E, Rooks M, et al. 2010. Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*. *Nat. Genet.* 42:6–9; author reply, 9–10
8. Berezikov E, Robine N, Samsonova A, Westholm JO, Naqvi A, et al. 2011. Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence. *Genome Res.* 21:203–15
9. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
10. Biggin MD. 2011. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* 21:611–26
11. Billi AC, Freeberg MA, Day AM, Chun SY, Khivansara V, Kim JK. 2013. A conserved upstream motif orchestrates autonomous, germline-enriched expression of *Caenorhabditis elegans* piRNAs. *PLOS Genet.* 9:e1003392
12. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816
13. Boley N, Stoiber MH, Booth BW, Wan K, Hoskins RA, et al. 2014. Genome guided transcript construction from integrative analysis of RNA sequence data. *Nat. Biotechnol.* 32:341–46
14. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* 512:453–56
15. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. 2013. Dynamic integration of splicing within gene regulatory pathways. *Cell* 152:1252–69
16. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, et al. 2010. Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res.* 21:193–202
17. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512:393–99
18. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10:1213–18
19. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, et al. 2009. Unlocking the secrets of the genome. *Nature* 459:927–30
20. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, et al. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147:358–69

21. Chen J, Miller A, Kirchmaier AL, Irudayaraj JM. 2012. Single-molecule tools elucidate H2A.Z nucleosome composition. *J. Cell Sci.* 125:2954–64
22. Chen RA, Down TA, Stempor P, Chen QB, Egelhofer TA, et al. 2013. The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures. *Genome Res.* 23:1339–37
23. Chen RA, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J. 2014. Extreme HOT regions are CpG-dense promoters in *C. elegans* and humans. *Genome Res.* 24:1138–46
24. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, et al. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods* 9:609–14
25. Chen Z-X, Sturgill D, Qu C, Jiang H, Park S, et al. 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res.* 24:1209–23
26. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308:1149–54
27. Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, et al. 2011. The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Res.* 21:301–14
28. Curtis HJ, Sibley CR, Wood MJ. 2012. Mirtrons, an emerging class of atypical miRNA. *Wiley Interdiscip. Rev. RNA* 3:617–32
29. Darnell RB. 2011. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA* 1:266–86
30. de Almeida SF, Grosso AR, Koch F, Fenouil R, Carvalho S, et al. 2011. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nat. Struct. Mol. Biol.* 18:977–83
31. de Wit E, Braunschweig U, Greil F, Bussemaker HJ, van Steensel B. 2008. Global chromatin domain organization of the *Drosophila* genome. *PLOS Genet.* 4:e1000045
32. de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26:11–24
33. Deal RB, Henikoff JG, Henikoff S. 2010. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* 328:1161–64
34. Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, et al. 2013. DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* 503:371–76
35. Ding Q, MacAlpine DM. 2010. Preferential re-replication of *Drosophila* heterochromatin in the absence of geminin. *PLOS Genet.* 6:e1001112
36. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, et al. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18:1433–45
37. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101–8
38. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21
39. Dodd IB, Micheelsen MA, Sneppen K, Thon G. 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* 129:813–22
40. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, et al. 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13:R53
41. Dorn R, Reuter G, Loewendorf A. 2001. Transgene analysis proves mRNA trans-splicing at the complex *mod*(*mdg4*) locus in *Drosophila. PNAS* 98:9724–29
42. Eaton ML, Prinz JA, MacAlpine HK, Tretyakov G, Kharchenko PV, MacAlpine DM. 2011. Chromatin signatures of the *Drosophila* replication program. *Genome Res.* 21:164–74
43. Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, et al. 2011. An assessment of histone-modification antibody quality. *Nat. Struct. Mol. Biol.* 18:91–93
44. Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, et al. 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* 159:188–99
45. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, et al. 2006. Complex loci in human and mouse genomes. *PLOS Genet.* 2:e47
46. Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* 28:817–25

47. Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15:7–21

48. Findlay GD, MacCoss MJ, Swanson WJ. 2009. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Res.* 19:886–96

49. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, et al. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *PNAS* 109:21330–35

50. Frise E, Hammonds AS, Celniker SE. 2010. Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Mol. Syst. Biol.* 6:345

51. Gapp K, Jawaid A, Sarkies P, Bohacek J, Pelczar P, et al. 2014. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat. Neurosci.* 17:667–69

52. Gent JI, Schvarzstein M, Villeneuve AM, Gu SG, Jantsch V, et al. 2009. A *Caenorhabditis elegans* RNA-directed RNA polymerase in sperm development and endogenous RNA interference. *Genetics* 183:1297–314

53. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330:1775–87

54. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* 512:445–48

55. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* 320:1077–81

56. Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. *Nature* 461:206–11

57. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473–79

58. Gu SG, Pak J, Guang S, Maniar JM, Kennedy S, Fire A. 2012. Amplification of siRNA in *Caenorhabditis elegans* generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint. *Nat. Genet.* 44:157–64

59. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477:295–300

60. Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, et al. 2009. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *Caenorhabditis elegans*. *PNAS* 106:18674–79

61. Heger P, Marin B, Schierenberg E. 2009. Loss of the insulator protein CTCF during nematode evolution. *BMC Mol. Biol.* 10:84

62. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. *PNAS* 108:18318–23

63. Hilgers V, Lemke SB, Levine M. 2012. ELAV mediates 3′ UTR extension in the *Drosophila* nervous system. *Genes Dev.* 26:2259–64

64. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. 2009. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res.* 19:657–66

65. Ho JW, Jung YL, Liu T, Alver BH, Lee S, et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature* 512:449–52

66. Horiuchi T, Giniger E, Aigaki T. 2003. Alternative *trans*-splicing of constant and variable exons of a *Drosophila* axon guidance gene, *lola*. *Genes Dev.* 17:2496–501

67. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* 21:182–92

68. Huang S, Litt M, Felsenfeld G. 2005. Methylation of histone H4 by arginine methyltransferase PRMT1 is essential in vivo for many subsequent histone modifications. *Genes Dev.* 19:1885–93

69. Ikegami K, Egelhofer TA, Strome S, Lieb JD. 2010. *Caenorhabditis elegans* chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. *Genome Biol.* 11:R120

70. Jacquier A. 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 10:833–44

71. Jung YL, Luquette LJ, Ho JW, Ferrari F, Tolstorukov M, et al. 2014. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.* 42:e74

72. Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, et al. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468:811–14
73. Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. 2014. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* 11:203–9
74. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471:480–85
75. Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26:1351–59
76. Khare SP, Habib F, Sharma R, Gadewal N, Gupta S, Galande S. 2011. HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* 40:D337–42
77. King FJ, Szakmary A, Cox DN, Lin H. 2001. *Yb* modulates the divisions of both germline and somatic stem cells through *piwi*- and *hh*-mediated mechanisms in the *Drosophila* ovary. *Mol. Cell* 7:497–508
78. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* 41:376–81
79. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, et al. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493:231–35
80. Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC. 2012. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res.* 22:1634–45
81. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22:1813–31
82. Lau NC, Robine N, Martin R, Chung WJ, Niki Y, et al. 2009. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res.* 19:1776–85
83. Lee H, McManus CJ, Cho D-Y, Eaton ML, Renda F, et al. 2014. DNA copy number evolution in *Drosophila* cell lines. *Genome Biol.* 15:R70
84. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851–58
85. Li JJ, Huang H, Bickel PJ, Brenner SE. 2014. Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.* 24:1086–101
86. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93
87. Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, et al. 2011. Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.* 21:227–36
88. Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. 2014. DNA replication and transcription programs respond to the same chromatin cues. *Genome Res.* 24:1102–14
89. MacAlpine HK, Gordan R, Powell SK, Hartemink AJ, MacAlpine DM. 2010. *Drosophila* ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Res.* 20:201–11
90. MacArthur S, Li XY, Li J, Brown JB, Chu HC, et al. 2009. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* 10:R80
91. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, et al. 2010. The landscape of *C. elegans* 3′UTRs. *Science* 329:432–35
92. McManus CJ, Duff MO, Eipper-Mains J, Graveley BR. 2010. Global analysis of trans-splicing in *Drosophila*. *PNAS* 107:12975–79
93. Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, et al. 2010. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci.* 11:14
94. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.* 23:812–25
95. Miura SK, Martins A, Zhang KX, Graveley BR, Zipursky SL. 2013. Probabilistic splicing of *Dscam1* establishes identity at the level of single neurons. *Cell* 155:1166–77
96. ModENCODE Consort, Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330:1787–97
97. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, et al. 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *PNAS* 103:12027–32

98. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, et al. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* 471:527–31

99. Ni X, Zhang YE, Negre N, Chen S, Long M, White KP. 2012. Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLOS Biol.* 10:e1001420

100. Nimura K, Ura K, Shiratori H, Ikawa M, Okabe M, et al. 2009. A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf-Hirschhorn syndrome. *Nature* 460:287–91

101. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, et al. 2011. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. *Genome Res.* 21:245–54

102. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* 130:89–100

103. Okamura K, Robine N, Liu Y, Liu Q, Lai EC. 2011. R2D2 organizes small regulatory RNA pathways in *Drosophila*. *Mol. Cell. Biol.* 31:884–96

104. Pac. Biosci. 2014. Data release: preliminary de novo haploid and diploid assemblies of Drosophila melanogaster. *PacBio Blog*, Jan. 13. **http://blog.pacificbiosciences.com/2014/01/data-release-preliminary-de-novo.html**

105. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40:1413–15

106. Park D, Lee Y, Bhupindersingh G, Iyer VR. 2013. Widespread misinterpretable ChIP-seq bias in yeast. *PLOS ONE* 8:e83506

107. Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10:669–80

108. Paulsen MT, Veloso A, Prasad J, Bedi K, Ljungman EA, et al. 2013. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* 67:45–54

109. Rando OJ. 2012. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.* 22:148–55

110. Rechtsteiner A, Ercan S, Takasaki T, Phippen TM, Egelhofer TA, et al. 2010. The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny. *PLOS Genet.* 6:e1001091

111. Richardson SR, Morell S, Faulkner GJ. 2014. L1 retrotransposons and somatic mosaicism in the brain. *Annu. Rev. Genet.* 48:1–27

112. Riddle NC, Jung YL, Gu T, Alekseyenko AA, Asker D, et al. 2012. Enrichment of HP1a on *Drosophila* chromosome 4 genes creates an alternate chromatin structure critical for regulation in this heterochromatic domain. *PLOS Genet.* 8:e1002954

113. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, et al. 2011. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.* 21:147–63

114. Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81:145–66

115. Ruby JG, Jan CH, Bartel DP. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* 448:83–86

116. Ruby JG, Jan CH, Player C, Axtell MJ, Lee W, et al. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127:1193–207

117. Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, et al. 2009. A regulatory circuit for *piwi* by the large Maf gene *traffic jam* in *Drosophila*. *Nature* 461:1296–99

118. Saloura V, Cho H-S, Kyiotani K, Alachkar H, Zuo Z, et al. 2015. WHSC1 promotes oncogenesis through regulation of NIMA-related kinase-7 in squamous cell carcinoma of the head and neck. *Mol. Cancer Res.* 13:293–304

119. Schwartz YB, Linder-Basso D, Kharchenko PV, Tolstorukov MY, Kim M, et al. 2012. Nature and function of insulator protein binding sites in the *Drosophila* genome. *Genome Res.* 22:2188–98

120. Shindo Y, Nozaki T, Saito R, Tomita M. 2013. Computational analysis of associations between alternative splicing and histone modifications. *FEBS Lett.* 587:516–21

121. Smibert P, Miura P, Westholm JO, Shenker S, May G, et al. 2012. Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. *Cell Rep.* 1:277–89

122. Stoeckius M, Maaskola J, Colombo T, Rahn HP, Friedlander MR, et al. 2009. Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. *Nat. Methods* 6:745–51

123. Sturgill D, Malone JH, Sun X, Smith HE, Rabinow L, et al. 2013. Design of RNA splicing analysis null models for post hoc filtering of *Drosophila* head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinform.* 14:320

124. Tagu D, Colbourne JK, Negre N. 2014. Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC Genomics* 15:490

125. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *PNAS* 110:18602–7

126. Tome JM, Ozer A, Pagano JM, Gheba D, Schroth GP, Lis JT. 2014. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods* 11:683–88

127. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:562–78

128. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511–15

129. Van Nostrand EL, Kim SK. 2013. Integrative analysis of *C. elegans* modENCODE ChIP-seq data sets to infer gene regulatory interactions. *Genome Res.* 23:941–53

130. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–76

131. Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* 30:1095–106

132. Wen J, Mohammed J, Bortolamiol-Becet D, Tsai H, Robine N, et al. 2014. Diversity of miRNAs, siRNAs, and piRNAs across 25 *Drosophila* cell lines. *Genome Res.* 24:1236–50

133. Westholm JO, Ladewig E, Okamura K, Robine N, Lai EC. 2012. Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. *RNA* 18:177–92

134. Yamamoto S, Jaiswal M, Charng WL, Gambin T, Karaca E, et al. 2014. A *Drosophila* genetic resource of mutants to study mechanisms underlying human genetic diseases. *Cell* 159:200–14

135. Yang Q, Hua J, Wang L, Xu B, Zhang H, et al. 2013. MicroRNA and piRNA profiles in normal human testis detected by next generation sequencing. *PLOS ONE* 8:e66809

136. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, et al. 2012. LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* 47:648–55

137. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, et al. 2012. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol. Evol.* 4:427–42

138. Zentner GE, Henikoff S. 2014. High-resolution digital profiling of the epigenome. *Nat. Rev. Genet.* 15:814–27

139. Zhao Q, Rank G, Tan YT, Li H, Moritz RL, et al. 2009. PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat. Struct. Mol. Biol.* 16:304–11

140. Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, et al. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLOS Genet.* 6:e1000848

141. Zhu S, Wang G, Liu B, Wang Y. 2013. Modeling exon expression using histone modifications. *PLOS ONE* 8:e67448

Annual Review of
Genomics and
Human Genetics

Volume 16, 2015

Contents

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://www.annualreviews.org/errata/genom