Identification of functional elements and regulatory circuits by Drosophila modENCODE


* = equal contributors, co-first authors. + = equal contributors, co-second authors, listed alphabetically. ** = equal contributors, listed alphabetically.

One-sentence summary

Integrative analysis of genome-wide functional datasets in Drosophila melanogaster reveals protein coding, non-coding, regulatory, and chromatin elements, and leads to regulatory networks predictive of gene function and activity.

Abstract

To gain insights into how genomic information is translated into cellular and developmental programs, the Drosophila modENCODE project is comprehensively mapping transcripts, histone modifications, chromosomal proteins, transcription factors, replication proteins and cell cycle timing, and nucleosome properties across a developmental time-course and in multiple cell lines. We have generated more than 700 datasets and discovered protein coding, non-coding, RNA regulatory, replication, and chromatin elements, tripling the annotated portion of the Drosophila genome. Correlated activity patterns of these elements reveal a functional regulatory network, which predicts putative new functions for genes, reveals stage-specific and tissue-specific regulators, and enables gene expression prediction. Our results provide a foundation for directed experimental and computational studies in Drosophila and related species, and a model for systematic data integration towards comprehensive genomic and functional annotation.
Introduction

Several years after the complete sequencing of many species it is still unclear how to translate genomic information into a functional map of cellular and developmental programs. The ENCyclopedia of DNA Elements (ENCODE) (1) and Model Organism ENCODE (modENCODE) (2) projects use diverse genomic assays to comprehensively annotate the *Homo sapiens* (human), *Drosophila melanogaster* (fly), and *Caenorhabditis elegans* (worm) genomes, through systematic generation and computational integration of functional genomic datasets.

Previous genomic studies in flies have made seminal contributions to our understanding of basic biological mechanisms and genome functions, facilitated by genetic, experimental, computational, and manual annotation of the euchromatic and heterochromatic genome (3), small genome size, short life cycle and a deep knowledge of development, gene function, and chromosome biology. The functions of approximately 40% of the protein and non-protein coding genes (FlyBase 5.12, ref site) have been determined from cDNA collections (4, 5), manual curation of gene models (6), gene mutations and comprehensive genome-wide RNAi screens (7-9), and comparative genomic analyses (10, 11).

The Drosophila modENCODE project has generated more than 700 datasets that profile transcripts, histone modifications and physical nucleosome properties, general and specific transcription factors (TFs), and replication programs in cell lines, isolated tissues, and whole organisms across several developmental stages (Fig. 1). Here, we computationally integrate these datasets and report:

* Improved and additional genome annotations, including full-length protein-coding genes and peptides as short as 21 amino acids.
* Non-coding transcripts, including 132 candidate structural RNAs and 1608 non-structural transcripts.
* Additional Ago-associated small RNA genes and pathways, including new miRNAs encoded within protein-coding exons and endogenous small interfering RNAs from 3’ untranslated regions.
* Chromatin ‘states’ defined by combinatorial patterns of 18 chromatin marks, which are associated with distinct functions and properties.
* Regions of high TF occupancy and replication activity with likely epigenetic regulation.
* Mixed TF and miRNA regulatory networks with hierarchical structure and enriched feed-forward loops.
Co-expression and co-regulation based functional annotations for nearly 3,000 genes.

Predictive models of gene expression levels.

Overview of datasets

Our datasets provide an extensive description of the transcriptional, epigenetic, replication and regulatory landscapes of the Drosophila genome (table S1). Experimental assays include: RNA-seq, capturing-small and large RNAs and splice variants; ChIP-chip and ChIP-seq, profiling chromosomal and RNA binding or processing proteins; tiling-arrays, identifying and measuring replication patterns, nucleosome solubility and turnover; and genomic DNA sequencing, measuring copy number variation. Most assays were conducted in the sequenced strain y; cn bw sp (12) with multiple developmental samples (30 for RNA expression and twelve for TF and histone studies), and in cultured cells, predominantly with four lines (S2, BG3, Kc, and Cl.8; table S2).

Annotation of gene transcripts and their promoter regions

To comprehensively characterize transcribed sequences, we performed RNA-seq using poly(A)+ and total RNA, CAGE, RACE and produced ESTs (table S1) (13-15) and cDNAs. These data support over 90% of annotated genes, exons, and splice junctions, and provide experimental evidence for a total of 17,000 protein coding and non-coding genes, of which 1938 are previously unannotated. In addition to genes, we discovered 52,914 new or modified exons (65% supported by cDNAs), 22,965 new splice junctions in 14,016 distinct alternative transcripts (35% supported by cDNAs, RT-PCR products and long poly(A)+ RNAseq (13)). Overall, 74% of annotated genes show at least one new or modified exon or alternative splice form despite extensive previous annotation efforts, illustrating the importance of probing additional cell types. Of the 21,071 newly predicted exons expressed in S2 cells, 89% are associated with chromatin signatures characteristic of transcribed regions (16).

Our Transcription Start Site (TSS) annotation characterized the shapes and TSS distributions for 56% of annotated genes (70% of embryonically expressed genes). We also discovered and validated 2075 alternative promoters for known genes. Of 427 discovered alternative promoters adjacent to active S2 cell transcripts, 72.5% are supported by promoter-associated chromatin marks in that cell type (17), confirming predictions, and suggesting that these regions contain regulatory elements. Similarly, comparison to chromatin marks in whole animals yielded 1,117 additional validated promoters (18).
We detect all but 1,498 (9.9%) of previously annotated *D. melanogaster* genes (FB5.12) in either the poly(A)+ or total RNA-seq samples. Undetected genes include members of multi-copy gene families (e.g. ribosomal RNAs, paralogs, snoRNAs, tRNAs) and those with known low or constrained expression. We discovered new snoRNAs, scaRNAs and pri-miRNA transcripts in the total embryonic RNA-seq data, even without including larval, pupal or adult samples.

**Protein-coding, structural and non-coding transcripts**

We searched for evolutionary signatures of conserved protein-coding DNA sequences in alignments of 12 Drosophila genomes (11, 19) and for similarity to known proteins. Only 57 of 1,938 new gene models (16) contain a complete, conserved open reading frame likely to represent unidentified protein-coding genes (Fig. 2A). An additional 81 are likely to be incompletely reconstructed coding genes, because they contain at least one protein-coding exon, but lack clearly identifiable translation start or stop sites (16). These 138 genes show nearly 6-fold lower average expression than known protein-coding genes (FPKM of 6.7 vs. 34.8), and 40% have expression restricted to late larvae, pupae and adult males, providing a potential explanation for why they were missed in previous annotations. In addition to lacking protein-coding selection, the remaining 1800 gene models also lack similarity to known proteins (16).

We looked for properties of non-coding RNAs (ncRNAs) among the 1740 transcripts (excluding 60 snoRNA and miRNA transcripts) detected by RNA sequencing that do not appear to encode proteins. We examined folding thermodynamics and comparative evidence of local secondary structures in the predicted ncRNAs and in 140 ncRNAs listed in FB5.12 that do not belong to major classes of structural RNAs, such as miRNAs and snoRNAs. RNAz (20) predicted high-confidence structures for 132 transcripts (7.6%) suggesting conserved function as structural RNAs, similar to the fraction (7.8%) observed for structured transcripts in FB5.12 ncRNAs. We revealed candidate structural RNAs in the newly predicted transcripts (Fig. 2B), as well as novel structural elements in well studied ncRNAs, including roX2 and HSRw (fig. S1)(16). However, the lack of highly structured regions in the vast majority of ncRNAs suggest functions independent of secondary structure.

**Argonaute-associated small regulatory RNAs**

Our analysis of deeply sequenced ~18-28 nt RNAs dramatically extended the catalog of Argonaute (Ago)-dependent small regulatory RNAs(21), including microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-associated RNAs (piRNAs). In the canonical miRNA
pathway, ~21-24 nt RNAs are cleaved from hairpin precursors by Drosha and Dicer-1 RNase III enzymes and loaded into AGO1 effector complexes to repress mRNA targets. We annotated 61 additional canonical miRNAs, 12 of which derived from the antisense strands of known miRNA loci (22), which may provide an efficient route for the evolution of new miRNA activities. We unexpectedly detected miRNAs that overlap mRNAs, including 9 cases where conserved protein-coding regions harbor RNA hairpins cleaved into miRNA/miRNA* duplexes, and detected in AGO1 complexes (e.g. Fig. 2C). It remains to be seen whether these mRNA-resident miRNAs have detectable trans-regulatory activities, affect their host transcripts in cis, or are simply neutral substrates. We identified 15 additional mirtrons that generate miRNAs by splicing of short hairpin introns (23). We defined up to 7 hybrid mirtrons bearing 3' tails, which appear to require processing by the exosome prior to dicing (24). In total, we recognize at least three miRNA biogenesis strategies, producing miRNAs from at least 240 genomic loci.

We and others recognized several classes of endogenous siRNAs (endo-siRNAs), 21-nt RNAs that are processed by Dicer-2 RNase III enzyme and preferentially loaded into AGO2 (25-30). Endo-siRNAs derive from three distinct sources: from diverse transposable elements (TEs), whose activity they restrict; from seven genomic regions encoding long inverted repeat transcripts, which direct the cleavage of specific mRNA targets; and from bi-directionally transcribed regions. This last class mostly comprises convergent transcripts that overlap in their 3' UTRs, termed 3' cis-natural antisense transcripts (3' cis-NATs). Our current analysis doubled the number of 3' cis-NAT-siRNA regions to 237, comprising nearly a quarter of overlapping 3' UTRs (Table S4).

Lastly, piRNAs are ~24-30 nt RNAs bound by the largely gonadal Piwi-class Argonautes, Piwi, Aubergine (Aub) and AGO3. The majority of piRNAs match TEs in sense or antisense orientation and are essential to repress their activity (31). While many Drosophila piRNAs map uniquely to tens of "master loci" that serve as genetic repositories for TE defense (31), we found that the 3' UTRs of hundreds of cellular transcripts also generate abundant Piwi-loaded primary piRNAs in somatic ovarian follicle cells (32-34). This suggests that beyond transposon control, the piRNA pathway may play more general roles in cellular gene regulation.

Large-scale organization of the chromatin landscape

Eukaryotic genomes are organized into large domains (~10kb-Mbs) that exhibit distinct chromatin properties, such as heterochromatic regions which cover one third of the genome and are typically known for transcriptional silencing (35). Our analyses show that the chromatin composition, organization, and boundaries of heterochromatin display surprising complexity and
plasticity among cell types (36). We find surprisingly active heterochromatic regions, with expression of 45% of pericentric heterochromatin genes (compared to 50% for euchromatic genes), and enrichment for both ‘active’ and ‘silent’ marks in active heterochromatin genes. Conversely, we find that domains enriched for heterochromatic marks (e.g. H3K9me2) cover a surprisingly large proportion of euchromatic sequences (12% in BG3 cells and 6% in S2) (36).

We identified large domains with similar replication patterns by characterizing the Drosophila DNA replication program in cell lines, and observed that the temporal replication program is determined by local chromatin environment {Kharchenko, submitted #33};Eaton, submitted #197} and the density of replication initiation factors (37). We also found that specific euchromatic regions up to 300kb were under-replicated in a tissue-specific manner in the polytene salivary glands, larval midgut and fatbodies (38) suggesting copy number variation may help regulate gene expression levels.

**Chromatin signatures characteristic of functional elements**

Many genomic regulatory regions are difficult to identify due to a lack of characteristic sequence signatures, but are often marked by specific histone modifications, variants, and other epigenetic factors (39, 40). To identify such signatures, we assayed 18 histone modifications and variants by ChIP-chip in multiple cell lines (17) and developmental stages (18), and defined the physical properties of nucleosomes (41, 42). Correlating this information with the transcriptome data sets, binding site profiles for replication factors, insulator-binding proteins, and TFs identified chromatin signatures related to active genes (Fig. 3A). TSS-proximal regions were marked by H3K4me3 enrichment (43), depletion of nucleosome density, increased nucleosome turnover, and enrichment in the pellet chromatin fraction (41, 42). Gene bodies showed a 3’-biased enrichment of H3K36me3 and H3K79me1 marks, and H2B ubiquitination covering the entire transcribed region. Moreover, large introns are enriched for H3K36me1, H3K18ac and H3K27ac, specific chromatin remodelers, high nucleosome turnover, the H3.3 histone variant, and DNase I hypersensitive sites, all suggestive of regulatory functions (17). These features are generally absent from short genes, and from genes with a low fraction of intronic sequence. Most transcriptionally-silent genes lack pronounced chromatin signatures, except when positioned within Pc domains (H3K27me3), or heterochromatin (H3K9me2/3, HP1a, H3K23ac depletion) (36).

Positional correlation analysis identified relationships between histone marks and nucleosome physical properties. Active marks (e.g. H3K27Ac, RNA Pol II, H3K4me3) correlate with high chromatin solubility and high nucleosome turnover rates, while marks associated with silent
Chromatin (e.g. H3K27me3, H1, H3K9me2/3) show the opposite, correlating with increased nucleosome density (fig. S2). High chromatin solubility indicates less stable nucleosomes (42), and high levels of nucleosome turnover are indicative of a dynamic chromatin structure (41), consistent with the biological functions associated with the corresponding marks.

We mapped origins of replication activated in early S phase of the cell cycle, and binding sites of the origin recognition complex (ORC), a conserved replication initiation factor that exhibits little, if any, sequence specificity in vitro (44, 45). ORC-associated sequences are often found at TSSs and depleted for bulk nucleosomes, but are enriched for the variant histone H3.3 (37) and undergo active nucleosome turnover (CATCH-IT) (41). These findings suggest that local nucleosome occupancy and organization are determinants of ORC binding in Drosophila, as in yeast (46, 47). By subdividing the ORC sites into TSS-proximal and TSS-distal sites, we found that local enrichment for GAGA factor (GAF), and H4Ac tetra, H3K27Ac, H4K8Ac and H3K18Ac are common to both, whereas H3K36me1 appears to be specific for TSS-distal ORC sites (Fig. 3A). ORC marks sites of cohesin complex loading in Drosophila (37); H3K36me1, which is also enriched at cohesin sites (17), may be required in the absence of TSS-associated marks to promote ORC binding and subsequent cohesin loading (48, 49).

Insulator elements and proteins (e.g. CP190, CTCF, SUHW and BEAF) block enhancer-promoter interactions and restrict the spread of histone modifications (50). Analysis of the genomic distributions of insulator proteins showed that BEAF32, CP190, and ZW5 preferentially bind upstream of TSSs, while SUHW binds almost exclusively distal to TSSs, with CTCF binding both equally (51). Insulator regions displayed distinct chromatin signatures (Fig. 3A), but most of the variation is explained by the differences between TSS-proximal and distal chromatin contexts, suggesting that specific marks are not required for insulator binding or function. However, nucleosome depletion is a common feature of both TSS-proximal and distal insulator binding sites, as in mammals (52), a property that may facilitate insulator binding or reflect the ability of insulator proteins to displace nucleosomes.

Chromatin-based annotation of functional elements

Chromatin signatures associated with TSSs and transcribed regions (43) identified genes and promoters missed by transcript-based annotation. We developed a predictive model for active promoters in cell lines using positional enrichments of 18 histone modifications, ORC complex localization, and nucleosome stability and turnover in the 1kb regions surrounding validated active promoters. Our logistic regression classifier achieved 93.7% sensitivity at a 21.5% False Discovery Rate (FDR)(fig. S4), and predicted 2,203 additional promoter positions at least 500bp
away from annotated TSSs (16). These included promoters for 10 primary miRNA transcripts of which 7 were also identified by RNAseq (13). We also used H3K36me3/H2B-ubiquitination signatures (fig. S3) to identify 53 transcribed gene bodies outside annotated genes, of which 11 are additionally supported by promoter predictions (e.g. Fig. 3B). These included 4 primary miRNA transcripts of which 3 are also supported by RNAseq (13) and 1 is also supported by our promoter predictions (for mir-317).

Chromatin signatures also identify functional elements involved in other chromosomal processes such as duplication and segregation. We identified 133 sites in BG3 and 78 sites in S2 cells that contained large (>10kbp) intergenic domains of H3K36me1. In BG3 cells, 90% and 68% of the intergenic H3K36me1 domains overlapped with cohesin (17, Misulovin, 2008 #93) and early origin activity, respectively, as observed for a 20kb region upstream of the bi gene (Fig. 3C, fig. S5). Although only 15% of early replication origins appear to be defined by intergenic H3K36me1 domains, the overlap with cohesion enrichment (17) suggests a shared mechanism to ensure faithful chromosome inheritance.

**De novo discovery of combinatorial chromatin states**

Multiple histone modifications act in concert to determine genome functions producing combinatorial chromatin ‘states’ (53). We used two unsupervised, multivariate Hidden Markov Models (HMMs) to segment the genome based on the combinatorial patterns of 18 histone modifications in S2 and BG3 cells (Fig. 4, fig. S6) (17). Not seeking a ‘true’ number of distinct chromatin states, we instead sought to identify models that balance resolution and interpretability given the available chromatin marks, as more states led to increased enrichment for specific genomic features but captured progressively smaller fractions of each type of feature (fig. S7).

From these considerations, we focused on a nine state intensity-based model reflecting broad classes of chromatin function (continuous model states c1 to c9), and a 30 state model that identifies combinatorial patterns at a finer resolution (discrete model states d1 to d30) (16). (Fig. 4, left panel). These showed distinct functional and genomic enrichments (Fig. 4, right panel) associated with different chromosomes (4, male X), regulatory elements (promoters, enhancers), gene length and exonic structure (e.g. long first introns), gene function (e.g. developmental regulators), gene expression levels (high or medium, low, or silent).

Intergenic regions and silent genes are associated with state d30 (c9) in euchromatin (covering 51% of the genome and lacking enrichments for any of the marks examined), and with states d26, d28-d29 (c7-c8) in heterochromatin (characterized by H3K9me2/3 enrichment and
H3K23ac depletion). These states lack enrichments for other mapped factors (e.g. insulators, HDACs, TFs), and exhibit low levels of chromatin solubility and nucleosome turnover.

In contrast, expressed genes display numerous and complex enrichments for several factors and chromatin properties. Most active TSSs were associated with state c1, defined by known promoter-associated marks H3K4me3 and H3K9ac (43). Other active TSSs were instead associated with H3K4me1, H3K36me1 and multiple acetylations (d13). Even within c1, some TSSs showed higher association with nucleosome turnover, group 1 insulator proteins and HDACs (d1, d3), while others were associated with heterochromatic genes of medium (d5) or low expression (d6).

The state analysis also captured the associated between ORC binding and TSSs for both euchromatin and heterochromatin, as well as the correlation between early origins and open chromatin in euchromatic regions. However, ORC binding is largely limited to a subset of TSS-associated states (d1, d5, d6, d13, d17 and not d3, d24), and some states enriched for ORC binding are not found at TSSs (d11, d14, d21). Early origins are primarily associated with states c3 (active intron, enhancer) and c4 (open chromatin), and often display distinct state enrichments from ORC binding in accord with the broad domains they cover, compared to the near nucleotide resolution of the ORC binding data.

Our states showed some similarities with the recently published 5 ‘colors’ of chromatin from DAM-ID mapped chromosomal proteins in Kc cells (54), but even highly-specific states were sometimes split across multiple colors (fig. S8). This suggests a more complex picture with many highly-specific chromatin states with specific functional enrichments.

**Chromatin and motif properties of high-occupancy TF binding sites**

Extensive overlap in the binding profiles of multiple TFs has revealed High-Occupancy Target (HOT) regions or hotspots (18, 55-59). Using the binding profiles of 41 TFs in early embryo development we assigned a ‘TF complexity’ score to each of 38,562 distinct TF binding sites (16), corresponding to the number of distinct TFs bound (from 1 to ~21), resulting in 2,006 hotspots with TF complexity 8 or greater, corresponding to approximately 10 overlapping factors bound. We correlated these regions with our and other datasets, to gain insights into the possible mechanisms of HOT region establishment, and how they may impact or be affected by chromatin properties.

We studied the enrichment of regulatory motifs for 32 TFs for which we have both genome-wide bound regions and well-established regulatory motifs (Fig. 5A). We sorted each TF based on its average complexity (the average number of TFs that co-bind (18)), which ranges from 10.8 for
KNI to 1.3 for FTZ-F1. We studied the relative enrichment of each factor’s known motif in bound regions and found eight factors (KNI, DLL, GT, PRD, KR, SNA, DA, TWI) with average complexity greater than 4 that showed significant differences in motif enrichment at varying complexity levels. In all eight cases, motif matches were preferentially found in regions of lower complexity suggestive of non-specific binding. For an additional nine TFs, bound regions were enriched in the known motif, but no bias for lower-complexity regions was found, and for another ten factors the known motif did not show a significant enrichment in bound regions, suggesting that either the motif is incorrect, or a larger fraction of TFs than previously expected binds in non-sequence-specific ways.

We found a strong correlation between HOT spots of increasing TF complexity and decreased nucleosome density (fig. S9A), increased nucleosome turnover (fig. S9B), and histone variant H3.3 which is associated with nucleosome displacement (fig. S9C), but a surprising depletion in previously-annotated enhancers (18) suggesting potentially distinct roles for these elements. Enrichment for HOT regions was observed across a wide range of complexity values for several chromatin states associated with TSS and open chromatin regions (d1, d5, d6, d13, d14, d21); whereas some states (d3, d24) were enriched only at lower complexity (fig. S9D). In contrast, transcriptional elongation (d7-d9), intergenic (d30), and heterochromatic states (d26, d27, d29) were strongly depleted across all complexity ranges. We also found concordance between HOT regions and ORC binding sites (Fig. 5B), with the likelihood of ORC binding increasing monotonically with the complexity of the TF bound regions. Coupled with the lack of a detectable specific sequence for ORC binding in Drosophila (37), this suggests hotspots as an alternative mechanism for ORC regulation via non-specific binding in high-accessibility regions, and widespread interplay between chromatin regulation, TF binding, and DNA replication. Given the high agreement between embryo and cell line datasets, we propose hotpots are stable genomic regions, kept open via recruitment of specific chromatin marks or remodelers, and facilitating binding of additional TFs at their motifs or non-specifically.

We looked for potential ‘driver’ motifs that may be recognized by TFs potentially involved in establishing HOT regions (Fig. 5C). Applying our motif discovery pipelines (18) within bound regions of varying complexity resulted in seven distinct motifs associated with hotspots of different complexities. Motifs M2 and M3 were similar to the BEAF-32 and Trl/GAF insulator motifs, suggesting interplay between hotspots and insulator proteins. Motif M1 differed in only one position from the known Sna motif, and was strongly enriched for high-complexity regions (Fig. 5C), while the Sna motif was depleted in Sna-bound regions of higher complexity (Fig. 5A) suggesting that the single-nucleotide difference may be important for recognition. The other four
motifs did not match any known TFs, suggesting that yet-uncharacterized potential sequence-specific regulators may be involved in the establishment of hotspots.

**Fraction of the genome assigned to candidate functions**

We assigned candidate functions to the fraction of the non-repetitive genome covered by the datasets, excluding large blocks of repeats and low-complexity sequences (Fig. 6A). Protein-coding exons cover 21% of the genome, and adding Argonaute-associated small regulatory RNAs, UTRs, other ncRNAs, bases covered by PolII, the binding sites of TFs and other chromatin interacting factors brings the total genome coverage to 73%. Inclusion of Pc and ORC binding sites, and derived chromatin states, brings the total genome coverage to 81.5%, and the addition of transcribed intronic positions raises the total coverage to over 89% (Fig. 6A). Compared to previous annotations (FB5.12), we have increased coverage of the *Drosophila* genome with putative associated functions by 26.3% (47 Mb). Euchromatic regions had much higher coverage than heterochromatic regions (90.6% vs 69.5%), when comparing the respective non-repetitive portions.

We next determined the overlap between our predicted functional elements and PhastCons evolutionarily conserved elements across 12 Drosophila species, mosquito, honeybee, and beetle (60). These elements cover 38% of the *D. melanogaster* genome in 1.2 million blocks, over which we repeated our previous individual and cumulative calculations. 32% of constrained bases are covered by protein-coding exons alone, increasing to a cumulative total of 80% for transcribed and regulatory elements, and 91.8% after inclusion of specific chromatin states (Fig. 6A). Nearly all modENCODE-defined functional elements were more likely to cover constrained bases than is expected by chance (fig. S10), with the exception of some less active chromatin states, and, as expected, introns, UTRs, and non-coding RNAs (61), provide additional independent evidence for the predicted elements.

Overlap among the annotations produced by different types of elements resulted in dense multiple coverage (Fig. 6B), even for regions that previously lacked any annotation (Fig. 6C). Even though the genome coverage average is 2.8 datasets, 10.8% of the genome is covered by 15 or more datasets, and coverage peaks at 103 datasets overlapping a single region on chromosome 3R. We found strong positive correlations between bound regulators and transcribed element densities, and regulators and chromatin element densities (fig. S11). In the case of chromatin datasets, additional chromatin marks resulted in higher accuracy in chromatin state recovery (fig. S12), and we expect similar impact of additional datasets on other classes of functional elements.
TF targets and physical regulatory network inference

We examined the network of regulatory relationships between TFs, miRNAs and their target genes. ‘Nodes’ in these networks represent the transcriptional and post-transcriptional regulators and target genes, and ‘edges’ or ‘connections’ represent their directed regulatory relationships. We inferred a physical regulatory network of TF binding and miRNA targeting, where connections represent physical contact between regulators and genomic regions of their target genes.

The structural properties of the physical regulatory network was inferred from the experimentally-derived binding profiles of 76 TFs (table S5), and genome-wide occurrences of 77 distinct evolutionarily-conserved miRNA seed motifs for 105 miRNAs (16). The structure of the resulting network shows high connectivity and rapid spread of regulatory information, requiring traversal of only ~2 regulatory connections on average between any two genes, and no more than 5 connections between any pair of genes. Target genes are regulated by ~12 TFs on average and can have up to 54 regulatory TFs (16). The most heavily targeted genes are associated with increased pleiotropy, as measured by the number of distinct functional processes and tissues with which they are associated (16).

The physical regulatory network includes both pre- and post-transcriptional regulators, identifying the interplay between these two types of regulation. We organized the TFs of the physical regulatory network into five levels (Figs. 7A, fig. S13) on the basis of the relative proportion of TF targets versus TF regulators for each TF (62), and augmented this network with the miRNA regulators most closely interacting with each level. The presumed ‘master regulator’ TFs at the top level targeted almost all the other TFs in the network, while only 8% of lower level edges pointed ‘upward’ to higher levels, supporting a hierarchical nature and suggesting little direct feedback control of master regulators among the TFs surveyed. We also observed that even though the number of TF targets decreases for TFs at lower levels of the hierarchy, the number of their miRNA targets increases (0.58 miRNA targets per TF for two topmost levels versus 1.55 for two lowest levels, fold enrichment of 2.66). This suggests that at least some feedback from lower levels to master regulators may occur indirectly through miRNA regulators.

We next searched for significantly overrepresented network connectivity patterns, or ‘network motifs’ (Fig. 7B) likely to represent building blocks of gene regulation (63). We found eight network motifs in the physical regulatory network (64), of which five correspond to TF co-operation (motifs 1, 2, 4, 7, 8), confirming observations of co-binding and co-targeting (55-59).
In all five motifs, at least two TFs bind each other’s promoter regions, suggesting extensive positive and negative feedback. Two other motifs correspond to mixed feed-forward loops involving cooperation of TFs and miRNAs (motifs 3, 6), which can lead to different delay properties in the expression of target genes depending on the activating or repressive action of the TF. Lastly, one motif (motif 5) corresponds to a feedback-loop of a downstream TF targeting an upstream TF through a miRNA, as also observed as a means for feedback in the hierarchical network layout (16).

Dataset integration predicts a functional regulatory network

We integrated the physical network with patterns of coordinated activity of regulators and targets to derive a functional regulatory network (fig. S14A). Although TF binding is strongly associated with the true regulatory targets, binding alone can occur without a sequence-specific TF-motif interaction, and does not always result in changes in gene expression (58). Thus, a functional regulatory network should consider both binding and its functional consequences, such as changes in expression or chromatin, which are correlated with gene function (fig. S15). Neither network is a strict subset of the other, as some physical connections may not lead to functional changes and functional connections may be indirect or simply missing in the physical regulatory map.

We integrated multiple types of evidence including-conserved sequence motifs of 104 TFs in promoter regions across the genome (table S5), ChIP-based TF binding for 76 factors, and the correlation between chromatin marks and gene expression patterns of regulators and their target genes (fig. S16). We combined these lines of evidence with unsupervised machine learning to infer the confidence of each regulatory edge between 707 proteins classified as TFs (16) and 14,444 targets for which at least one line of evidence was available (16).

We compared the resulting functional network to the physical network inferred from TF binding, a predicted physical network constructed from motif occurrences, and the REDfly literature-curated functional network (16). The functional network included a similar number of target genes as both the binding and motif physical networks (~10,000 targets each), but more regulators overall (576 vs. 104 and 76, respectively), and more regulators per target (24 vs. 7 and 13, respectively) (fig. S14B). The functional network showed similarity to both the motif and binding networks, which were both used as input evidence; connections of the functional network showed a greater than 4-fold enrichment compared to both networks, even though the two only showed a 1.6-fold enrichment to each other’s connections (fig. S14C). Compared to
either the motif or the binding network, the functional network showed the strongest connectivity similarity to the REDfly network, even though it was not specifically trained to match known edges.

The functional regulatory network showed increased biological significance compared to both the motif and binding networks, including increased functional similarity of co-targeted genes, expression correlation of co-targeted genes, and protein-protein interactions (fig. S14D) (16). The REDfly network slightly outperformed the functional network, confirming the relevance of the metrics. However, the functional network contains 100-fold more targets (9,436 vs. 88) and 1000-fold more connections (230k vs. 233), suggesting it will be more valuable for predicting gene function and gene expression at the genome scale.

**Predicting gene function from the functional regulatory network**

We provided candidate functional annotations for genes that lack gene ontology (GO) terms on the basis that targets of similar regulators and with similar expression are likely to share similar functions. We probabilistically assigned genes to 34 expression clusters (fig. S15 (16)), and predicted likely functional GO terms for every gene using a guilt-by-association approach that using GO terms of annotated genes to predict likely functions of unannotated genes, allowing for multiple annotation predictions for each gene (16). This resulted in a higher predictive power than using expression or regulators alone (fig. 8A). At FDR<0.25, we predicted GO terms for 1286 previously-unannotated genes and additional terms for 1586 previously-annotated genes (fig S17, Table S6, Data S15). Generally, tissue-specific enrichments of new GO predictions matched those of known genes in the same GO terms (fig. S18), providing an independent validation of our approach.

**Predicting stage-specific regulators of gene expression**

We predicted stage-specific regulators of gene expression on the basis of transcriptional changes during development. With the Dynamic Regulatory Events Miner (DREM) (65), we searched for splits (a point at which previously co-expressed genes begin to exhibit divergence into two or three distinct expression patterns) among a set of over 6000 genes with the largest expression changes occurring during the developmental time-course (Fig. 9A, S19). We mined the physical and functional regulatory networks to predict stage-specific regulators from the over-representation of regulator targets along specific trajectories or “paths” from each split (16)). Several predictions agreed with literature support. For example, Tin was associated with early increase in expression and enriched for early organ development genes (p<10^{-53}) (66) and
E2f2 was associated with early decrease in expression and enriched for cell cycle genes \((p<10^{-100})\) \((67)\).

To provide additional support for regulator predictions made using the physical network, we examined the time-course expression profiles of the regulators, which were not directly used in the prediction scheme. Even though several caveats could hinder this analysis, the time-course expression of the regulators was often consistent with DREM’s predictions. For example, a sharp decline in SU(HW) expression coincides with sharp expression increase of its targets (Fig. 9A), consistent with a repressive role \((68)\). We generally observed a significant correspondence among the stage-specific expression changes of predicted regulators at developmental stages that correspond with concomitant expression changes in their target genes. Regulators predicted to be associated with a split had on average a significantly greater absolute expression change than those not associated with a split \((P<10^{-10}, \text{fig. S19})\) \((16)\).

**Predicting cell type-specific regulators of chromatin activity**

We computed enrichments of conserved regulatory motif instances in cell-type specific annotations for 22 chromatin factors in both S2 and BG3 cells. We defined signatures of cell-type specific activators and repressors likely involved in establishing the chromatin differences between S2 and BG3 cells (Fig. 9B) by comparing these enrichments to the expression patterns of the TFs that recognize these motifs in the same cell types \((16)\). Activators were defined as TFs whose cell-type specific expression coincided with activation of their predicted targets, and repressors were defined as TFs whose cell-type specific expression was correlated with repression of their predicted targets. This resulted in 1-8 predicted regulators for each cell, including for example CREBA as a predicted S2 activator, H as a predicted BG3 repressor, and Homeobox-like as a predicted BG3 activator.

For most regulatory motifs, enrichment in activating chromatin marks was coupled with depletion in repressive chromatin marks. This coupling leads to more robust predictions of activators and repressors, and also enables a high-level distinction between active and repressive chromatin marks that agrees with previous studies and with our chromatin state analysis \((\text{(17, 18) Fig. 4})\). For a small number of motifs, however, the chromatin enrichments did not show a consistent picture of opposite enrichments in activating vs. repressive marks. These could be false positives and not actually associated with chromatin regulation, or they could be active in other cell types and not relevant to the distinction between S2 and BG3 chromatin marks.

**Predicting target gene expression from regulator expression**
Developmental regulatory programs are defined by multiple interacting regulators contributing to observed changes in gene or region activity (69). We sought to predict the specific expression levels of target genes across numerous stages and cell lines on the basis of the expression levels of its regulators. With the 30 distinct measurements of expression levels obtained by RNA-seq across development (13), we predicted the expression level of each target gene as a linear combination of its regulators, as defined by the functional regulatory network (Fig. 9C). We split the time-course into ten intervals of three samples each and learned stable coefficients for linear combination of TFs across nine intervals to predict expression in the tenth (16).

We predicted the expression levels of 1,991 genes better than random control networks (23.6% of genes), a 2.5-fold enrichment (control networks perform better on 9.5% of genes) (figs. S20, S21). In contrast, physical networks showed almost no predictive value over the randomized networks (table S7), suggesting that they are best used when combined with additional information for inferring the functional regulatory networks.

Genes whose expression levels are ‘predictable’ from the expression levels of their regulators (those with consistently lower errors than random) may be more precisely regulated, and thus associated with less noisy expression patterns. Indeed, the expression correlation between the 30-time-point data set used for expression prediction (13), and an independently-generated 12-time-point data set sampled at longer intervals (18), was significantly higher for predictable genes compared to unpredictable genes (KS test P-value < 1e-7, fig. S22). These results validate our methodology for gene expression prediction, and suggest that unpredictable genes may be due to intrinsic variability in gene expression levels.

We also tested if the regulatory models obtained with whole embryo time-course datasets can predict gene expression under novel conditions; specifically the Clone8, Kc, BG3, and S2 cell lines. For each ‘predictable’ gene the expression levels of their regulators were combined as dictated by the weights learned in the time-course experiment, and used to predict target gene expression. The expression of 932 predictable genes also showed better than random predictions (compared to 296 genes for the binding network and 214 genes for the motif network). Overall, 62% of embryo-defined predictable genes were also predictable in cell lines, compared to only 10-15% for embryo-based unpredictable genes, providing further validation of our methodology.

Our results suggest that the primary datasets are highly relevant for inferring functional regulatory relationships that are predictive of expression (figs 9C, S20, S23). However, genome-scale gene expression prediction remains an enormously difficult problem as only a quarter of
all genes was predictable, which we expect to improve with additional datasets generated from increasingly genome-scale projects.

Discussion

This first phase of the modENCODE project has provided the foundation for integrative studies of metazoan biology, enhancing existing genome annotations, broadening the number and diversity of small RNA genes and pathways, revealing chromatin domains and signatures, and elucidating the interplay between replication, chromatin and TF binding in high-occupancy regions of likely non-specific binding. Our resulting annotations together cover 82% of the genome, a nearly 4-fold increase compared to previously-annotated protein-coding exons, and have important implications for interpreting the molecular basis of genetically-linked phenotypes.

Our integrative analysis revealed connections between elements in physical and functional regulatory networks, enabling the prediction of gene function, tissue and stage-specific regulators, and gene expression levels. While our initial results are promising, only one quarter of all genes showed predictable expression, suggesting the need for continued mapping of regulatory interconnections and functional datasets, as well as new predictive models.

It remains to be seen how general regulatory principles elucidated here will be conserved across the animal kingdom and especially in the human, through comparison across the ENCODE and modENCODE projects. Towards this end, we are expanding our exploration of functional elements, cell types and developmental stages, and prioritizing orthologous assays and conditions across species. Given the extensive conservation of biological molecules and processes between flies and vertebrates (70), these will not only improve our understanding of fly biology, but can serve as a template for understanding of human biology and disease.

Supporting Online Material includes Supplementary Methods, Supporting online text, Supplementary Figs. S1 to S23, Supplementary Tables S1 to S7, and Supplementary Data S1 to S17. These can be downloaded from modencode.org and www.sciencemag.org

References cited
16. Supporting text and materials and methods are available as supporting material on Science Online.
17. P. V. Kharchenko et al., *Nature*, (Submitted).
Acknowledgements. Supported by the National Science Foundation under Grant # 0937060 to the Computing Research Association for the CIFellows Project (SR), the National Science Foundation under award no. 0905968 (JE), NSERC Fellowship (BA), Tamer Kahveci (FA), the Japan Society for the Promotion of Science (KO), the Swedish Research Council (QD), a NIH NRSA postdoctoral fellowship (CAB), a NDSEG fellowship (RS), an Erwin Schrödinger Fellowship of the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (SW), a Leukemia and lymphoma society fellowship (SW), a Lilly-LSRF fellowship (CDB), a Natural Sciences and Engineering Research Council of Canada Post-Doctoral Fellowship (CGA), Affymetrix for partial support (TGR), a fellowship from the Swiss National Science Foundation (DM), by NHGRI U01 HG004258 (OK), German Research Foundation grant WI 3628/1-1 (SW), an HHMI Damon Runyon Cancer Research fellowship (JTN), the Indiana Genomics Initiative (TCK), Harold Smith and the NIDDK genomics core laboratory (BO), contract U01 HG004271 from NHGRI under DOE contract DE-AC02-05CH11231 (SEC), U01-HG004261 (EL), 5U01HG004279 (DMM), U01 HG004258 (GHK), U01HG004264 (KPW), and NHGRI 5R01HG004037-03, RC2 HG005639, an NSF CAREER award, and the Sloan Foundation (MK).

Correspondence should be addressed to: manoli@mit.edu (Integrative analysis), celniker@fruitfly.org (Transcripts), karpen@fruitfly.org (Chromatin), kpwhite@uchicago.edu (TFs), david.macalpine@duke.edu (Replication), laie@mskcc.org (Small RNAs), steveh@fhcrc.org (Nucleosomes), lincoln.stein@gmail.com (Data availability).


**Author affiliations:**

1. Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02140, USA.
2. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
3. Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St. Boston, Massachusetts 02115, USA.
4. Institute for Genomics &amp; Systems Biology, Department of Human Genetics, The University of Chicago, 920 E. 58th Street, Chicago, Illinois 60637, USA.
5. Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA.
6. Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mailstop 64-121, Berkeley, California 94720 USA.
7. Department of Molecular Genetics, University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada.
8. Sloan-Kettering Institute, 1275 York Avenue, Box 252, New York, New York 10065, USA.
9. Department of Genome Dynamics, Lawrence Berkeley National Laboratory, and Department of Molecular and Cell Biology, University of California, Berkeley, One Cyclotron Road, Berkeley, California 94720, USA.
10. Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK.
11. Department of Medicine, Brigham & Women's Hospital, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.
12. Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda MD 20892, USA.

13. Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK.

14. Cold Spring Harbor Laboratory, Howard Hughes Medical Institute, Cold Spring Harbor, New York 11724, USA.

15. Department of Genetics and Developmental Biology, University of Connecticut Stem Cell Institute, 263 Farmington, Connecticut 06030-6403, USA.

16. Department of Biology CB-1137, Washington University, Saint Louis, Missouri 63130, USA.

17. Affymetrix Inc, Santa Clara, California 95051, USA.

18. Division of Extramural Research, National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Suite 4076, Bethesda, Maryland 20892-9305, USA.

19. Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto Ontario M5G 0A3, Canada.

20. Department of Genetics and Drosophila RNAi Screening Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.

21. Center for Genome Sciences, Washington University, 4444 Forest Park Boulevard, Saint Louis, Missouri 63108, USA.

22. Department of Biology, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA.

23. Department of Biology, Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

24. Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences & University Medical Center Utrecht, Utrecht, The Netherlands.

25. Center for Genomics and Bioinformatics, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA.

26. Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA.

27. Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA.

28. Division of Biological Sciences, Section of Cell & Developmental Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093 USA.

29. Genome Sciences Centre, BC Cancer Agency, Vancouver Canada.

30. Department of Genetics, University of Cambridge, CB2 3EH, UK; Cambridge Systems Biology Centre, Tennis Court Road, Cambridge CB2 1QR, UK.

31. Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854, USA.

32. Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.
33. Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA.
34. Machine Learning Group, Université Libre de Bruxelles, CP212, Brussels 1050, Belgium.
35. Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown, MA, 02129, USA.
36. Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA.
37. Department of Plant & Microbial Biology, University of California, Berkeley, California 94720, USA.
38. Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA 92039, USA.
39. Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA.
40. Center for Biomolecular Science and Engineering, School of Engineering and Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA.
41. Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY 11794, USA.

Figure legends:

Figure 1. Overview of Drosophila modENCODE datasets. Range of genomic elements and trans factors studied, with relevant techniques and resulting genome annotations.

Figure 2 - Coding and non-coding genes and structures. A. Extended region of male specific expression in chromosome 2R including new protein-coding and non-coding transcripts. MIP03715 contains two short open reading frames (ORFs) of 23 and 21 codons. ORF multispecies alignments (color coded) show abundant synonymous (bright green) and conservative (dark green) substitutions and a depletion of nonsynonymous substitutions (red), indicative of protein-coding selection (dN/dS < 1 for both, P < 10^{-7} and P<10^{-11} respectively, likelihood ratio test). Surrounding regions show abundant stop codons (blue, magenta, yellow) and frame-shifted positions (orange). B. New transcribed region in Chromosome 3R (26,572,290-26,573,456) identified by RNA-Seq and supported by promoter-specific and transcription-associated chromatin marks shows RNA secondary structure conservation in eight Drosophila species. C. Example of a new miRNA derived from a protein-coding exon of CG6700, with 21-23 nt RNAs indicative of Drosha/Dicer-1 processing and also recovered in AGO1-Immunoprecipitate libraries from S2 cells and adult heads indicative of Argonaute loading. Evolutionary evidence suggests protein-coding constraint, no conservation for the mature arm, and conservation of the star arm.
Figure 3: Chromatin-based annotation of functional elements A. Average enrichment profiles of histone marks, chromosomal proteins and physical chromatin properties at genes, origins of replications, insulator proteins and TF binding positions. Each panel shows 4kb centered at a specified location, either proximal to TSS (prox.) or distal (dist.). B. Example of new transcript predicted by chromatin signatures associated with promoter (red track) and gene bodies (blue box) and supported by cDNA evidence. Strong RNA PolII and H3K4me3 peaks in the promoter region and strong H2B-ubiquitination extending towards the previously annotated luna gene are confirmed by RNA-seq junction reads that were not used in the prediction. C. Intergenic H3K36me1 chromatin signatures predict replication activity. Enrichment of multiple chromatin marks were used to identify putative large (>10kbp) intergenic H3K36me1/H3K18ac domains located outside of annotated genes. While these marks generally correspond to long introns within transcripts, their intergenic domains were enriched for replication activity (fig. S5). In this example from BG3 cells, such a domain was found upstream of the bi locus and is associated with early replication, contains an early origin, is enriched for ORC binding, and is further supported by NippedB binding.

Figure 4. Discovery and characterization of chromatin states and their functional enrichments. Combinatorial patterns of chromatin marks in S2 and BG3 cells reveal chromatin states associated with different classes of functional elements. Discrete model (states d1-d30) captures presence/absence information and continuous model (states c1-c9) also incorporates mark intensity information (21). States were learned solely from mapped locations of marks (left panel) and associated with modENCODE-defined elements (right panel) with most pronounced patterns in euchromatin (green) and heterochromatin (blue) illustrated here and additional variations shown in fig. S6.

Figure 5. High-occupancy TF binding regions and their relationship to motifs, ORC, and chromatin. A. Enrichment of known motifs for regions bound by corresponding TF, sorted by average complexity, denoting the number of distinct TFs bound in the same region. For eight TFs, motifs are depleted (blue) for higher-complexity regions, suggesting non-sequence-specific recruitment. In seven of eight cases, known motifs were enriched in bound regions ('Enrich') suggesting sequence-specific recruitment in lower-complexity regions. For each factor, binding sites were highly reproducible between replicates ('Reprod'). B. ORC vs. TF Complexity. Relationship between HOT spot complexity (x-axis) and enrichment in ORC binding (y-axis). C. Discovered motifs. Discovered motifs in high or low complexity regions (boxed range) and their enrichment in regions of higher (red) or lower (blue) complexity. M1-M5 are candidate 'drivers' of HOT region establishment.
Figure 6. Genome coverage by modENCODE datasets. A. Unique (bars) and cumulative (lines) coverage of non-repetitive (blue line) and conserved genome (red line). B. Multiple coverage for datasets grouped into transcribed elements (red), bound regulators (blue) and chromatin domains (green) (16). Across all three classes (black), 10.8% of the genome is covered 15 or more times, and 69.5% is covered at least twice. C. Increased coverage in Chr2R region with no prior annotation (left half), now showing multiple overlapping datasets. Coverage by different tracks is highly clustered (fig. S11), with some regions showing little coverage and others densely covered by many types of data.

Figure 7. Properties of the physical regulatory network A. Hierarchical view of mixed ChIP-based/miRNA physical regulatory network which combines transcriptional regulation by 76 TFs (green) from ChIP experiments and post-transcriptional regulation by 52 miRNAs (red). TFs are organized in a five level hierarchy on the basis of their relative proportion of TF targets versus TF regulators. miRNAs are separated into two groups, the ones that are regulated by TFs (left) and the ones that only regulate TFs (right). The horizontal position of the TFs in each level shows whether they regulate miRNAs (left), have no regulation to or from miRNAs (middle), or do not regulate but are targeted by miRNAs (right). Different shades of green and red represent the total number of target genes for TFs and miRNAs respectively (darker node = more targets). 92% of TF regulatory connections are ‘downstream’ connections from higher levels to lower levels (green), and only 8% are upstream (blue). miRNA regulatory connections are red. B. Highly-enriched network motifs in mixed physical regulatory network including TFs (green), miRNAs (red) and target genes (black). For each motif, five examples are shown, with known activators colored blue, known repressors colored red, and other TFs in black.

Figure 8. Gene function prediction from co-expression and co-regulation patterns. Receiver operator characteristic (ROC) curves for gene ontology terms with predicted new members, and area under the curve (AUC) statistic for each. False negatives for each GO term are predictions for genes previously annotated for ‘incompatible’ GO terms, defined as pairs of GO terms that have less than 10% common genes relative to the union of their gene sets.

Figure 9. Predictive models of regulator, region, and gene activity. A. Dynamic regulatory map produced by DREM predicts stage-specific regulators associated with expression changes (y-axis, log space relative to first time point) across developmental stages (x-axis) (16). Each path (colored lines) indicates the average expression of a group of genes (solid circles) and its standard deviation (size of circle). Predicted bifurcation events, or splits (open circles) are numbered 1 through 19, with inset image showing expression level of each individual gene going through the split, and associated regulators indicated (‘H’: higher-expression path; ‘L’:
lower-expression path; ‘M’: middle path), from the physical network (black) or the functional regulatory network (blue), ranked. Uncolored inset shows the expression of repressor su(Hw), whose expression decrease coincides with expression increase of its targets (red star). B. Predicted S2 activators (top group) or repressors (bottom group), based on the coherence between relative expression of the TF in S2 vs. BG3 (yellow/green) and the relative motif enrichment (red) or depletion (blue) in S2 vs. BG3 for activating marks (left columns) or repressive marks (right columns). C. Predicting gene expression levels from the expression of inferred regulators. For each target gene, we predicted its expression level as a linear combination of the expression level of its regulators, as dictated by the functional regulatory network. Each 30-stage time-course dataset was split into 10 groups of 3 stages each, and weights were learned on 9 and evaluated on the 10th. The predicted expression is shown (dotted blue line) compared to the true expression (top of shaded area) for target genes. The expression of the top 5 positive and negative regulators is shown for each target gene, ranked by their contribution to the expression prediction (the weights of the linear regression model). Top row shows examples from the top 25 scoring reconstructions, second row shows examples from 30-100, third row from 100-500 and fourth row from 500-1488, ordered by their prediction quality scores (rank in upper right corner), evaluated as the averaged squared error between predicted and true expression levels across the time-course. Color indicates the sign of correlation of a regulator and a target, red for positive (+) correlation, green for negative (-) correlation. Color scale is proportional to strength of correlation or anti-correlation. An expanded set of examples is found in fig S23.
Figure 1 - Overview of modENCODE datasets
Figure 2 - Examples of new protein-coding and non-coding genes
Figure 3 - Chromatin-based annotation of functional elements
<table>
<thead>
<tr>
<th>Histone Marks</th>
<th>Chromatin States</th>
<th>State Annotation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K36me3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K79me1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K79me2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K4me2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K4me3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K9me3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K27me3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K9me2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K9ac</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4K16ac</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K18ac</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K27ac</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDACs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDACs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDACs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDACs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 - Chromatin states and their functional enrichments
Figure 5 - Hotspot enrichment for known motifs, ORC, and discovered motifs
Figure 6 - Genome Coverage by modENCODE datasets
Figure 7 - Properties of the mixed Transcription Factor/microRNA regulatory network
Figure 8 - Gene function prediction performance

- oxidative phosphorylation 0.91
- spindle elongation 0.86
- reg. of RNA metabolic process 0.82
- organ morphogenesis 0.81
- pattern specification process 0.81
- regulation of transcription 0.81
- imaginal disc development 0.80
- organ development 0.78
- nervous system development 0.78
- cell differentiation 0.78
- muscle structure development 0.73
**Figure 9 - Predictive models of regulator, gene and region activity**