Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE

The modENCODE Consortium,* Sushmita Roy,[1,2]† Jason Ernst,[1,2]† Peter V. Kharchenko,[3]† Pouya Kheradpour,[1,2]† Nicolas Negre,[4]† Matthew L. Eaton,[5]† Jane M. Landolin,[6]† Christopher A. Bristow,[1,2]† Lijia Ma,[4]† Michael F. Lin,[1,2]† Stefan Washietl,[1]† Bradley I. Arshinoff,[7,18]† Ferhat Ay,[1,33]† Patrick E. Meyer,[1,30]† Nicolas Robine,[8]† Nicole L. Washington,[9]† Luisa Di Stefano,[1,31]† Eugene Berezikov,[23]‡ Christopher D. Brown,[4]‡ Rogerio Candeias,[1]‡ Joseph W. Carlson,[6]‡ Adrian Carr,[10]‡ Irwin Jungreis,[1,2]‡ Daniel Marbach,[1,2]‡ Rachel Sealfon,[1,2]‡ Michael Y. Tolstorukov,[3]‡ Sebastian Will,[1]‡ Artyom A. Alekseyenko,[11] Carlo Artieri,[12] Benjamin W. Booth,[6] Angela N. Brooks,[28] Qi Dai,[8] Carrie A. Davis,[13] Michael O. Duff,[14] Xin Feng,[13,18,35] Andrey A. Gorchakov,[11] Tingting Gu,[15] Jorja G. Henikoff,[8] Philipp Kapranov,[16] Renhua Li,[17] Heather K. MacAlpine,[5] John Malone,[12] Aki Minoda,[6] Jared Nordman,[22] Katsutomo Okamura,[8] Marc Perry,[18] Sara K. Powell,[5] Nicole C. Riddle,[15] Akiko Sakai,[29] Anastasia Samsonova,[19] Jeremy E. Sandler,[6] Yuri B. Schwartz,[3] Noa Sher,[22] Rebecca Spokony,[4] David Sturgill,[12] Marijke van Baren,[20] Kenneth H. Wan,[6] Li Yang,[14] Charles Yu,[6] Elise Feingold,[17] Peter Good,[17] Mark Guyer,[17] Rebecca Lowdon,[17] Kami Ahmad,[29] Justen Andrews,[21] Bonnie Berger,[1,2] Steven E. Brenner,[28,32] Michael R. Brent,[20] Lucy Cherbas,[21,24] Sarah C. R. Elgin,[15] Thomas R. Gingeras,[13,16] Robert Grossman,[4] Roger A. Hoskins,[6] Thomas C. Kaufman,[21] William Kent,[34] Mitzi I. Kuroda,[11] Terry Orr-Weaver,[22] Norbert Perrimon,[19] Vincenzo Pirrotta,[27] James W. Posakony,[26] Bing Ren,[26] Steven Russell,[10] Peter J. Park,[3] Susan E. Celniker,[6]§‖ Steven Henikoff,[25]§‖ Gary H. Karpen,[6,28]§‖ Eric C. Lai,[8]§‖ David M. MacAlpine,[5]§‖ Lincoln D. Stein,[18]§‖ Kevin P. White,[4]§‖ Manolis Kellis[1,2]‖

To gain insight into how genomic information is translated into cellular and developmental programs, the *Drosophila* model organism Encyclopedia of DNA Elements (modENCODE) project is comprehensively mapping transcripts, histone modifications, chromosomal proteins, transcription factors, replication proteins and intermediates, and nucleosome properties across a developmental time course and in multiple cell lines. We have generated more than 700 data sets and discovered protein-coding, noncoding, RNA regulatory, replication, and chromatin elements, more than 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- 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 also a model for systematic data integration toward comprehensive genomic and functional annotation.

Several years after the complete genetic 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* (fruit fly), and *Caenorhabditis elegans* (worm) genomes, through systematic generation and computational integration of functional genomic data sets.

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 ~40% of the protein- and nonprotein-coding genes [FlyBase 5.12 (4)] have been determined from cDNA collections (5, 6), manual curation of gene models (7), gene mutations and comprehensive genome-wide RNA interference screens (8–10), and comparative genomic analyses (11, 12).

The *Drosophila* modENCODE project has generated more than 700 data sets 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 data sets and report (i) improved and additional genome annotations, including full-length protein-coding genes and peptides as short as 21 amino acids; (ii) noncoding transcripts, including 132 candidate structural RNAs and 1608 nonstructural transcripts; (iii) additional Argonaute (Ago)–associated small RNA genes and pathways, including new microRNAs (miRNAs) encoded within protein-coding exons and endogenous small interfering RNAs (siRNAs) from 3′ untranslated regions; (iv) chromatin "states" defined by combinatorial patterns of 18 chromatin marks that are associated with distinct functions and properties; (v) regions of high TF occupancy and replication activity with likely epigenetic regulation; (vi) mixed TF and miRNA regulatory networks with hierarchical structure and enriched feed-forward loops; (vii) coexpression- and co-regulation–based functional annotations for nearly 3000 genes; (viii) stage- and tissue-specific regulators; and (ix) predictive models of gene expression levels and regulator function.

Overview of data sets. Our data sets provide an extensive description of the transcriptional, epigenetic, replication, and regulatory landscapes of the *Drosophila* genome (table S1). Experimental assays include high-throughput RNA sequencing (RNA-seq), capturing-small and large RNAs and splice variants; chromatin immunoprecipitation (ChIP)–chip and ChIP followed by high-throughput sequencing (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. We conducted most assays in the sequenced strain *y; cn bw sp* (13), with multiple developmental samples (30 for RNA expres-

sion and 12 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, cap analysis of gene expression, rapid amplification of cDNA ends, and produced expressed sequence tags (table S1) (14–16) and cDNAs. These data support more than 90% of annotated genes, exons, and splice junctions and provide experimental evidence for a total of 17,000 protein-coding and noncoding genes, of which 1938 are previously unannotated. In addition to genes, we discovered 52,914 previously undescribed or modified exons (65% supported by cDNAs) and 22,965 new splice

junctions in 14,016 distinct alternative transcripts [35% supported by cDNAs, reverse transcription polymerase chain reaction products, and long poly(A)$^+$ RNA-seq (14)]. Overall, 74% of annotated genes show at least one previously undescribed 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 (17).

We also characterized the shapes and transcription start site (TSS) distributions for 56% of annotated genes (70% of embryonically expressed genes). We 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 (18), confirming predictions and suggesting that these regions contain regulatory elements. Similarly, comparison to chromatin marks in whole animals yielded 1117 additional validated promoters (19).

We detect all but 1498 (9.9%) of previously annotated *D. melanogaster* genes (4) in either the poly(A)$^+$ or total RNA-seq samples. Undetected genes include members of multicopy gene families [e.g., ribosomal RNAs, paralogs, small nucleolar RNAs (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 alone, even without including larval, pupal, or adult samples.
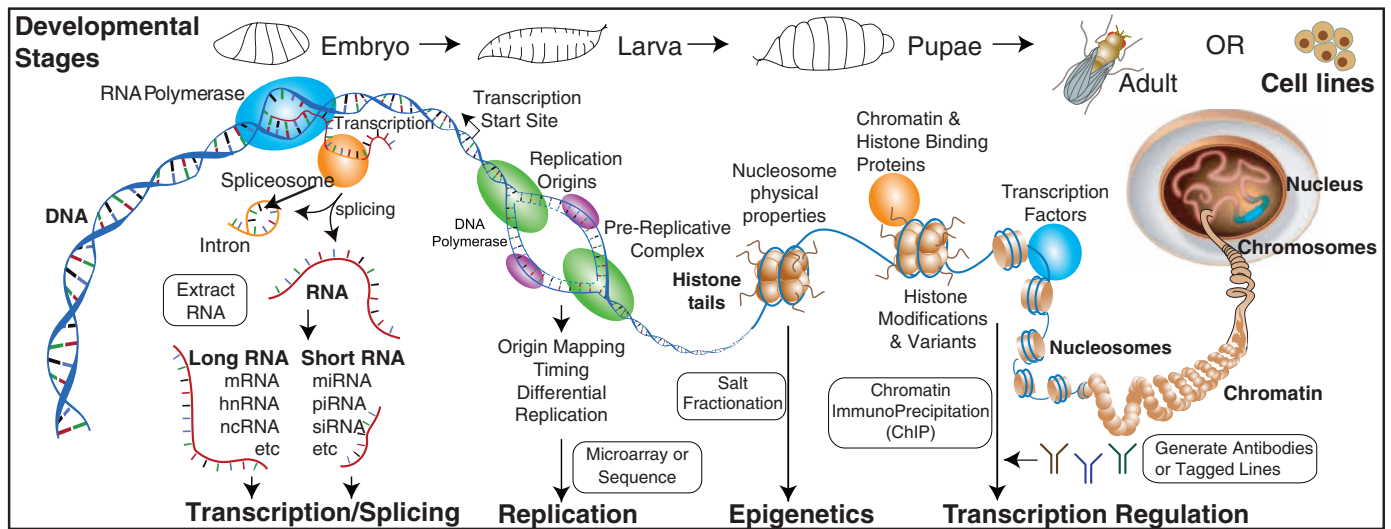

Fig. 1. Overview of *Drosophila* modENCODE data sets. Range of genomic elements and trans factors studied, with relevant techniques and resulting genome annotations. hnRNA, heterogeneous nuclear RNA.

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02140, USA. [3]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA. [4]Institute for Genomics and Systems Biology, Department of Human Genetics, The University of Chicago, 900 East 57th Street, Chicago, IL 60637, USA. [5]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710, USA. [6]Department of Genome Dynamics, Lawrence Berkeley National Laboratory (LBNL), 1 Cyclotron Road, Berkeley, CA 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, NY 10065, USA. [9]Genome Sciences Division, LBNL, 1 Cyclotron Road, Berkeley, CA 94720, USA. [10]Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK. [11]Department of Medicine and Department of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [12]Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institutes of Health (NIH), Bethesda, MD 20892, USA. [13]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [14]Department of Genetics and Developmental Biology, University of Connecticut Stem Cell Institute, 263 Farmington,

CT 06030–6403, USA. [15]Department of Biology CB-1137, Washington University, Saint Louis, MO 63130, USA. [16]Affymetrix, Santa Clara, CA 95051, USA. [17]Division of Extramural Research, National Human Genome Research Institute, NIH, 5635 Fishers Lane, Suite 4076, Bethesda, MD 20892–9305, USA. [18]Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario M5G 0A3, Canada. [19]Department of Genetics and *Drosophila* RNAi Screening Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [20]Center for Genome Sciences, Washington University, 4444 Forest Park Boulevard, Saint Louis, MO 63108, USA. [21]Department of Biology, Indiana University, 1001 East 3rd Street, Bloomington, IN 47405–7005, USA. [22]Whitehead Institute, Cambridge, MA 02142, USA. [23]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences and University Medical Center Utrecht, Utrecht, Netherlands. [24]Center for Genomics and Bioinformatics, Indiana University, 1001 East 3rd Street, Bloomington, IN 47405–7005, USA. [25]Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA. [26]Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [27]Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, USA. [28]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [29]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA.

[30]Machine Learning Group, Université Libre de Bruxelles, CP212, Brussels 1050, Belgium. [31]Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown, MA 02129, USA. [32]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA. [33]Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA. [34]Center for Biomolecular Science and Engineering, School of Engineering and Howard Hughes Medical Institute (HHMI), University of California Santa Cruz, Santa Cruz, CA 95064, USA. [35]Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY 11794, USA.

*The complete list of authors appears at the end of the paper.
†These authors contributed equally to this work.
‡These authors contributed equally to this work (listed alphabetically).
§These authors contributed equally to this work (listed alphabetically).
||To whom correspondence should be addressed. E-mail: manoli@mit.edu (M.K.) (integrative analysis); celniker@fruitfly.org (S.E.C.) (transcripts); karpen@fruitfly.org (G.H.K.) (chromatin); kpwhite@uchicago.edu (K.P.W.) (transcription factors); david.macalpine@duke.edu (D.M.M.) (replication); laie@mskcc.org (E.C.L.) (small RNAs); steveh@fhcrc.org (S.H.) (nucleosomes); lincoln.stein@gmail.com (L.D.S.) (data availability)

Protein-coding, structural, and noncoding transcripts. We searched for evolutionary signatures of conserved protein-coding DNA sequences in alignments of 12 *Drosophila* genomes (*12*, *20*) and for similarity to known proteins. Only 57 of 1938 previously undescribed gene models (*17*) contain a complete, conserved open reading frame (ORF) likely to represent unidentified protein-coding genes (Fig. 2A). An additional 81 gene models 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 (*17*). These 138 genes show nearly sixfold lower average expression than known protein-coding genes [fragments per kilobase of transcript per million fragments sequenced (FPKM) of 6.7 versus 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. For the remaining 1800 gene models, we find no evidence of protein-coding selection using PhyloCSF and no similarity to known protein sequences using blastx, suggesting that they are unlikely to represent protein-coding genes (*20*).

We looked for properties of noncoding 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 FlyBase (*4*) that do not belong to major classes of structural RNAs, such as miRNAs and snoRNAs. We predicted high-confidence structures for 132 transcripts (7.6%) using the RNAz program (*21*), suggesting conserved function as structural RNAs, similar to the fraction (7.8%) of transcripts with predicted structure observed in FlyBase ncRNAs (*4*). We revealed candidate structural RNAs in the newly

predicted transcripts (Fig. 2B), as well as previously unidentified structural elements in well-studied ncRNAs, including sex-chromosome dosage compensation regulator roX2 and heat-shock regulator HSRω (fig. S1) (*17*). However, the lack of highly structured regions in the vast majority of ncRNAs suggests functions independent of secondary structure.

Argonaute-associated small regulatory RNAs. Our analysis of deeply sequenced ~18- to 28-nucleotide (nt) RNAs dramatically extended the catalog of Ago-dependent small regulatory RNAs (*22*), including miRNAs, siRNAs, and piwi-associated RNAs (piRNAs). In the canonical miRNA pathway, ~21- to 24-nt RNAs are cleaved from hairpin precursors by Drosha and Dicer-1 ribonuclease (RNase) III enzymes and loaded into AGO1 effector complexes to repress mRNA targets. We annotated 61 additional canonical miRNAs, 12 of which are derived from the antisense strands of known miRNA loci (*23*), which may provide an efficient route for evolution of new miRNA activities. We unexpectedly detected miRNAs that overlap mRNAs, including nine cases where conserved protein-coding regions harbor RNA hairpins cleaved into duplexes of miRNA and partner strand miRNA* species, many of which are found 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 the cis configuration, or are simply neutral substrates. We identified 15 additional mirtrons that generate miRNAs by splicing of short hairpin introns (*24*), doubling the number of known cases from 14 to 29. We defined up to seven hybrid mirtrons bearing 3′ tails, which appear to require processing by the exosome before dicing (*25*). 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 (*26–31*). Endo-siRNAs derive from three distinct sources: (i) diverse transposable elements (TEs), whose activity they restrict; (ii) seven genomic regions encoding long inverted-repeat transcripts, which direct the cleavage of specific mRNA targets; and (iii) bi-directionally transcribed regions. This last class mostly comprises convergent transcripts that overlap in their 3′ untranslated regions (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, including nearly one-quarter of overlapping 3′ UTRs (table S4).

Lastly, piRNAs are ~24- to 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 (*32*). Though many *Drosophila* piRNAs map uniquely to tens of master loci that serve as genetic repositories for TE defense (*32*), we found that the 3′ UTRs of hundreds of cellular transcripts also generate abundant Piwi-loaded primary piRNAs in somatic ovarian follicle cells (*33–35*). This suggests that beyond transposon control, the piRNA pathway may play a more general role in cellular gene regulation.

Large-scale organization of the chromatin landscape. Eukaryotic genomes are organized into large domains (~10 kb to megabases) that exhibit distinct chromatin properties, such as heterochromatic regions that cover one-third of the genome and are typically known for transcriptional silencing (*36*). Our analyses show that the chromatin composition, organization, and boundaries of heterochromatin display surprising complexity and plasticity among cell types (*37*). We find surprisingly active heterochromatic regions,
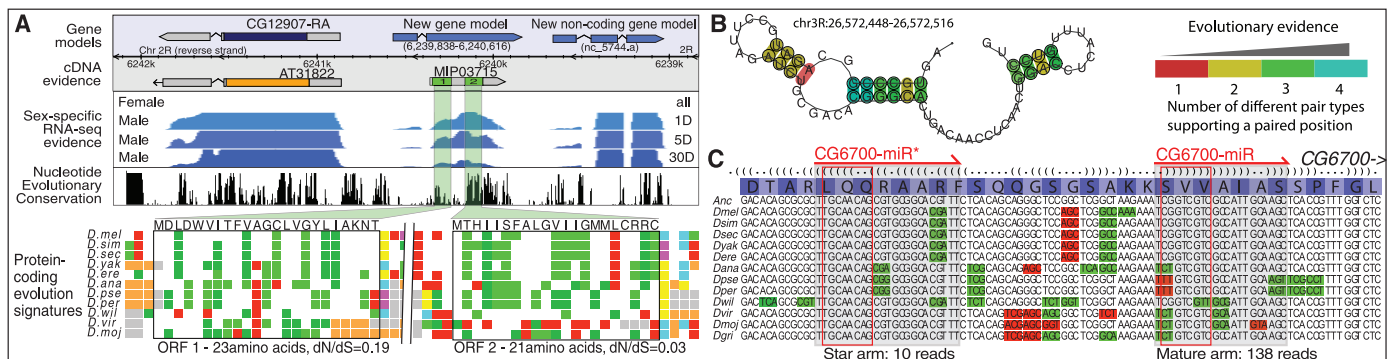

Fig. 2. Coding and noncoding genes and structures. (**A**) Extended region of male-specific expression in chromosome 2R including new protein-coding and noncoding transcripts. MIP03715 contains two short ORFs of 23 and 21 codons, respectively. ORF multispecies alignments (color coded) show abundant synonymous (bright green) and conservative (dark green) substitutions and a depletion of nonsynonymous substitutions (red), indicative of protein-coding selection [ratio of nonsynonymous to synonymous substitutions (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**) A transcribed region in chromosome 3R (26,572,290 to 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- to 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. Red boxes indicate 8-mer "seed" sequence potentially mediating 3′ UTR targeting.

with expression of 45% of pericentric heterochromatin genes (compared with 50% for euchromatic genes), and enrichment for both active and silent marks in active heterochromatic 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) (37).

We identified large domains with similar replication patterns by characterizing the *Drosophila* DNA replication program in cell lines, and we observed that the temporal replication program is determined by local chromatin environment (18, 38) and the density of replication initiation factors (39). We also found that specific euchromatic regions up to 300 kb were under-replicated in a tissue-specific manner in the polytene salivary glands, larval midgut, and fat bodies (40), which suggests that copy-number variation may help regulate gene expression levels.

Chromatin signatures characteristic of functional elements. Many genomic regulatory regions are difficult to identify because of a lack of characteristic sequence signatures, but they are often marked by specific histone modifications, variants, and other epigenetic factors (41, 42). To identify such signatures, we assayed 18 histone modifications and variants by ChIP-chip in multiple cell lines (18) and developmental stages (19), and we defined the physical properties of nucleosomes (43, 44). We correlated this information with gene annotations, transcriptome data sets, binding site profiles for replication factors,

insulator-binding proteins, and TFs to characterize chromatin signatures of each type of element (Fig. 3A). TSS-proximal regions were marked by H3K4me3 enrichment (45), depletion of nucleosome density, increased nucleosome turnover, and enrichment in the pellet chromatin fraction (43, 44). Gene bodies showed H2B ubiquitination covering the entire transcribed region and a 3′-biased enrichment of H3K36me3 and K3K79me1 marks. 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 (18). 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) (37).

Positional correlation analysis identified relationships between histone marks and nucleosome physical properties. Active marks [e.g., H3K27Ac, RNA polymerase II (RNA Pol II), H3K4me3] correlate with high chromatin solubility and high nucleosome-turnover rates, whereas 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 (44), and high levels of nucleosome turnover are indicative of a dynamic chromatin

structure (43), consistent with the biological functions associated with the corresponding marks.

We mapped origins of replication activated early in the 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 (46, 47). ORC-associated sequences are often found at TSSs and depleted for bulk nucleosomes, but are enriched for the variant histone H3.3 (39) and undergo active nucleosome turnover (43). These findings suggest that local nucleosome occupancy and organization are determinants of ORC binding in *Drosophila*, as in yeast (48, 49). By subdividing the ORC sites into TSS-proximal and -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* (38); H3K36me1, which is also enriched at cohesin sites (18), may be required in the absence of TSS-associated marks to promote ORC binding and subsequent cohesin loading (50, 51).

Insulator elements and proteins (e.g., CP190, CTCF, SUHW, and BEAF) block enhancer-promoter interactions and restrict the spread of histone modifications (52). Analysis of the genomic distributions of insulator proteins showed that BEAF32, CP190, and ZW5 preferentially bind upstream of TSSs, whereas SUHW binds
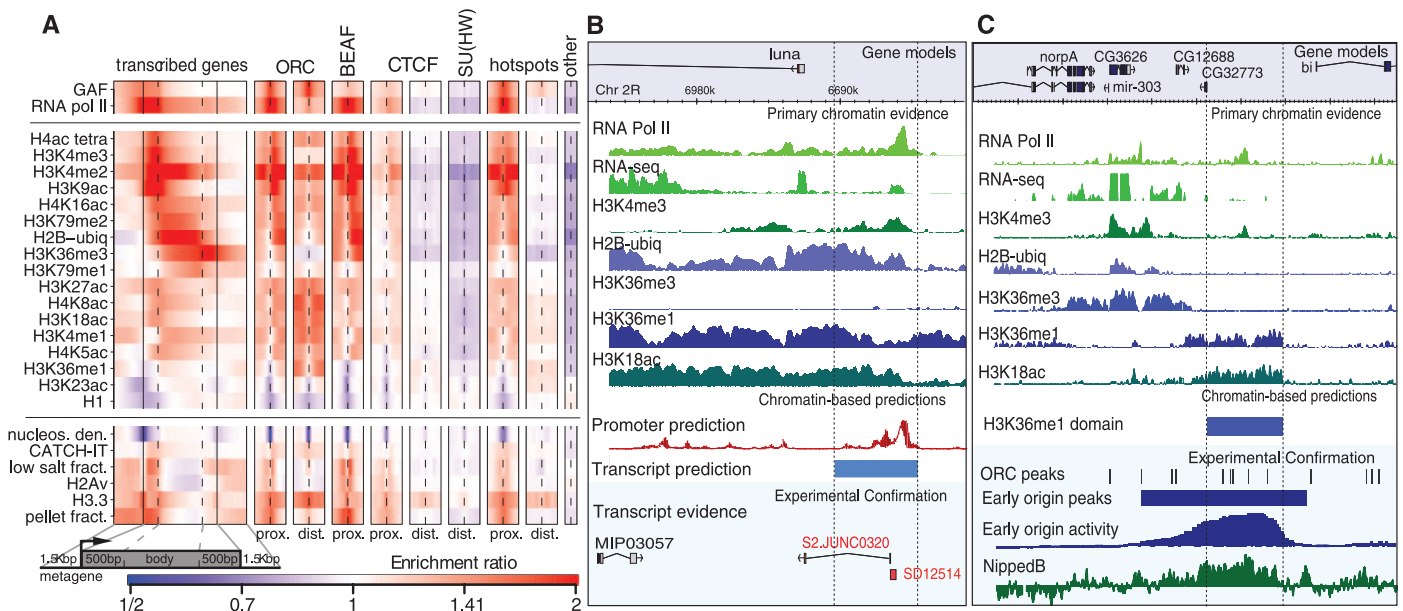

Fig. 3. Chromatin-based annotation of functional elements. (**A**) Average enrichment profiles of histone marks, chromosomal proteins, and physical chromatin properties at genes, origins of replications, insulator proteins, and TF binding positions. Each panel shows 4 kb centered at a specified location, either proximal to TSS (prox.) or distal (dist.). (**B**) Example of a transcript predicted by chromatin signatures associated with promoter (red trace) and gene bodies (blue box) and supported by cDNA evidence. Strong RNA Pol II and H3K4me3 peaks in the promoter region and strong H2B ubiquitination extending toward 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 (>10 kbp) intergenic H3K36me1/H3K18ac domains located outside of annotated genes. Although 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.

almost exclusively distal to TSSs, with CTCF binding both equally (53). 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 (54), 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 (45) 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 marks, ORC complex localization, and nucleosome stability and turnover in the 1-kb 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 2203 additional promoter positions at least 500 base pairs (bp) away from annotated TSSs (17). These included promoters for 10 primary miRNA transcripts, of which 7 were also identified by RNA-seq (14). We also used H3K36me3/H2B-ubiquitination signatures (fig. S3) to identify 53 transcribed gene bodies outside annotated genes, 11 of which are additionally supported by promoter predictions (e.g., Fig. 3B). These included four primary miRNA transcripts, of which three are also supported by RNA-seq (14) and one 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 (>10-kbp) intergenic domains of H3K36me1. In BG3 cells, 90 and 68%

of the intergenic H3K36me1 domains overlapped with cohesin (18) and early origin activity, respectively, as observed for a 20-kb region upstream of the *bi* gene (Fig. 3C and fig. S5). Although only 15% of early replication origins appear to be defined by intergenic H3K36me1 domains, the overlap with cohesion enrichment (18) 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 (55). We used two unsupervised, multivariate hidden Markov models to segment the genome on the basis of the combinatorial patterns of 18 histone marks in S2 and BG3 cells (Fig. 4 and fig. S6) (18). We did not seek a true number of distinct chromatin states; instead, we 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 9-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) (Fig. 4, left panel) (17). These showed distinct functional and genomic enrichments (Fig. 4, right panel) associated with different chromosomes (chromosome 4, male X), regulatory elements (promoters, enhancers), gene length and exonic structure (e.g., long first introns), gene function (e.g., developmental regulators), and 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, and d29 (c7 and c8) in heterochromatin (characterized by H3K9me2/3 enrichment and H3K23ac depletion). These states lack enrichments for other mapped factors [e.g., insulators, histone deacetylases (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 (45). Other active TSSs were additionally enriched for H3K36me1 and multiple acetylations (d13). Even within c1, some TSSs showed higher association with nucleosome turnover, group 1 insulator proteins and HDACs (d1, d3), whereas others were associated with heterochromatic genes of medium (d5) or low expression (d6).

The state analysis also captured the correlation 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 or 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 with the near nucleotide resolution of the ORC binding data.

Our states showed some similarities with the recently published five "colors" of chromatin from DNA adenine methyltransferase identification–mapped chromosomal proteins in Kc cells (56), 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.
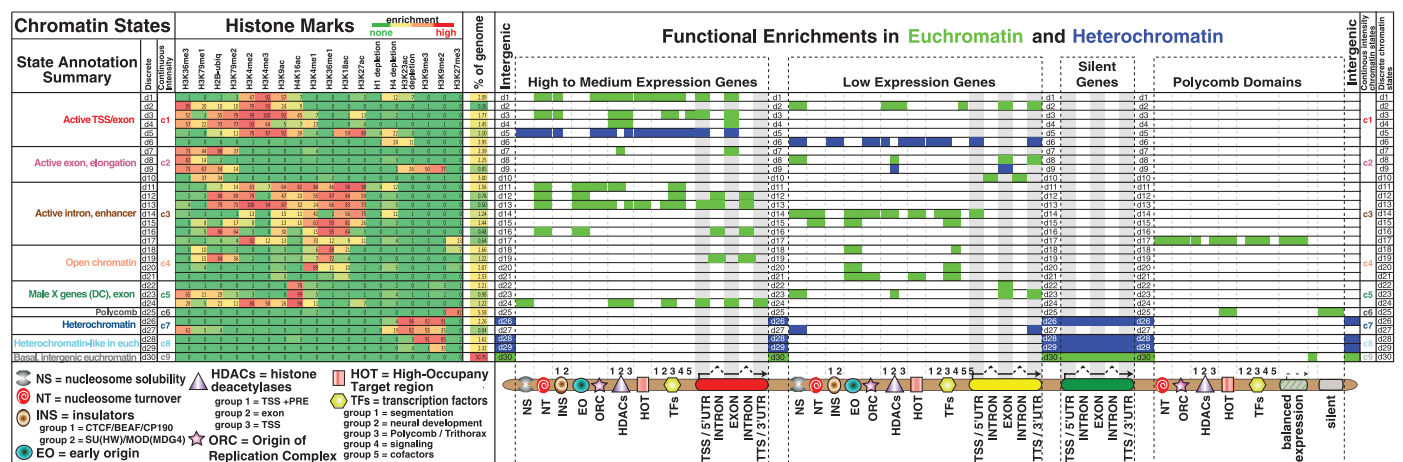
Fig. 4. Discovery and characterization of chromatin states and their functional enrichments. Combinatorial patterns of chromatin marks in S2 and BG3 cells reveal chromatin states associated with different classes of functional elements. A discrete model (states d1 to d30) captures the presence/absence information, and a continuous model (states c1 to c9) also incorporates mark intensity information (22). States were learned solely from mapped locations of marks (left) and were associated with modENCODE-defined elements (right) with most pronounced patterns in euchromatin (green) and heterochromatin (blue) shown here (additional variations shown in fig. S6).

Chromatin and motif properties of high-occupancy TF binding sites. Extensive overlap in the binding profiles of multiple TFs has revealed highly occupied target (HOT) regions or hotspots (19, 57–61). 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 corresponding to the number of distinct TFs bound (from 1 to ~21), resulting in 1962 hotspots with TF complexity of eight or greater, corresponding to ~10 overlapping factors bound (19). We correlated these regions with our and other data sets to gain insight 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 on the basis of its average complexity [the average number of TFs that co-bind (19)], 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, and TWI) with average complexity greater than four 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, which is suggestive of non-specific binding. For an additional 9 TFs, bound regions were enriched in the known motif, but no bias for lower-complexity regions was found; for another 10 factors, the known motif did not show a substantial 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) (19), 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 (19), suggesting potentially distinct roles for these elements. We observed enrichment for HOT regions 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 and d24) were enriched only at lower complexity (fig. S9D). In contrast, transcriptional elongation (d7 to 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* (39), this suggests hotspots as an alternative mechanism for ORC localization via nonspecific binding in high-accessibility regions, as well as widespread interplay between chromatin regulation, TF binding, and DNA replication. Given the high agreement between embryo and cell-line data sets, we propose that hotspots are stable genomic regions, kept open via recruitment of specific chromatin marks or remodelers, that facilitate binding of additional TFs at their motifs or nonspecifically.

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 (19) 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), whereas 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 nonrepetitive genome covered by the data sets, 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 Pol II, 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 more than 89% (Fig. 6A). Compared with previous annotations [FlyBase (4)], 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 versus 69.5%) in a comparison of the respective nonrepetitive portions.

We next determined the overlap between our predicted functional elements and PhastCons evolutionarily conserved elements across 12 *Drosophila* species, mosquitoes, honeybees, and beetles (62). These elements cover 38% of the *D. melanogaster* genome in 1.2 million blocks, over which
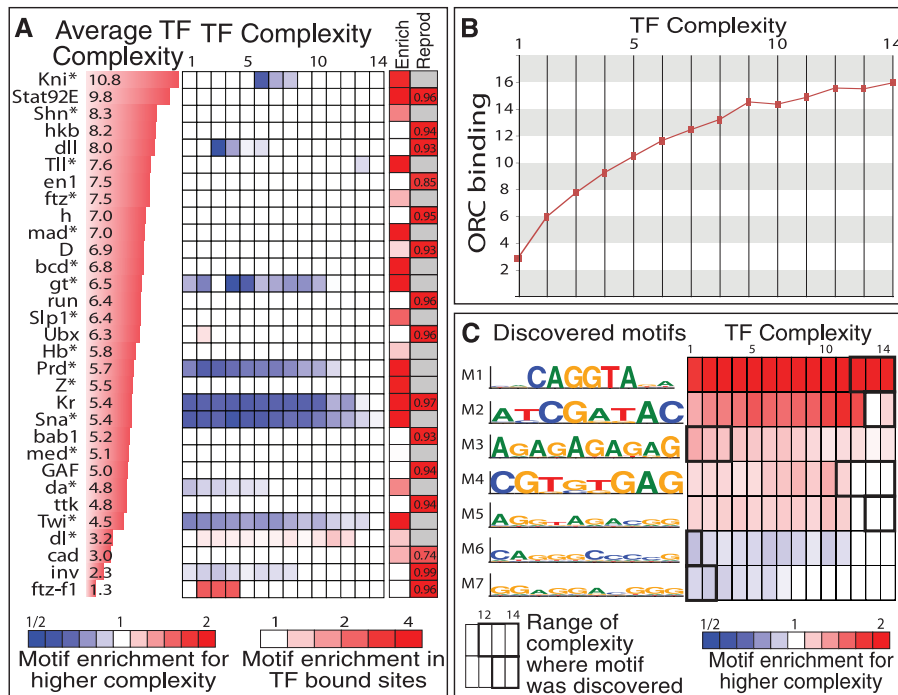

Fig. 5. High-occupancy TF binding regions and their relation to motifs, ORC, and chromatin. (**A**) Enrichment of known motifs for regions bound by corresponding TF, sorted by average complexity, denoting the number of distinct TFs bound in the same region. For eight TFs, motifs are depleted (blue) for 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 versus TF complexity. The relation between HOT spot complexity (x axis) and enrichment in ORC binding (y axis). (**C**) Discovered motifs in high- or low-complexity regions (boxed range) and their enrichment in regions of higher (red) or lower (blue) complexity. M1 to M5 are candidate "drivers" of HOT region establishment.

we repeated our previous individual and cumulative calculations. Thirty-two percent 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, providing additional independent evidence for the predicted elements (fig. S10). The only exceptions were some less active chromatin states, as expected, and introns, UTRs, and ncRNAs (63) providing 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 data sets, 10.8% of the genome is covered by 15 or more data sets, and coverage peaks at 103 data sets overlapping a single region on chromosome 3R. We found strong positive correlations between bound regulators and transcribed element densities, as well as regulators and chromatin element densities (fig. S11). In the case of chromatin data sets, additional chromatin marks resulted in higher accuracy in chromatin-state recovery (fig. S12), and we expect similar additional data sets to have an effect 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. In these networks, "nodes" represent the transcriptional and posttranscriptional 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 were 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 (17). The structure of the resulting network shows high connectivity and rapid spread of regulatory information, requiring traversal of only ~two regulatory connections, on average, between any two genes and no more than five connections between any pair of genes. Target genes are regulated by ~12 TFs, on average, and can have up to 54 regulatory TFs (17). 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 (17).

The physical regulatory network includes both pre- and posttranscriptional regulators, identifying the interplay between these two types of regulation. We organized the TFs of the physical regulatory network into five levels (Fig. 7A and fig. S13) on the basis of the relative proportion of TF targets versus TF regulators for each TF (64), and we 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 of the other TFs in the network, whereas 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 the two topmost levels versus 1.55 for the two lowest levels, fold enrichment of 2.66). This suggests that at least some feedback from the lower levels to the 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 (65). We found eight network motifs in the physical regulatory network (66), five of which correspond to TF cooperation (motifs 1, 2, 4, 7, and 8), confirming observations of cobinding and cotargeting (57–61). 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 and 6), which can lead to different delay properties in the expression of target genes depending
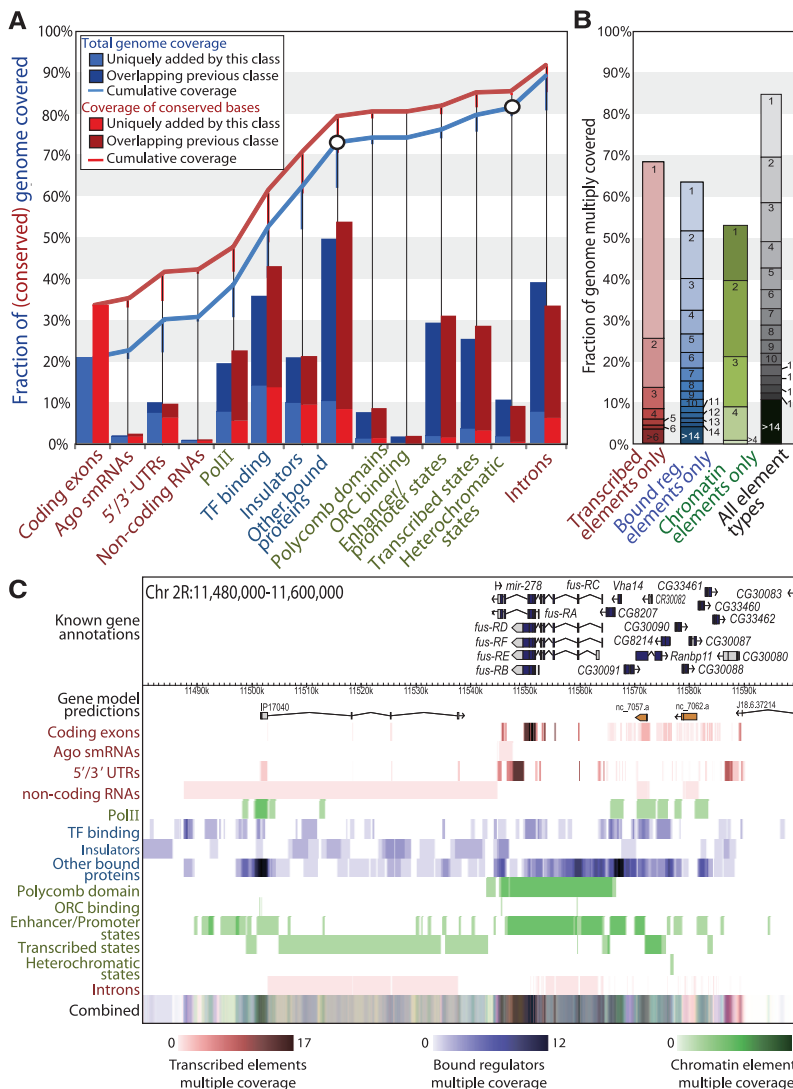

Fig. 6. Genome coverage by modENCODE data sets. (**A**) Unique (bars) and cumulative (lines) coverage of nonrepetitive (blue line) and conserved (red line) genomes. (**B**) Multiple coverage for data sets grouped into transcribed elements (red), bound regulators (blue), and chromatin domains (green) (17). 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 a Chr2R region with no prior annotation (left half), now showing multiple overlapping data sets. Coverage by different tracks is highly clustered (fig. S11), with some regions showing little coverage and others densely covered by many types of data.

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, which is also observed as a means for feedback in the hierarchical network layout (*17*).

Data set 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 (*60*). 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 (*17*) and 14,444 targets for which at least one line of evidence was available (*17*).

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 (*17*). 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 versus 104 and 76, respectively) and more regulators per target (24 versus 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 more than fourfold enrichment in both networks, even though the two only showed a 1.6-fold enrichment to each other's connections (fig. S14C). Compared with 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 relevance compared with both the motif and binding networks, including increased functional similarity, increased expression correlation, and increased protein-protein interactions of cotargeted genes (fig. S14D) (*17*). The REDfly network slightly outperformed the functional network, confirming the relevance of the metrics. However, the functional network contains 100 times more targets (9436 versus 88) and 1000 times more connections (231,181 versus 233) than the REDfly network, 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) (*17*) and predicted likely functional GO terms for every gene with a guilt-by-association approach that uses GO terms of annotated genes to predict likely functions of unannotated genes, allowing for multiple annotation predictions for each gene (*17*). This resulted in a higher predictive power than the use of expression or regulators alone (Fig. 8). 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 set S15). In general, 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) (*67*), we searched for splits (a point at which previously coexpressed genes begin to exhibit divergence into two or three distinct expression patterns) among a set of more than 6000 genes with the largest expression changes occurring during the developmental time course (Fig. 9A and fig. 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 (*17*). Several predictions agreed with literature support. For example, TIN, a known regulator of organ development (*68*), was a predicted regulator of genes with an early increase in expression and enriched for organ development ($P < 10^{-53}$), and E2F2, a known cell-cycle regulator (*69*), was a predicted regulator of genes with an early decrease in expression and enriched for cell-cycle function ($P < 10^{-100}$).
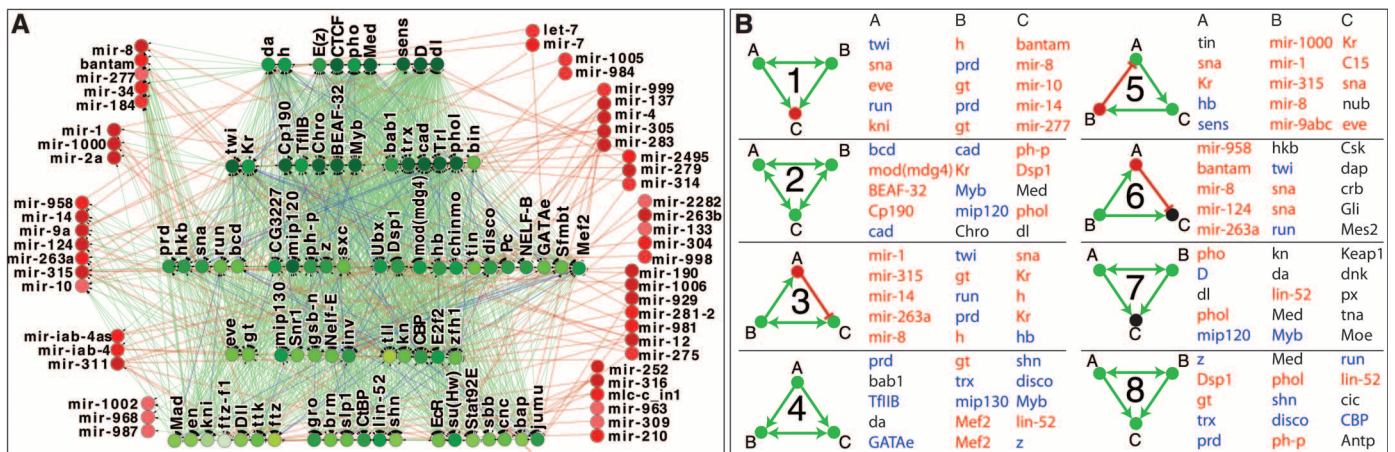

Fig. 7. Properties of the physical regulatory network. (**A**) Hierarchical view of mixed ChIP-based/miRNA physical regulatory network that combines transcriptional regulation by 76 TFs (green) from ChIP experiments and posttranscriptional regulation by 52 miRNAs (red). TFs are organized in a five-level 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 nodes indicate more targets). Ninety-two percent 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 a mixed physical regulatory network including TFs (green), miRNAs (red), and target genes (black). For each motif, five examples are shown. Known activators, blue; known repressors, red; other TFs, black.

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 (70). We generally observed a notable 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 ($P < 10^{-10}$) (fig. S19) (17).

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 signa-tures of cell-type–specific activators and repressors probably 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 (17). 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 one to eight predicted regulators for each cell, including, for example, CREBA as a predicted S2 activator, H as a predicted BG3 repressor, and factors with the stereotypical homeo-box binding motif (HOX-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 (Fig. 4) (18, 19). For a small number of motifs, however, the chromatin enrichments did not show a consistent picture of opposite enrichments in activating versus 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.

Fig. 8. Gene function prediction from coexpression and co-regulation patterns. Receiver operator characteristic curves for GO terms with predicted new members and area-under-the-curve statistics. 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.
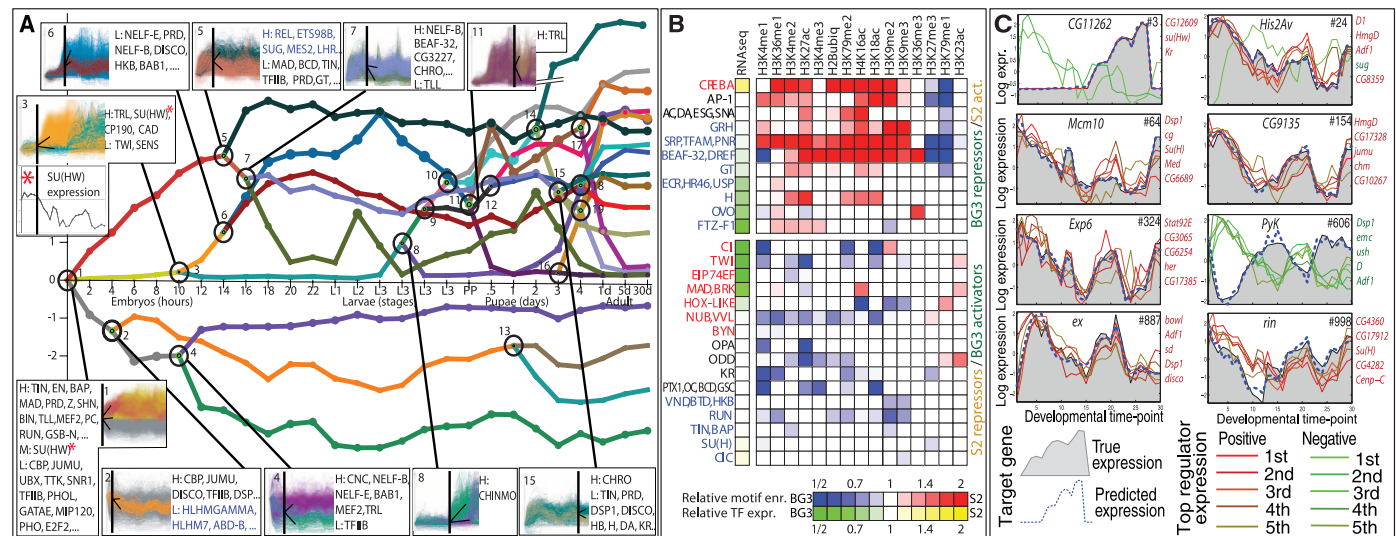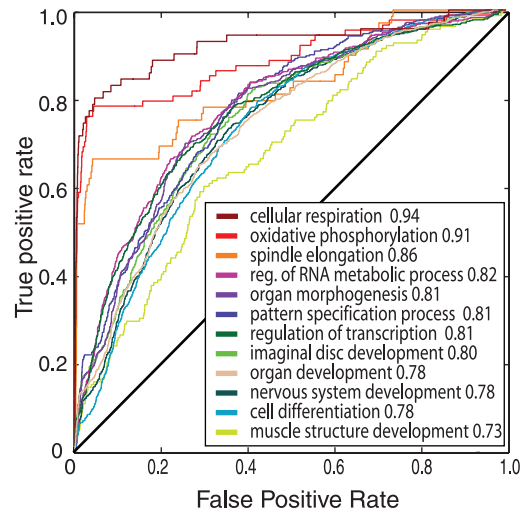

cellular respiration 0.94
oxidative phosphorylation 0.91
spindle elongation 0.86
reg. of RNA metabolic process 0.82
organ morphogenesis 0.81
pattern specification process 0.81
regulation of transcription 0.81
imaginal disc development 0.80
organ development 0.78
nervous system development 0.78
cell differentiation 0.78
muscle structure development 0.73

Fig. 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) (17). 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. The colored insets show the expression level of each individual gene going through the split and ranked regulators from the physical (black) or functional (blue) regulatory network associated with the higher (H), lower (L), or middle (M) path. The uncolored inset shows the expression of repressor SU(HW), whose expression decrease coincides with an expression increase of its targets (red asterisk). (**B**) Predicted S2 activators (top group) or repressors (bottom group), based on the coherence between relative expression of the TF in S2 (yellow) versus BG3 (green) and the relative motif enrichment (red) or depletion (blue) in S2 versus BG3 for activating (left columns) or repressive marks (right columns). (**C**) True (top of shaded area) and predicted (dotted blue line) expression levels for target genes, from the expression levels of inferred activators (red) and repressors (green). Only the top five positive and negative regulators are shown, ranked by their contribution to the expression prediction (weight of linear-regression model). Examples are shown from 8 of 1487 predictable genes, ranked by prediction quality scores (rank in upper right corner), evaluated as the averaged squared error between predicted and true expression levels across the time course. An expanded set of examples is shown in fig. S23.

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 (*71*). 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 their regulators. With the 30 distinct measurements of expression levels obtained by RNA-seq across development (*14*), we represented 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 10 intervals of three samples each and learned stable coefficients for linear combinations of TFs across 9 intervals to predict expression in the tenth (*17*).

We predicted the expression levels of 1991 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 and 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 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 (*14*) and an independently generated 12–time-point data set sampled at longer intervals (*19*) was significantly higher for predictable genes compared with unpredictable genes (Kolmogorov-Smirnov 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 whether the regulatory models obtained with whole-embryo time-course data sets can predict gene expression under novel conditions: specifically the Cl.8+, Kc167, BG3, and S2-DRSC cell lines. For each "predictable" gene, the expression levels of its 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 with 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 with only 10 to 15% for embryo-based unpredictable genes, providing further validation of our methodology.

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

Discussion. This first phase of the mod-ENCODE 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. Together, our resulting annotations cover 82% of the genome, a nearly fourfold increase compared with 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. Though 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 data sets, as well as new predictive models.

It remains to be seen how the general regulatory principles elucidated here will be conserved across the animal kingdom and especially in humans, through comparison across the ENCODE and modENCODE projects. Toward 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 (*72*), these will not only improve our understanding of fly biology, but can also serve as a template for understanding of human biology and disease.

References and Notes

1. www.genome.gov/10005107
2. S. E. Celniker *et al.*, *Nature* **459**, 927 (2009).
3. R. A. Hoskins *et al.*, *Science* **316**, 1625 (2007).
4. Compared to FlyBase release 5.12 (October 2008), available at http://fb2008_09.flybase.org/
5. M. Stapleton *et al.*, *Genome Biol.* **3**, RESEARCH0080 (2002).
6. K. H. Wan *et al.*, *Nat. Protoc.* **1**, 624 (2006).
7. R. Drysdale, FlyBase Consortium, *Methods Mol. Biol.* **420**, 45 (2008).
8. G. Dietzl *et al.*, *Nature* **448**, 151 (2007).
9. S. Mohr, C. Bakal, N. Perrimon, *Annu. Rev. Biochem.* **79**, 37 (2010).
10. H. J. Bellen *et al.*, *Genetics* **167**, 761 (2004).
11. *Drosophila* 12 Genomes Consortium, *Nature* **450**, 203 (2007).
12. A. Stark *et al.*, *Nature* **450**, 219 (2007).
13. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
14. B. Graveley, *Nature*, 10.1038/nature09715.
15. L. Cherbas *et al.*, *Genome Res.*, 10.1101/gr.112961.110.
16. R. A. Hoskins *et al.*, *Genome Res.*, 10.1101/gr.112466.110.
17. Supplemental text and materials and methods are available as supporting material on *Science* Online.
18. P. V. Kharchenko *et al.*, *Nature*, 10.1038/nature09725.
19. TF binding, hotspots, TF motif instances, promoter and enhancer validations, 12-point expression, and chromatin time course are available at www.cistrack.org.
20. M. F. Lin *et al.*, *Genome Res.* **17**, 1823 (2007).
21. S. Washietl, I. L. Hofacker, P. F. Stadler, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2454 (2005).
22. V. N. Kim, J. Han, M. C. Siomi, *Nat. Rev. Mol. Cell Biol.* **10**, 126 (2009).
23. E. Berezikov *et al.*, *Genome Res.*, 10.1101/gr.116657.110.
24. W.-J. Chung *et al.*, *Genome Res.*, 10.1101/gr.113050.110.
25. A. S. Flynt, J. C. Greimann, W. J. Chung, C. D. Lima, E. C. Lai, *Mol. Cell* **38**, 900 (2010).
26. W. J. Chung, K. Okamura, R. Martin, E. C. Lai, *Curr. Biol.* **18**, 795 (2008).
27. Y. Kawamura *et al.*, *Nature* **453**, 793 (2008).
28. B. Czech *et al.*, *Nature* **453**, 798 (2008).
29. M. Ghildiyal *et al.*, *Science* **320**, 1077 (2008); 10.1126/science.1157396.
30. K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, E. C. Lai, *Cell* **130**, 89 (2007).
31. K. Okamura *et al.*, *Nature* **453**, 803 (2008).
32. J. Brennecke *et al.*, *Cell* **128**, 1089 (2007).
33. N. C. Lau *et al.*, *Genome Res.* **19**, 1776 (2009).
34. N. Robine *et al.*, *Curr. Biol.* **19**, 2066 (2009).
35. K. Saito *et al.*, *Nature* **461**, 1296 (2009).
36. J. C. Eissenberg, G. Reuter, *Int Rev. Cell. Mol. Biol.* **273**, 1 (2009).
37. N. C. Riddle *et al.*, *Genome Res.*, 10.1101/gr.110098.110.
38. M. L. Eaton *et al.*, *Genome Res.*, 10.1101/gr.116038.110.
39. H. K. MacAlpine, R. Gordân, S. K. Powell, A. J. Hartemink, D. M. MacAlpine, *Genome Res.* **20**, 201 (2010).
40. J. Nordman, S. Li, T. Eng, D. MacAlpine, T. L. Orr-Weaver, *Genome Res.*, 10.1101/gr.114611.110
41. G. C. Hon, R. D. Hawkins, B. Ren, *Hum. Mol. Genet.* **18**, R195 (2009).
42. T. Kouzarides, *Cell* **128**, 693 (2007).
43. R. B. Deal, J. G. Henikoff, S. Henikoff, *Science* **328**, 1161 (2010).
44. S. Henikoff, J. G. Henikoff, A. Sakai, G. B. Loeb, K. Ahmad, *Genome Res.* **19**, 460 (2008).
45. N. D. Heintzman *et al.*, *Nat. Genet.* **39**, 311 (2007).
46. S. Vashee *et al.*, *Genes Dev.* **17**, 1894 (2003).
47. D. Remus *et al.*, *Cell* **139**, 719 (2009).
48. M. L. Eaton, K. Galani, S. Kang, S. P. Bell, D. M. MacAlpine, *Genes Dev.* **24**, 748 (2010).
49. N. M. Berbenetz, C. Nislow, G. W. Brown, *PLoS Genet.* **6**, e1001092 (2010).
50. P. J. Gillespie, T. Hirano, *Curr. Biol.* **14**, 1598 (2004).
51. T. S. Takahashi, P. Yiu, M. F. Chou, S. Gygi, J. C. Walter, *Nat. Cell Biol.* **6**, 991 (2004).
52. K. S. Scott, P. K. Geyer, *EMBO J.* **14**, 6258 (1995).
53. N. Nègre *et al.*, *PLoS Genet.* **6**, e1000814 (2010).
54. C. Jin *et al.*, *Nat. Genet.* **41**, 941 (2009).
55. J. Ernst, M. Kellis, *Nat. Biotechnol.* **28**, 817 (2010).
56. G. J. Filion *et al.*, *Cell* **143**, 212 (2010).
57. S. MacArthur *et al.*, *Genome Biol.* **10**, R80 (2009).
58. A. Carr, M. D. Biggin, *EMBO J.* **18**, 1598 (1999).
59. C. Moorman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12027 (2006).
60. X. Y. Li *et al.*, *PLoS Biol.* **6**, e27 (2008).
61. R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, E. E. Furlong, *Nature* **462**, 65 (2009).
62. A. Siepel *et al.*, *Genome Res.* **15**, 1034 (2005).
63. S. Meader, C. P. Ponting, G. Lunter, *Genome Res.* **20**, 1335 (2010).
64. H. Yu, M. Gerstein, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14724 (2006).
65. U. Alon, *Nat. Rev. Genet.* **8**, 450 (2007).
66. S. Wernicke, F. Rasche, *Bioinformatics* **22**, 1152 (2006).
67. J. Ernst, O. Vainas, C. T. Harbison, I. Simon, Z. Bar-Joseph, *Mol. Syst. Biol.* **3**, 74 (2007).
68. E. E. M. Furlong, E. C. Andersen, B. Null, K. P. White, M. P. Scott, *Science* **293**, 1629 (2001); 10.1126/science.1062660.
69. L. A. Lee, T. L. Orr-Weaver, *Annu. Rev. Genet.* **37**, 545 (2003).
70. D. A. Harrison, D. A. Gdula, R. S. Coyne, V. G. Corces, *Genes Dev.* **7**, 1966 (1993).
71. E. H. Davidson *et al.*, *Science* **295**, 1669 (2002).
72. A. C. Spradling, *Genetics* **174**, 1 (2006).
73. This work was supported by the National Human Genome Research Institute as part of the modENCODE project under RC2HG005639 (M.K.), U01HG004271 (S.E.C.), U01HG004258 (G.H.K.), U01HG004264 (K.P.W.),

Complete Author List

Kellis (integration): Sushmita Roy, Jason Ernst, Pouya Kheradpour, Christopher A. Bristow, Michael F. Lin, Stefan Washietl, Ferhat Ay, Patrick E. Meyer, Luisa Di Stefano, Rogerio Candeias, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Manolis Kellis

Celniker (transcription): Jane M. Landolin, Joseph W. Carlson, Benjamin Booth, Angela N. Brooks, Carrie A. Davis, Michael O. Duff, Philipp Kapranov, Anastasia A. Samsonova, Jeremy E. Sandler, Marijke J. van Baren, Kenneth H. Wan, Li Yang, Charles Yu, Justen Andrews, Steven E. Brenner, Michael R. Brent, Lucy Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Norbert Perrimon, Peter Cherbas, Brenton R. Graveley, Susan E. Celniker, Charles L. G. Comstock, Alex Dobin, Jorg Drenkow, Sandrine Dudoit, Jacqueline Dumais, Delphine Fagegaltier, Srinka Ghosh, Kasper D. Hansen, Sonali Jha, Laura Langton, Wei Lin, David Miller, Aaron E. Tenney, Huaien Wang, Aarron T. Willingham, Chris Zaleski, Dayu Zhang

Karpen (chromatin): Peter V. Kharchenko, Michael Y. Tolstorukov, Artyom A. Alekseyenko, Andrey A. Gorchakov, Tingting Gu, Aki Minoda, Nicole C. Riddle, Yuri B. Schwartz, Sarah C. R. Elgin, Mitzi I. Kuroda, Vincenzo Pirrotta, Peter J. Park, Gary H. Karpen, David Acevedo, Eric P. Bishop, Sarah E. Gadel, Youngsook L. Jung, Cameron D. Kennedy, Ok-Kyung Lee, Daniela Linder-Basso, Sarah E. Marchetti, Gregory Shanower

White (transcription factors): Nicolas Nègre, Lijia Ma, Christopher D. Brown, Rebecca Spokony, Robert L. Grossman, James W. Posakony, Bing Ren, Steven Russell, Kevin P. White, Richard Auburn, Hugo J. Bellen, Jia Chen, Marc H. Domanus, David Hanley, Elizabeth Heinz, Zirong Li, Folker Meyer, Steven W. Miller, Carolyn A. Morrison, Douglas A. Scheftner, Lionel Senderowicz, Parantu K. Shah, Sarah Suchy, Feng Tian, Koen J. T. Venken, Robert White, Jared Wilkening, Jennifer Zieba

MacAlpine (replication): Matthew L. Eaton, Heather K. MacAlpine, Jared T. Nordman, Sara K. Powell, Noa Sher, Terry L. Orr-Weaver, David M. MacAlpine, Leyna C. DeNapoli, Queying Ding, Thomas Eng, Helena Kashevsky, Sharon Li, Joseph A. Prinz

Lai (small RNAs): Nicolas Robine, Eugene Berezikov, Qi Dai, Katsutomo Okamura, Eric C. Lai, Qi Dai, Gregory J. Hannon, Martin Hirst, Marco Marra, Michelle Rooks, Yongjun Zhao

Henikoff (nucleosomes): Jorja G. Henikoff, Akiko Sakai, Kami Ahmad, Steven Henikoff, Terri D. Bryson

Stein (data coordination center): Bradley I. Arshinoff, Nicole L. Washington, Adrian Carr, Xin Feng, Marc D. Perry, William J. Kent, Suzanna E. Lewis, Gos Micklem, Lincoln D. Stein, Galt Barber, Aurelien Chateigner, Hiram Clawson, Sergio Contrino, Francois Guillier, Angie S. Hinrichs, Ellen T. Kephart, Paul Lloyd, Rachel Lyne, Sheldon McKay, Richard A. Moore, Chris Mungall, Kim M. Rutherford, Peter Ruzanov, Richard Smith, E. O. Stinson, Zheng Zha

Oliver (comparative transcription): Carlo G. Artieri, Renhua Li, John H. Malone, David Sturgill, Brian Oliver, Lichun Jiang, Nicolas Mattiuzzo

RNA structure: Sebastian Will, Bonnie Berger

Program management: Elise A. Feingold, Peter J. Good, Mark S. Guyer, Rebecca F. Lowdon

REPORTS

High-Flux Solar-Driven Thermochemical Dissociation of CO$_2$ and H$_2$O Using Nonstoichiometric Ceria

William C. Chueh,[1] Christoph Falter,[2] Mandy Abbott,[1] Danien Scipio,[1] Philipp Furler,[2] Sossina M. Haile,[1]* Aldo Steinfeld[2,3]*

Because solar energy is available in large excess relative to current rates of energy consumption, effective conversion of this renewable yet intermittent resource into a transportable and dispatchable chemical fuel may ensure the goal of a sustainable energy future. However, low conversion efficiencies, particularly with CO$_2$ reduction, as well as utilization of precious materials have limited the practical generation of solar fuels. By using a solar cavity-receiver reactor, we combined the oxygen uptake and release capacity of cerium oxide and facile catalysis at elevated temperatures to thermochemically dissociate CO$_2$ and H$_2$O, yielding CO and H$_2$, respectively. Stable and rapid generation of fuel was demonstrated over 500 cycles. Solar-to-fuel efficiencies of 0.7 to 0.8% were achieved and shown to be largely limited by the system scale and design rather than by chemistry.

Long-term storage and long-range transport of the vast, yet intermittent and unevenly distributed, solar energy resource is essential for a transition away from fossil energy (*1*).

[1]Materials Science, California Institute of Technology, MC 309-81, Pasadena, CA 91125, USA. [2]Department of Mechanical and Process Engineering, Eidgenössische Technische Hochschule (ETH) Zürich, 8092 Zürich, Switzerland. [3]Solar Technology Laboratory, Paul Scherrer Institute, 5232 Villigen PSI, Switzerland.

*To whom correspondence should be addressed. E-mail: smhaile@caltech.edu (S.M.H.); aldo.steinfeld@ethz.ch (A.S.)

Chemical fuels, derived from CO$_2$ and/or H$_2$O, offer exceptional energy density and convenience for transportation, but their production using solar energy input has remained a grand challenge (*2–9*). Solar-driven thermochemical approaches to CO$_2$ and H$_2$O dissociation inherently operate at high temperatures and use the entire solar spectrum; as such, they provide an attractive path to solar fuel production at high rates and efficiencies in the absence of precious metal catalysts (*10*). In contrast to direct thermolysis of CO$_2$ and H$_2$O, two-step ther-

mochemical cycles using metal oxide redox reactions further bypass the CO-O$_2$ or H$_2$-O$_2$ separation problem (*11*). Among candidate redox materials, ferrite-based oxides exhibit relatively slow reaction rates, degradation in rates because of sintering, and losses because of uncontrolled volatilization, whereas ZnO, SnO$_2$, and analogous volatile oxides that sublime during decomposition require rapid quenching of gaseous products to avoid recombination (*10–18*). Cerium oxide (ceria) has emerged as a highly attractive redox active material choice for two-step thermochemical cycling because it displays rapid fuel production kinetics and high selectivity (*17*, *19–24*), where such features result, in part, from the absence of distinct oxidized and reduced phases. However, ceria-based thermochemical studies to date have largely been limited to bench-top demonstrations of components or individual steps of the solar fuel production cycle; assessment of cyclability has been limited, and the energy conversion efficiency has remained uncertain because of the relatively low gravimetric fuel productivity inherent to the nonstoichiometric process. Here, we demonstrate high-rate solar fuel production from both CO$_2$ and H$_2$O using a solar reactor subjected directly to concentrated radiation under realistic operating conditions relevant to large-scale industrial implementation, without the need for complex material microstructures and/or system design (e.g., additional quench or separation steps). The results provide compelling evidence for the viability of thermochemical approaches to solar fuel