Transcriptional Landscape of the Human and Fly Genomes: Nonlinear and Multifunctional Modular Model of Transcriptomes

A.T. Willingham, S. Dike, J. Cheng, J.R. Manak, I. Bell, E. Cheung, J. Drenkow, E. Dumais, R. Duttagupta, M. Ganesh, S. Ghosh, G. Helt, D. Nix, A. Piccolboni, V. Sementchenko, H. Tammana, P. Kapranov, the ENCODE Genes and Transcripts Group,* and T.R. Gingeras

Affymetrix, Inc., Santa Clara, California 95051

Regions of the genome not coding for proteins or not involved in *cis*-acting regulatory activities are frequently viewed as lacking in functional value. However, a number of recent large-scale studies have revealed significant regulated transcription of unannotated portions of a variety of plant and animal genomes, allowing a new appreciation of the widespread transcription of large portions of the genome. High-resolution mapping of the sites of transcription of the human and fly genomes has provided an alternative picture of the extent and organization of transcription and has offered insights for biological functions of some of the newly identified unannotated transcripts. Considerable portions of the unannotated transcription observed are developmental or cell-type-specific parts of protein-coding transcripts, often serving as novel, alternative 5′ transcriptional start sites. These distal 5′ portions are often situated at significant distances from the annotated gene and alternatively join with or ignore portions of other intervening genes to comprise novel unannotated protein-coding transcripts. These data support an interlaced model of the genome in which many regions serve multifunctional purposes and are highly modular in their utilization. This model illustrates the underappreciated organizational complexity of the genome and one of the functional roles of transcription from unannotated portions of the genome.

WIDESPREAD RECOGNITION OF THE PHENOMENA OF EXTENSIVE AND COMPLEX PATTERNS OF TRANSCRIPTION THROUGHOUT THE GENOMES OF MANY SPECIES

Within the past 5 years, multiple large-scale, unbiased experimental approaches have identified surprisingly large amounts of RNA transcription far exceeding that estimated to be required for the production of messenger RNA for known proteins. This transcriptional "dark matter" has been observed in (1) large-scale full-length cDNA sequencing (Okazaki et al. 2002; Carninci et al. 2005), (2) mapping of 3′ ends with serial analysis of gene expression (Chen et al. 2002; Saha et al. 2002), (3) mapping of 5′ ends by cap analysis of gene expression (Carninci et al. 2005), and (4) analysis of expression by massively parallel signature sequencing (Jongeneel et al. 2005). Whereas these approaches have been key to analyzing widespread transcription, genomic tiling arrays have made substantial contributions by being both unbiased in their interrogation coverage (i.e., not limited to annotated regions) and sensitive with detection of low-copy-number transcripts (for review, see Johnson et al. 2005). In 2002, the first systematic and unbiased analysis of transcription across human chromosomes 21 and 22 was carried out. Strikingly, the sites of transcription across these chromosomes was determined to be at least an order of magnitude greater than that observed for annotated protein-coding genes (Kapranov et al. 2002). Implicit in this discovery was that a significant proportion of transcribed RNA was likely noncoding, suggesting an unanticipated degree of RNA complexity and possible novel function roles for such noncoding (nc)RNAs.

Careful review of earlier studies of the complexity and characteristics of transcribed RNA in eukaryotic cells finds evidence of widespread transcription heralded decades ago (for discussion, see Willingham and Gingeras 2006). These original studies were focused on large RNAs (i.e., >200–300 nucleotides) and arrived at the common conclusion that the complexity of transcripts made by organisms ranging from sea urchins to humans not only seemed to be inexplicably sizable and complex, but also contained non-polyadenylated (poly(A)$^-$) RNAs that were more numerous than the standard polyadenylated (poly(A)$^+$) RNAs associated with protein-coding functions.

Current efforts are focused on mapping transcription at very high nucleotide resolutions across the entire human genome and further dissecting this RNA into classes and cellular compartments (T.R. Gingeras et al., in prep.), searching for detectable patterns of stable RNA structures (Washietl et al. 2005; Torarinsson et al. 2006) and cross-comparing transcriptional complexity between evolutionarily distant genomes (Khaitovich et al. 2006; Manak et al. 2006). These studies have come to challenge the relatively straightforward protein-coding-centric view of genome organization in which genes are structured in discrete loci with a few transcript isoforms being made in the locus in a linear and mostly nonoverlapping fashion. Such loci are canonically seen to contain regulatory promoters immediately adjacent to annotated 5′ ends of the encoded isoforms. As discussed in this paper, this "beads on a string" linear model is gradually being replaced by a more complicated interlaced architecture in which many discrete genomic loci (e.g., exons) serve a multitasking func-

*Group members listed in Acknowledgments.

tion in which sequences comprising such loci may also serve as promoters, and introns may serve as exons for overlapping transcripts on both strands (Gingeras 2006).

BRIEF BACKGROUND ABOUT TILING MICROARRAYS

In the work described here, we employed four different types of tiling arrays (for general review, see Mockler et al. 2005). The first contains a medium interrogation resolution of 20 nucleotides designed to interrogate the 44 regions of the human genome selected for the Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2004). Transcription for 11 cell lines and 12 tissues was profiled using this ENCODE tiling array. Second, a higher-resolution tiling array with an interrogation resolution of 5 nucleotides covering 10 human chromosomes (~30% of the genome) was developed (Cheng et al. 2005). This 98-array set contains 380 Mb of probe-coverage and, because of its size, sample analysis was limited to 8 developmentally diverse cell lines. However, RNA samples were fractionated based on cellular compartments (cytosolic versus nuclear) and physical structure (poly(A)$^+$ vs. poly(A)$^-$). Very recently, a third tiling array set was developed that allowed creation of a 91-array set containing about 1300 Mb of probe-coverage, which translates to 100% of the nonredundant human genome tiled at 5 nucleotides resolution (T.R. Gingeras et al., in prep.). Finally, a single-array, low-resolution (35 nucleotides, 3.2 Mb of coverage) tiling array was created for the *Drosophila* genome and, taking advantage of the wealth of developmental biology for this model organism, was used to profile the first 24 hours of embryogenesis in 2-hour time increments (Manak et al. 2006).

Analysis of data gathered from tiling arrays indicates detected sites of transcription (transcribed fragments) which are termed "transfrags" (TFs). The labeled targets hybridized to the tiling arrays are double-stranded (ds) cDNAs made, in most instances, from processed (e.g., capped, polyadenylated, and/or spliced) RNAs isolated from cells. Thus, the detected TFs represent the sum of all transcripts mapping to the interrogated positions. Since the target is ds-cDNA, there is no strand information present in the tiling array data obtained from such labeled samples. Intensity thresholds are determined by calculating the intensity distribution of probes that originate from bacterial negative control regions. This allows for an estimated 5% false-positive rate to be used to determine where sites of transcription are located (see Kampa et al. 2004).

MOST OF THE HUMAN GENOME IS TRANSCRIBED AS NUCLEAR PRIMARY (UNPROCESSED) RNAs

This transcriptional complexity with overlapping sense and antisense transcripts significantly complicates interpretations of genome organization and annotation. For example, a 500-kb region selected by the ENCODE project (ENr 233) on chromosome 15 (Fig. 1) contains only 15 RefSeq annotated genes (top track), and yet, significant amounts of overlapping transcription as detected by tiling arrays (see Affymetrix and Yale tracks) far exceed these limited annotations. Messenger RNAs cloned by the Mammalian Gene Collection (MGC), H-Invitational Gene Database (HInv-DB), and submitted to GenBank also highlight the large degree of alternative splice isoforms. Paired-end ditag (PET) sequencing by the Genome Institute of Singapore has identified the starts and ends of hundreds of thousands of mRNAs, and similar large-scale 5′ cap analysis of gene expression (CAGE) by the RIKEN has predicted scores of gene start sites. Taken together with expression data from tiling arrays generated by Yale and Affymetrix, this 500-kb region appears to be entirely transcribed in a series of overlapping and intertwined transcripts. Furthermore, such transcriptional complexity is not limited to this example region, but rather appears to be the case for most of the ENCODE regions in the human genome (ENCODE Project Consortium, in prep.). In fact, by taking all the annotated mRNAs and empirically detected processed RNAs present in the ENCODE regions and then extrapolating the amount of primary nuclear transcription required to produce these RNAs, these analyses have inferred that >90% of the genomic sequence can be transcribed as nuclear primary sequences when looking at the totality of all examined biological samples (ENCODE Project Consortium, in prep.). This is illustrated by the last two tracks of Figure 1.

GENOMIC TILING ARRAYS DETECT PROCESSED, CYTOSOLIC TRANSCRIPTS COVERING FIVE TIMES MORE OF THE GENOME THAN SEQUENCES OF ANNOTATED PROTEIN-CODING GENES

When the polyadenylated cytosolic RNA present in each of eight analyzed cell lines was mapped to the non-repeat portion of 10 human chromosomes (Cheng et al. 2005), approximately 4–5% of interrogated nucleotides were observed to be transcribed (Table 1). Furthermore, when the nonredundant union of all detected poly(A)$^+$ RNA present in the eight cell lines was created, the detected transcribed non-repeat portions of the genome increased to 10.14%, suggesting that a significant portion of this transcription is cell-line-specific. Given that all of the nucleotides present in the exons of protein-coding genes amounts to 1–2% of the total (including repeated portions) human genome, the observed amount of transcription derived from poly(A)$^+$ cytosolic RNAs in eight cell lines is conservatively estimated to be more than five-fold greater.

The RNA products of this transcription are not uniformly distributed within a cell, nor are poly(A)$^+$ RNAs the only transcripts produced. To gain additional information on cellular compartment and structure of transcribed RNAs, one cell line (HepG2) was selected, and RNA was fractionated into nuclear and cytosolic portions. Further characterization was made by selecting poly(A)$^+$ and poly(A)$^-$ RNAs from each compartment (Cheng et al. 2005). Approximately 10.2% and 51.3% of detected sequences were found exclusively in the cytosolic and

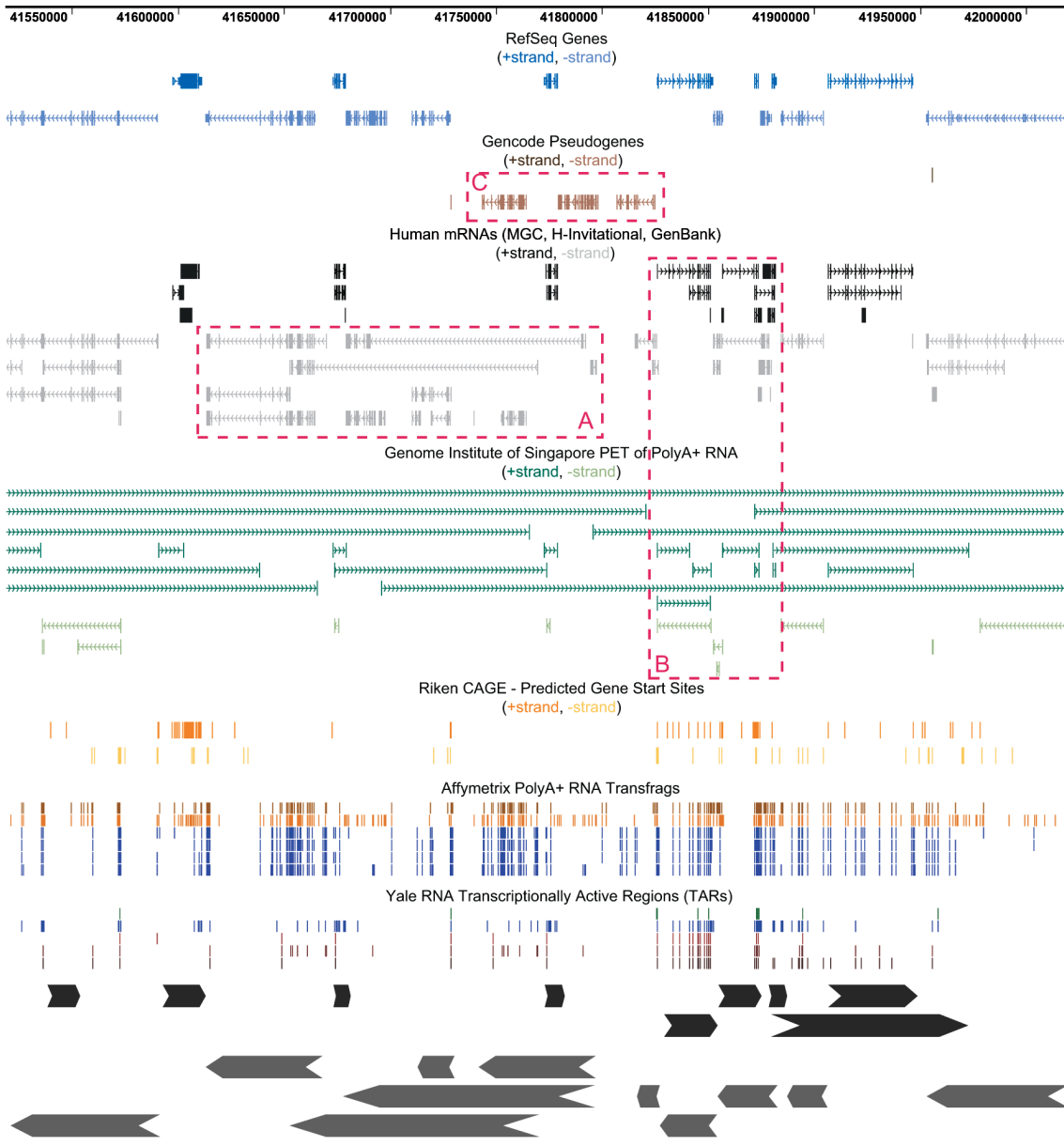


Figure 1. A 500-kb region of human chromosome 15 (ENCODE region ENr233) highlights the complexity of transcription observed across the genome. This region contains only 15 RefSeq annotated genes and yet contains (*A*) a variety of intertwined isoforms of annotated genes, (*B*) significant overlapping sense and antisense transcription, and (*C*) expressed pseudogenes (see Affymetrix transfrags). These examples are highlighted with red boxes. General areas of transcription are indicated by black (+ strand) and gray (– strand) arrow bars. Depicted annotations are based on the May 2004 human genome assembly and the UCSC Genome Browser (www.genome.ucsc.edu), where additional details on the represented data sets are available (Kent et al. 2002).

nuclear compartments (Table 2). This fivefold difference is striking, since it clearly indicates that most of the transcriptional products analyzed under these steady-state conditions never exit the nuclear compartment. These data also suggest that the sequences comprising the difference observed between the nuclear and cytosolic compartments are either synthesized and degraded, or that these products of widespread transcription have hitherto unknown nuclear functions. A total 19.4% and 43.7% of all detected transcription in both the nucleus and cytosol of HepG2 cells are exclusively poly(A)$^+$ or poly(A)$^-$, respectively (Table 2). The remainder of the detected transcription is bimorphic (detected as both poly(A)$^+$ and poly(A)$^-$). These data underscore the underappreciated prevalence (twofold greater overall) of poly(A)$^-$ transcription; for example, 84% of the cytosolic poly A$^-$ transcribed regions are located in unannotated genome regions. Furthermore, there is a renewed appreciation that many protein-coding transcripts exist within cells both with and without (or with shortened) polyadenylated 3′ termini. Although this bimorphic state has been previously described for some individual protein-coding transcripts (for list of genes, see Cheng et al. 2005), the extent of the number of genes exhibiting a bimorphic state was not understood.

Table 1. Poly(A)$^+$ RNA Detected in Eight Cell Lines

Sample	Coverage (bp)	% of all interrogated base pairs on the arrays
A-357	19, 330, 720	5.07
FHs 378Lu	18, 579, 012	4.87
Jurkat	18, 886, 873	4.96
NCCIT	21, 662, 254	5.68
PC-3	18, 242, 956	4.79
SK-N-AS	19, 088, 926	5.01
U-87 MG	14, 864, 583	3.90
HepG2	18, 899, 552	4.96
All cell lines (†)	38, 656, 627	10.14

Polyadenylated RNA map of the non-repeat portion 10 chromosomes representing ~30% of the human genome. For each cell line assayed, expressed poly(A)$^+$ RNA is listed as the number of nucleotides detected (coverage) and as a corresponding percentage of the total number of interrogated base pairs present on the microarrays. The nonredundant union of all detected poly(A)$^+$ RNA in the eight tested cell lines exceeds 10% of interrogated bases (†). Based on data in Cheng et al. (2005).

Table 2. Classes of RNAs Detected in HepG2 Cells

Sample	Coverage (bp)	% of all transcription detected in HepG2
Only in cytoplasmic fraction (‡)	6, 032, 310	10.25
Only in cytosolic poly(A)$^+$	1, 835, 709	3.12
Only in cytosolic poly(A)$^-$	3, 847, 281	6.53
Only in nuclear fraction (‡)	30, 207, 724	51.31
Only in nuclear (A)$^+$	5, 706, 194	9.69
Only in nuclear poly(A)$^-$	18, 237, 769	30.98
Only in poly(A)$^+$ RNA	11, 432, 433	19.42
Only in poly(A)$^-$ RNA	25, 747, 796	43.73
Both A$^+$ and A$^-$ RNA (†)	21, 693, 884	36.85

Percentages of HepG2 transcription detected in nuclear and cytosolic compartments. RNA from HepG2 cells was fractionated based on cellular compartment (nuclear vs. cytosolic) and polyadenylation (plus and minus). Total summing of transcription in each compartment (e.g., only in cytoplasmic fraction) includes the nonredundant contribution from both poly(A)$^+$ and poly(A)$^-$ fractions (‡). Sequences detected in both poly(A)$^+$ and (A)$^-$ are the nonredundant union (†). The number of nucleotides detected for poly(A)$^+$ RNA from nuclear (30,162,893 bp) and cytosolic (18,899,552 bp) fractions includes 15,936,128 bp present in both fractions. Based on data in Cheng et al. (2005).

However, is all this noncoding RNA biologically functional? Comparative genomic analysis of RNA structural elements has yielded intriguing clues and predicted thousands of functional ncRNAs. Conserved genomic sequences from several vertebrates were comparatively analyzed for base-pairing and thermodynamic stability contributions to structural conservation (Washietl et al. 2005). More than 30,000 RNA elements were identified in human, and nearly 1,000 were found conserved across vertebrates; furthermore, half of the structured RNA elements are located distant from known genes. In a separate study, Torarinsson et al. (2006) examined the approximately one-third of the nonrepetitive human genome that is not alignable with mouse, and a subset of these approximately 100,000 regions were analyzed for the presence of RNA structural elements. A significant number of "nonconserved" regions were found to have common RNA structure and, surprisingly, were twice as likely to overlap expressed transfrags as not to be expressed. Together, these studies begin to illustrate the value of profiling whole-genome transcription in multiple species and point the way to the genomic areas that will be the focus of additional studies.

UNANNOTATED ncRNA IS EXTENSIVELY UTILIZED DURING EARLY EMBRYOGENESIS OF *DROSOPHILA MELANOGASTER*

Within the first 24 hours of *Drosophila* embryogenesis, one oocyte and 15 nurse cells pattern and develop the complete musculature and nervous system for a larva which hatches about 24 hours postfertilization. This period of incredible transcriptional and developmental activity (e.g., early cell cycles are 7–8 minutes long) was interrogated with tiling arrays using samples gathered from 2-hour time points (Manak et al. 2006). In total over this 24-hour period, 27.6% of the 105.9-Mb nonrepetitive portion of the *Drosophila* genome is detected as transcribed RNAs with about 70% overlapping annotated gene structures (Fig. 2A). The 30% which is unannotated is almost half that observed for the human transcriptome and, if analysis is further restricted to intergenic transcription, the 7% observed in *Drosophila* is approximately fourfold lower. These differences are likely attributable to the significantly more compact fly genome (106 Mb, nonrepetitive) having a more complete annotation than the approximately tenfold larger human genome (1255 Mb, nonrepetitive). Interestingly, in *Drosophila* the unannotated regions are often expressed at specific and discrete time points, similar to what was observed between differing human cell lines. A comparison of the overall expression levels across the developmental time points, measured as the number of transcribed nucleotides, finds minimum total expression (8.2%) at the 4- to 6-hour time point which roughly corresponds with the lull between the degradation of the maternally contributed RNA and the initiation of zygotically produced transcripts (Fig. 2B). Conversely, maximal expression (16.1%) is observed at the 10- to 12-hour time point.

Direct comparisons of human and fly intergenic genomic sequences are made complicated by their evolutionary distances; however, a study comparing human and chimpanzee intergenic transcription highlights the value of broadening the scope of comparative genomics to include noncoding regions. Chimpanzee and human transcription were compared in three tissues and one cell line across 1% of their genomes (ENCODE regions) using tiling arrays (Khaitovich et al. 2006). Intergenic transcripts show patterns of tissue-specific conservation of expression comparable to protein-coding exons of known genes, suggesting intergenic transcripts evolved under similar levels of positive selection and functional constraint. Additionally, about half of observed intergenic transcription is differentially expressed between humans and chimps.

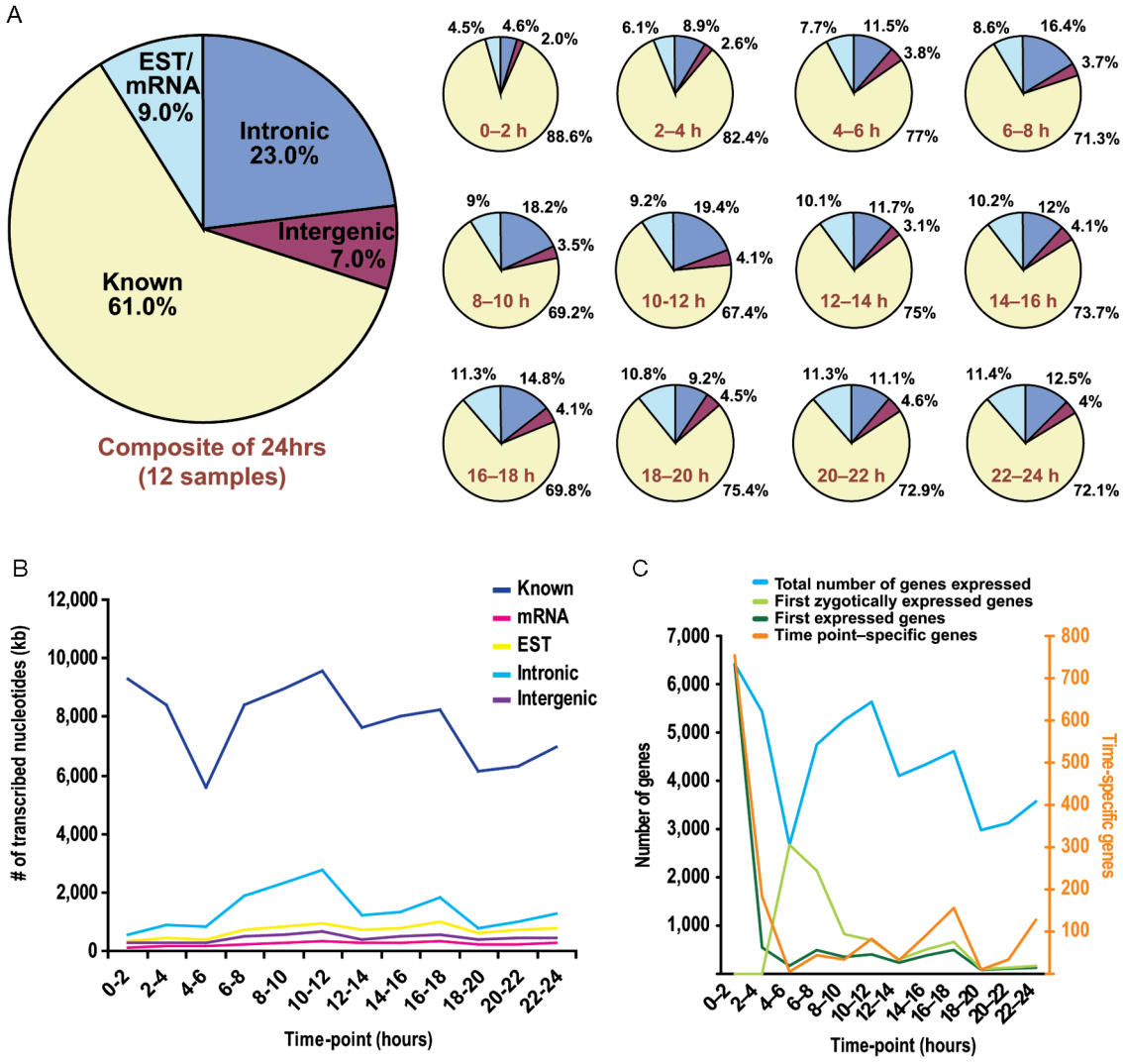


Figure 2. Analysis of transcription during *Drosophila* embryogenesis (0–24 hr). (*A*) Summary of annotated and unannotated (intronic and intergenic) transcription for each of 12 2-hr time points across the first 24 hrs of embryogenesis. (*B*) Comparison of the variation in overall expression levels (measured as number of transcribed kilobases of nucleotides) across time points for indicated classes of RNAs. (*C*) Relationship of expressed RefSeq genes: total number of expressed genes (*blue*), genes first expressed at a given time point (*dark green*), zygotically expressed genes (first 4 hrs disregarded due to maternal contributions of RNA) (*light green*), and genes expressed only in a specific time point (*orange*). (Based on data in Manak et al. 2006.)

Focusing on annotated fly protein-coding genes, a total of 9,808 genes are expressed during embryogenesis: 52.4% of a total of 18,716 known RefSeq genes (Fig. 2C). An average of approximately 4,400 are expressed in each time point, with an average duration of 10.8 hours. Despite the fact that average gene expression numbers in many hours, a significant number of genes (1,563; 16% of expressed genes) were found expressed at only one specific time point. It is worth noting that gene expression was computationally defined as the requirement that >70% of probes within an annotated gene be determined to be positive. Thus, it is likely that using this high level of stringency overlooks splice forms or other forms of partial gene expression and therefore underestimates overall transcription during embryogenesis.

ONE BIOLOGICAL FUNCTION OBSERVED FOR UNANNOTATED TRANSCRIPTION

A careful analysis of time-point-specific expression of genes and their correlation with adjacent unannotated transcription has proven a powerful method to understand the biological function of some unannotated transfrags. A convergence of a number of lines of evidence suggested that gene architectures are significantly more complicated than consensus annotations would suggest (see Fig. 1). Thus, we explored the possibility that this newly appreciated complex gene architecture could be related to some unannotated transcription and begin to address the important question concerning the biological function of the detected unannotated transcription. Specifically, one pos-

sibility was that some of the unannotated transfrags were previously unidentified parts of protein-coding genes. Interestingly, as suggested by the overlap of CAGE and Ditag data sets with transfrag data (Fig. 1), it is possible that some of the intergenic unannotated transfrags are novel 5′ transcriptional start sites (TSSs) which would allow the identification of novel regulatory regions, immediately 5′ to these transfrags. Therefore, two strategies were employed for the large-scale identification of alternative 5′ TSSs for protein-coding genes. For this study, both human and *Drosophila* genes were analyzed. Because of the limited number of protein-coding genes present within the ENCODE regions, a comprehensive empirical method relying on the use of the combination of 5′ RACE and tiling arrays was adopted, whereas in *Drosophila*, a whole-genome computational method was first developed to search for distal unannotated, developmental-stage-specific 5′ TSSs and upstream regulatory regions followed by the use of RT-PCR, cloning, and sequencing.

The hybridization of 5′ RACE reactions to tiling arrays forgoes time-consuming intermediate cloning and sequencing steps and permits the high-throughput assessment of gene structures. Products of the RACE reactions (RACEfrags) are resolved on a tiling array, thus providing all portions of a transcript that is connected to the site (index position) where the RACE primers are positioned (Kapranov et al. 2005). Thus, 5′ RACE reactions starting from a protein-coding exon proximal to the annotated 5′ end of a transcript would allow the identification of any novel exons that are upstream of the annotated end of the interrogated gene.

This strategy was applied to 399 genes contained within the 44 ENCODE regions that were subjected to 5′ RACE using RNA isolated individually from 12 human tissues. For the 359 loci with successful RACE products, almost half of the RACEfrags detected did not overlap with annotated exons. Nearly 80% of the genes were found to either have novel internal exons (between the index exon and the annotated 5′ end of the gene) or new 5′ extensions, with many of these new exons being tissue specific: 65.7% had 5′ extensions in at least one tissue, and 59.9% had new internal exons (ENCODE Project Consortium, in prep.). The distances to the tissue-specific unannotated 5′ extensions were surprisingly quite large, averaging about 108 kb for the new first intron, with 23% being ≥200 kb in size (Fig. 3A). These tissue-specific distal and often previously unannotated transcripts traverse several well-characterized protein-coding gene regions to reach the parent coding gene.

These findings are consistent with a large-scale analysis of TSSs conducted by Carninci et al., where sequencing of hundreds of thousands of CAGE tags culled from 145 mouse and 41 human libraries has permitted the quantitative analysis of the differential usage of promoters in numerous tissues (Carninci et al. 2006). Bidirectional promoters are found to be common, as are TSSs associated with internal exons, and the majority of protein-coding genes (58%) are shown to have alternative promoters. All told, TSSs are found to be abundant, tissue-specific, and bidirectional, further supporting emerging models for a complex overlapping network of transcription across the genome.

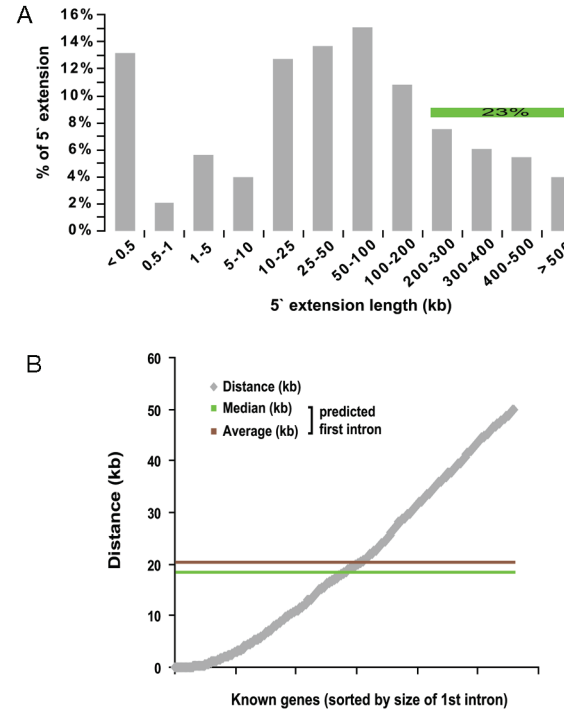


Figure 3. Many human and *Drosophila* genes have extensive, previously unannotated 5′ exons. (*A*) Summary of the size distribution of the longest 5′ RACE extensions per gene and tissue for 350 human genes within the ENCODE regions. The average size of these newly identified first introns is 108 kb; with 23% ≥200 kb in size. From data produced by the ENCODE Genes & Transcripts Analysis Group (ENCODE Project Consortium, in prep.). (*B*) Distance of most distal predicted 5′ site from the nearest co-regulated RefSeq gene (50-kb window) expressed in *Drosophila* embryonic development. Average size of predicted first introns was found to be 20.4 kb; ~12-fold larger than the average size of RefSeq annotated first introns. Median size of predicted new introns was 18.4 kb. (Based on data in Manak et al. 2006.)

What are the consequences of such extensive 5′ extensions? Sequencing of clones containing a subset of 5′ RACE extensions finds that approximately 45% likely add sequences to the protein open reading frame. In one striking example, a chimeric transcript isoform of DONSON, a gene of unknown function, was identified (Fig. 4A). The formation of such fusion transcripts has previously been reported (Kapranov et al. 2005). Tissue-specific RACEfrags were identified in brain, heart, kidney, and lung; but it was a small-intestine-specific product that was most intriguing, since it appeared to directly overlap exons from ATP50 about 330 kb upstream. A RT-PCR product was cloned and sequenced, confirming that three exons from ATP50 including the DNA-binding domain were being fused to at least two exons from DONSON. Independent paired end ditag (PET) data also confirm the presence of such an mRNA species. Such a combinatorial approach to building a protein takes full advantage of the modular potential of a genome's architecture and may prove to be a novel method for increasing complexity of the proteome.

For *Drosophila*, the availability of a developmental time course of whole-genome transcription provided the high-resolution temporal data that were used for predicting 5′ extensions based on the transcriptional co-regulation of unannotated regions with proximal (within 50 kb) protein-coding loci (Manak et al. 2006). More than 1000 genes were predicted to have 5′ extensions, and the distribution of lengths to the predicted 5′ sites for each are graphed in Figure 3B. As was found in the analysis of the human protein-coding genes, the average size of the predicted first introns was quite large: averaging 20.4 kb versus 1.7 kb for RefSeq annotated first introns. A subset of 180 co-regulated transfrags were then tested by RT-PCR and sequencing with a total verification rate of 78% for predicting new 5′ exons. Conservatively, 29% of the unannotated transcribed sequences function as misannotated constitutive exons or alternative exons of protein-coding genes. A total of 15.6% of the unannotated transcribed sequences appear to be TSSs used by 11.4% of the genes expressed during embryogenesis.

In one striking example, the RhoGAP gene was shown to be significantly larger than its annotated form, with novel 5′ exons adding more than 1300 amino acids to the protein and a TSS over 50 kb upstream of the annotated start site (Fig. 4B). This new annotation of RhoGAP now includes three P-element transposon insertions near the TSS and one lethal insertion within the gene, which adds a new mutant allele interrupting the transcript to the repertoire for future characterization of RhoGAP function. Given the abundance of intergenic P-element insertions mapped in the *Drosophila* genome, a continuation of this type of analysis would potentially allow the connection of "orphan" transposons with adjacent protein-coding loci. In summary, novel 5′ extensions are predicted to "convert" 16 Mb genomic sequence from intergenic to intronic and, when summed with all observed annotated and unannotated transcription, strongly suggest that >85% of the *Drosophila* genome is expressed as nuclear primary transcripts. This observation again underscores the observation made for the nuclear primary transcripts in the ENCODE regions.

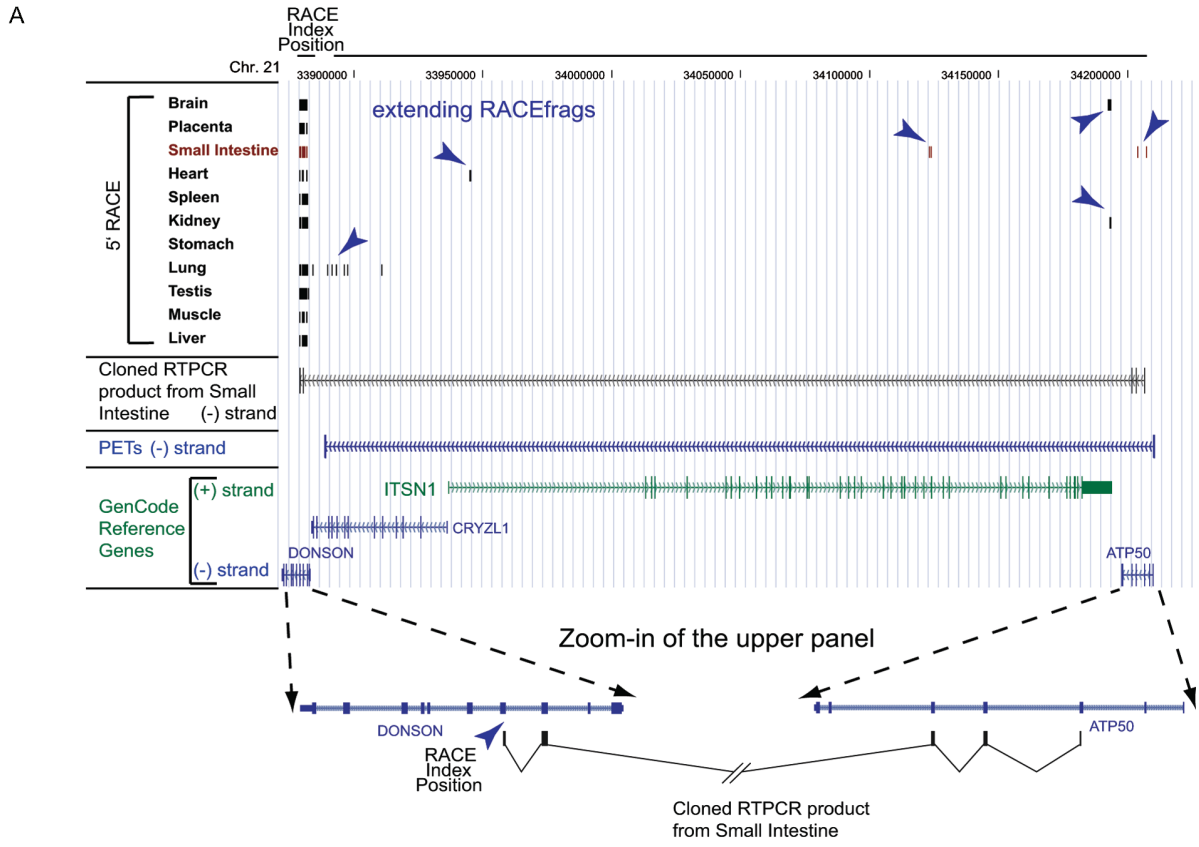
Overall, the abundance of unannotated 5′ TSSs is surprising. As discussed, these novel 5′ extensions can result in novel protein open reading frames and even "chimeric" transcripts composed of protein-coding exons contributed by several different genes. Alternative TSSs can introduce new promoters with different regulatory timing and expression strengths, some of which appear to be well-characterized promoters used by upstream genes. It is possible that the long-range transcription afforded by these 5′ extensions marks the genomic interval as transcriptional capable for DNA and chromatin modifying machineries. This may allow transcription of independently regulated coding and noncoding RNAs from the intronic regions or even processing of active RNAs from the transcribed intron itself. Last, large-scale analysis of TSSs has found that 28% (human) to 36% (mouse) of TSSs appear to be associated with noncoding transcripts (Carninci et al. 2006), adding an additional layer of transcriptional complexity.

IMPLICATIONS AND CONCLUSIONS

The emerging consensus from whole-genome studies of human and *Drosophila* transcription is that almost all of the non-repeat portions of both genomes are transcribed as nuclear primary RNAs: *Drosophila* >85% and human ENCODE regions about 93%. Additionally, cytosolic processed transcribed regions of the nonrepetitive human genome are also pervasive. As consequence of such widespread transcription, many regions of the genome consist of overlapping networks of intertwined, independently regulated transcripts originating from both strands. In *Drosophila*, approximately 9% (29% are previously missed unannotated exons of the total of 30% of the total annotated detected transcribed sequences) of the unannotated transcribed sequences appear to function as parts of protein-coding transcripts. In human, estimates of the percent of unannotated transcription which are likely to be portions of protein-coding transcripts are yet undetermined. Furthermore, ncRNAs are increasingly emerging as novel functional regulators in cellular processes such as microRNAs regulating cancers (Esquela-Kerscher and Slack 2006), NFAT signaling (Willingham et al. 2005), heat-shock sensing (Shamovsky et al. 2006), and multiple other instances (for review, see Mattick 2004; Storz et al. 2005; Zamore and Haley 2005; Goodrich and Kugel 2006). Especially interesting is the recent recognition that the most rapidly evolving human gene yet identified is in fact a ncRNA specifically expressed in the neocortex during a critical period of neurodevelopment where it is potentially involved in specifying structural aspects of the human cortex (Pollard et al. 2006).

Beyond their role as unannotated parts of mRNAs and functional ncRNAs, noncoding transcripts are likely acting on the level of gene regulation and genome architecture. Small RNAs have been implicated in transcriptional silencing and chromatin formation (for review, see Andersen and Panning 2003; Matzke and Birchler 2005). Noncoding transcription appears to encompass a number of locus control regions (LCRs), and the contributions of ncRNA transcription are investigated for the LCR of the human growth hormone (hGH) (Ho et al. 2006). The hGH LCR is 14.5 kb from the gene, and levels of hGH expression directly correlate with the levels of LCR transcription. This LCR transcription is bidirectional, does not seem to be directly connected to hGH as a 5′ extension, and, if inhibited by the insertion of a transcriptional terminator, results in reduced hGH transcription. RNA may play a role in the interchromosomal interactions of a single enhancer element with select odorant receptors (OR). Within single neurons, the *trans*-acting H-enhancer element on chromosome 14 colocalizes with only one of 1300 possible OR promoters present in the mouse genome and likely serves as a selection method for expression of a single receptor gene in individual sensory neurons (Lomvardas et al. 2006). Interestingly, H-enhancer DNA is shown to colocalize with OR receptor RNA, raising the possibility that transcribed RNA might contribute to stabilizing H-enhancer association with a single locus.

Of particular interest is the prevalence of small RNAs such as microRNAs, snoRNAs, and other regulatory RNAs



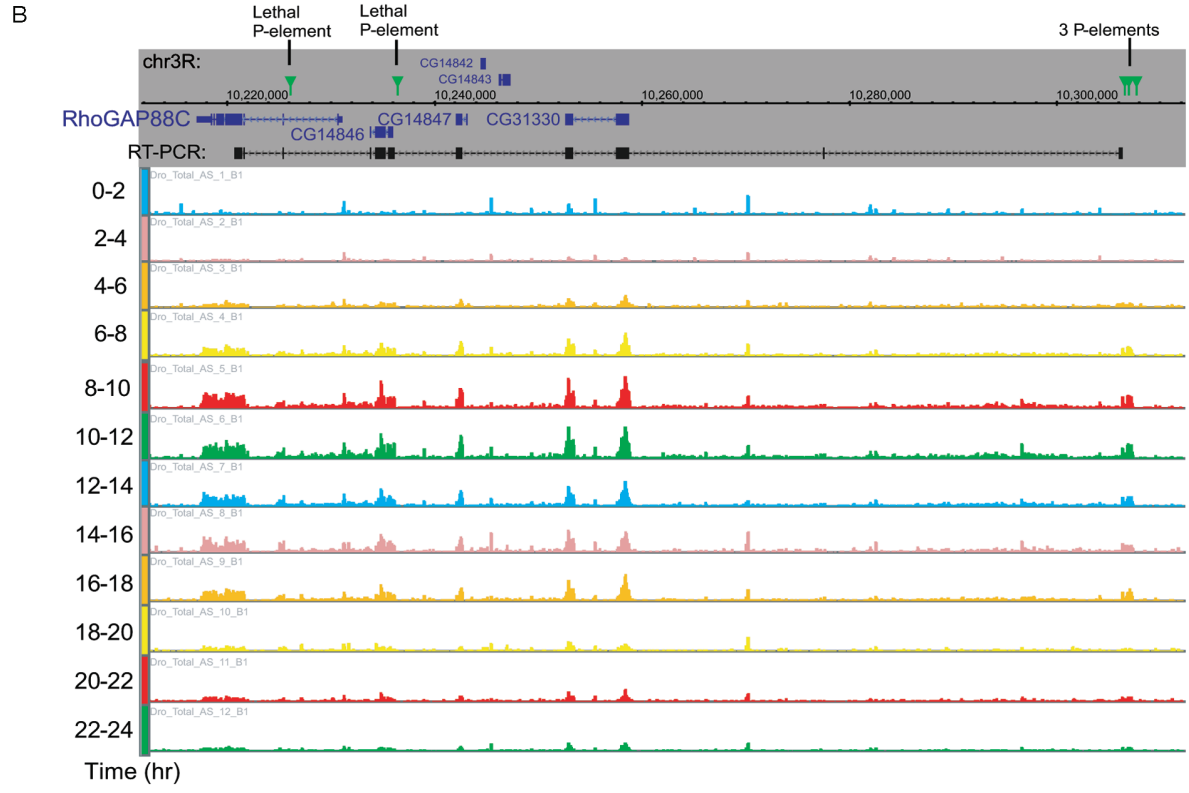


Figure 4. (*See facing page for legend.*)

(for review, see Mattick and Makunin 2005; Zamore and Haley 2005) as well as the surprising abundance of a newly discovered class of testes-specific approximately 30-bp RNAs which associate with the Piwi family and number in the tens of thousands (for review, see Kim 2006). Our group has begun to characterize the small RNA transcriptome on a whole-genome level using 5-bp resolution microarrays (T.R. Gingeras et al., in prep.). Known small RNA species such as miRNAs and snoRNAs are detected well, as are an astonishing number of novel small RNAs which could number over 400,000, averaging one small RNA every 3 kb. Ongoing investigations are exploring (1) small RNA biogenesis by perturbing proteins known to process small RNAs, (2) the relationship of these small RNAs to known gene annotations, and (3) the evolutionary conservation of these small RNAs. Clearly, as new technical approaches begin to answer old questions, they also elicit numerous new questions and insights into the organization and regulation of genomes.

ACKNOWLEDGMENTS

We are indebted to the ENCODE Genes & Transcripts Analysis Group for their invaluable computational, experimental contributions and discussions: Josep F. Abril, Tyler Alioto, Stylianos E. Antonarakis, Robert Baertsch, Peter Bickel, Ewan Birney, James B. Brown, Piero Carninci, Robert Castelo, Kuo Ping Chiu, Siew Woh Choo, Chiou Yu Choo, Jacqueline Chrast, Taane Clarke, France Denoeud, Emmanouil T. Dermitzakis, Mark C. Dickson, Olof Emanuelsson, Christoph Flamm, Paul Flicek, Sylvain Foissac, Adam Frankish, Claudia Fried, Mark Gerstein, James Gilbert, Roderic Guigó, Jörg Hackermüller, Jennifer Harrow, Yoshihide Hayashizaki, Charlotte N. Henrichsen, Jana Hertel, Heather Hirsch, Ivo L. Hofacker, Nancy Holroyd, Tim Hubbard, Chikatoshi Kai, Jun Kawai, Damian Keefe, Jan Korbel, Julien Lagarde, Zheng Lian, Jin Lian, Manja Lindemeyer, Todd M. Lowe, Caroline Manzano, Elliott H. Margulies, Nicholas Matthews, John S. Mattick, Kristin Missal, Zarmik Moqtaderi, Richard M. Myers, Ugrappa Nagalakshmi, Peter Newburger, Hong Sain Ooi, Sandeep Patel, Jakob S. Pedersen, Alexandre Reymond, Jane Rogers, Joel Rozowsky, Yijun Ruan, Albin Sandelin, Edward A. Sekinger, Atif Shahab, Michael Snyder, K.G. Srinivasan, Peter F. Stadler, Kevin Struhl, Wing-Kin Sung, David Swarbreck, Andrea Tanzer, Ruth Taylor, Daryl J. Thomas, Catherine Ucla, Stefan Washietl, Chia-Lin Wei, Matthew T. Weirauch, Sherman M. Weissman, Jiaqian Wu, Carine Wyss, Annie Yang, Xueqing Zhang, Xiao-Dong Zhao, Deyou Zheng, and Zhou Zhu. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400, and from the National Human Genome Research Institute, National Institutes of Health, under Grant No. U01 HG003147, and by Affymetrix, Inc.

REFERENCES

Andersen A.A. and Panning B. 2003. Epigenetic gene regulation by noncoding RNAs. *Curr. Opin. Cell Biol.* **15:** 281.

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N., Oyama R., Ravasi T., Lenhard B., Wells C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559.

Carninci P., Sandelin A., Lenhard B., Katayama S., Shimokawa K., Ponjavic J., Semple C.A., Taylor M.S., Engstrom P.G., Frith M.C., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626.

Chen J., Sun M., Lee S., Zhou G., Rowley J.D., and Wang S.M. 2002. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci.* **99:** 12257.

Cheng J., Kapranov P., Drenkow J., Dike S., Brubaker S., Patel S., Long J., Stern D., Tammana H., Helt G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636.

Esquela-Kerscher A. and Slack F.J. 2006. Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer* **6:** 259.

Gingeras T.R. 2006. The multitasking genome. *Nat. Genet.* **38:** 608.

Goodrich J.A. and Kugel J.F. 2006. Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* **7:** 612.

Ho Y., Elefant F., Liebhaber S.A., and Cooke N.E. 2006. Locus control region transcription plays an active role in long-range gene activation. *Mol. Cell* **23:** 365.

Johnson J.M., Edwards S., Shoemaker D., and Schadt E.E. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21:** 93.

Jongeneel C.V., Delorenzi M., Iseli C., Zhou D., Haudenschild C.D., Khrebtukova I., Kuznetsov D., Stevenson B.J., Strausberg R.L., Simpson A.J., and Vasicek T.J. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* **15:** 1007.

Kampa D., Cheng J., Kapranov P., Yamanaka M., Brubaker S., Cawley S., Drenkow J., Piccolboni A., Bekiranov S., Helt G., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14:** 331.

Kapranov P., Cawley S.E., Drenkow J., Bekiranov S., Strausberg R.L., Fodor S.P., and Gingeras T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916.

Figure 4. Examples of how unannotated transcriptional complexity can affect the landscape and composition of human and *Drosophila* genes. (*A*) 5′-RACE upstream of the DONSON gene identified tissue-specific 5′ exons (RACEfrags) spanning ~330 kb. A product specific to small intestine was shown by RT-PCR to be a novel fusion transcript incorporating three exons including the DNA-binding domain from the upstream ATP50 gene and at least two exons from DONSON. From data produced by the ENCODE Genes & Transcripts Analysis Group (ENCODE Project Consortium, in prep.). Gene annotations based on the May 2004 version of the human genome presented in the UCSC browser (Kent et al. 2002). (*B*) The annotated *Drosophila* RhoGAP gene is shown by transcriptional co-regulation and RT-PCR to have additional 5′ regions encompassing three separately in-silico annotated genes and a new unannotated 5′ exon. This extension includes four new P-element insertions and adds a new mutant allele interrupting the transcript for RhoGAP (this transposon was previously thought to disrupt an enhancer element). (Based on data in Manak et al. 2006.)

Kapranov P., Drenkow J., Cheng J., Long J., Helt G., Dike S., and Gingeras T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., and Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996.

Khaitovich P., Kelso J., Franz H., Visagie J., Giger T., Joerchel S., Petzold E., Green R.E., Lachmann M., and Paabo S. 2006. Functionality of intergenic transcription: An evolutionary comparison. *PLoS Genet.* (in press).

Kim V.N. 2006. Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev.* **20:** 1993.

Lomvardas S., Barnea G., Pisapia D.J., Mendelsohn M., Kirkland J., and Axel R. 2006. Interchromosomal interactions and olfactory receptor choice. *Cell* **126:** 403.

Manak J.R., Dike S., Sementchenko V., Kapranov P., Biemar F., Long J., Cheng J., Bell I., Ghosh S., Piccolboni A., and Gingeras T.R. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet*. **38:** 1151.

Mattick J.S. 2004. RNA regulation: A new genetics? *Nat. Rev. Genet.* **5:** 316.

Mattick J.S. and Makunin I.V. 2005. Small regulatory RNAs in mammals. *Hum. Mol. Genet.* **14:** R121.

Matzke M.A. and Birchler J.A. 2005. RNAi-mediated pathways in the nucleus. *Nat. Rev. Genet.* **6:** 24.

Mockler T.C., Chan S., Sundaresan A., Chen H., Jacobsen S.E., and Ecker J.R. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85:** 1.

Okazaki Y., Furuno M., Kasukawa T., Adachi J., Bono H., Kondo S., Nikaido I., Osato N., Saito R., Suzuki H., et al. (FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team). 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563.

Pollard K.S., Salama S.R., Lambert N., Lambot M.A., Coppens S., Pedersen J.S., Katzman S., King B., Onodera C., Siepel A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167.

Saha S., Sparks A.B., Rago C., Akmaev V., Wang C.J., Vogelstein B., Kinzler K.W., and Velculescu V.E. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20:** 508.

Shamovsky I., Ivannikov M., Kandel E.S., Gershon D., and Nudler E. 2006. RNA-mediated response to heat shock in mammalian cells. *Nature* **440:** 556.

Storz G., Altuvia S., and Wassarman K.M. 2005. An abundance of RNA regulators. *Annu. Rev. Biochem.* **74:** 199.

Torarinsson E., Sawera M., Havgaard J.H., Fredholm M., and Gorodkin J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16:** 885.

Washietl S., Hofacker I.L., Lukasser M., Huttenhofer A., and Stadler P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23:** 1383.

Willingham A.T. and Gingeras T.R. 2006. TUF love for "junk" DNA. *Cell* **125:** 1215.

Willingham A.T., Orth A.P., Batalov S., Peters E.C., Wen B.G., Aza-Blanc P., Hogenesch J.B., and Schultz P.G. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309:** 1570.

Zamore P.D. and Haley B. 2005. Ribo-gnome: The big world of small RNAs. *Science* **309:** 1519.