Science
AAAS

Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution

Jill Cheng, *et al.*

Science **308**, 1149 (2005);
DOI: 10.1126/science.1108625

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution

Jill Cheng,[1*] Philipp Kapranov,[1*] Jorg Drenkow,[1] Sujit Dike,[1] Shane Brubaker,[1] Sandeep Patel,[1] Jeffrey Long,[1] David Stern,[1] Hari Tammana,[1] Gregg Helt,[1] Victor Sementchenko,[1] Antonio Piccolboni,[1] Stefan Bekiranov,[1] Dione K. Bailey,[1] Madhavan Ganesh,[1] Srinka Ghosh,[1] Ian Bell,[1] Daniela S. Gerhard,[2] Thomas R. Gingeras[1]†

Sites of transcription of polyadenylated and nonpolyadenylated RNAs for 10 human chromosomes were mapped at 5–base pair resolution in eight cell lines. Unannotated, nonpolyadenylated transcripts comprise the major proportion of the transcriptional output of the human genome. Of all transcribed sequences, 19.4, 43.7, and 36.9% were observed to be polyadenylated, nonpolyadenylated, and bimorphic, respectively. Half of all transcribed sequences are found only in the nucleus and for the most part are unannotated. Overall, the transcribed portions of the human genome are predominantly composed of interlaced networks of both poly A+ and poly A– annotated transcripts and unannotated transcripts of unknown function. This organization has important implications for interpreting genotype-phenotype associations, regulation of gene expression, and the definition of a gene.

The current classification of protein-coding and noncoding genomic regions is based on intron-exon structures of well-characterized protein-coding genes. Noncoding genomic regions, which account for 98% to 99% of the human genome, consist of introns found within protein-coding transcripts and the intergenic regions between them (*1*, *2*). Recent observations indicate that noncoding regions are transcribed into polyadenylated, stable RNAs that are transported into the cytosol during development (*3*–*8*). The ENCODE consortium has suggested transcripts of unknown function (TUFs) as an unofficial name for these unannotated transcribed regions (*9*). Transcribed fragments (transfrags) are used to denote array-detected regions of transcription (*3*–*5*) representing exons of both well-characterized protein-coding genes and TUFs.

Although our understanding of poly A+ cytosolic TUFs has increased, much less is known about the synthesis sites of transcripts lacking 3′ polyadenylation (poly A–). Replication-dependent histone genes are currently considered to be the only transcripts synthesized exclusively as poly A– transcripts (*10*). However, 30 years ago, Milcarek *et al.* reported that approximately 30% of rapidly labeled polysomal-associated RNA in actinomycin D–inhibited HeLa cells was poly A– (*11*). Similarly, Salditt-Georgieff *et al.* reported that there were three times as many transcripts with 5′ cap structures as poly A+-containing transcripts localized with polysomes of Chinese hamster cells (*12*). Later studies revealed that many genes are transcribed as poly A+ RNAs, which under specific conditions are processed to reduce or totally remove the 3′ poly A sequences. Such RNAs are called "bimorphic" transcripts (*13*). The distribution of poly A+ and poly A– transcripts between the nucleus and cytosol is also relatively unexplored.

In this report, we examined approximately 30% of the human genome encoded in 10 human chromosomes (6, 7, 13, 14, 19, 20, 21, 22, X, and Y) and mapped the sites of transcription for poly A+ cytosolic RNA derived from eight cell lines. For one cell line (HepG2), maps were constructed for cytosolic and nuclear poly A– and poly A+ transcripts. The full-length structures of many TUFs have been determined by employing a rapid amplification of cDNA ends (RACE) technique and resolving the RACE products by using high-density arrays. These studies indicate that previously considered "junk" genomic regions encode multiple overlapping poly A+ and poly A– coding transcripts and TUFs.

Overview of sites of transcription of cytosolic poly A+ RNAs along 10 human chromosomes. High-density arrays using 25-mer oligonucleotides spaced every 5 bp on average (i.e., 20-bp overlap) provided an interrogation resolution at least seven times as high as that of previous studies (*3*–*6*, *14*). The consequences of conducting array-based interrogations every 5 bp include increased likelihood of detecting exons of shorter length, increased statistical confidence in determining whether a region is transcribed, and identification of specific hybridization patterns characteristic of 3′ ends of transcripts (fig. S1).

Five male and three female cell lines were selected as sources for mature (i.e., post-spliced) cytosolic poly A+ RNA. Maps were constructed with the lowest likelihood of signals being derived from cross-hybridization. Transfrag sequences that overlapped with pseudogene sequences or contained low-complexity repeat sequences were removed (*15*).

Approximately 9% of 74,180,611 total probe pairs detected transcription per cell line and per chromosome. Average positive probe percentiles for individual chromosomes ranged from 7.1% [chromosome (chr) 13] to 14.6% (chr 19) (table S1A). This number increased to 16.5% for a cumulative map, referred to as a "1 of 8 map" in which a positive probe must appear in at least one of eight cell lines. This is consistent with our previous results from chromosomes 21 and 22 (*3*). The average number of transfrags found per cell line and per chromosome was observed to be 16,864 (table S1B). The average and median lengths of observed transfrags were 115 and 78 nucleotides, respectively (table S1C). The number of transfrags increases to 31,443 in the 1 of 8 map, yet the average length of a transfrag, 124 bp, remains approximately the same.

On average, 18,694,360 nucleotides (4.9% of interrogated genomic nucleotides) are transcribed as cytosolic poly A+ RNA derived from 10 chromosomes of each cell line. In the 1 of 8 map, the number of transcribed cytosolic poly A+ nucleotides increases to 38,656,627 (10.1%). The 2.1-fold difference (4.9% versus 10.1%) indicates that a considerable proportion of the detected transcription is cell-line specific. This observation is consistent with earlier findings (*4*).

Correlation of detected sites of transcription with current genome-wide annotations. Maps created by using poly A+ cytosolic RNA from eight cell lines were compared with annotations from the University of California–Santa Cruz (UCSC) genome browser database (*16*, *17*). We found that 56.7% of the detected cytosolic poly A+ sequences from the 1 of 8 map do not overlap with any well-characterized exon, mRNA, or expressed sequence tag (EST) annotation (Fig. 1). With the exception of chromosomes 13, 19, and Y, the remaining seven chromosomes have similar proportions of assignment

[1] Affymetrix Inc., Santa Clara, CA 95051, USA. [2] Office of Cancer Genomics, National Cancer Institute, Bethesda, MD 20892, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: tom_gingeras@affymetrix.com

of transcribed nucleotides to the unannotated category. Annotation-dense chromosome 19 has a lower proportion of detected unannotated (33.4%) transcripts, with the smallest amount of detected transcribed unannotated sequences in the intergenic regions (13%). Chromosomes 13 and Y have higher proportions (69.5% and 80.3%, respectively) of unannotated transcribed sequences, most of which originate from intergenic regions. In general, the genomic positions of transcribed regions track with gene and exon densities for each of the chromosomes (fig. S2). With the exception of chromosomes Y and 19, 56.8% to 77.7% of detected transcribed regions (1 of 8 composite analysis) are derived from within genes (exons and introns). Overall, 31.8% of detected transcription originates from unannotated intergenic regions. The remainder of the 42% to 49% of the observed unannotated cytosolic poly A+ transcription is derived from the intronic regions within genes (26%) (Fig. 1).

Estimates of the fraction of positive probes interrogating well-characterized exons reveal a tendency toward a bimodal distribution with all probes being “on” (>90%) or “off” (<10%) (fig. S3). Approximately 68% of well-characterized exons (58,984) fall within one of the two peaks. Exons with partial positive probe coverage (i.e., >10% and <90%) may represent alternative exon structures compared with those described in the current annotation collections. Alternatively, the lack of correspondence between some of the exon annotations and detected transfrags may be attributed to inaccuracies in transfrag generation, sequencing errors, or misassembly of the human genome (*18*).

Figure S4 estimates the degree of cell-line-specific transcription by plotting the percentage of total nucleotides within known or novel transfrags against the number of cell lines expressing that transfrag. Two dominant populations of transfrags emerge: those expressed in one or two cell lines and those expressed in all cell lines.

Characterization and structure of transcripts containing unannotated transfrags. A combination of RACE, high-density arrays, and cloning/sequencing techniques was used to characterize transcripts containing unannotated transfrags (*15*) (fig. S5). Of 768 randomly selected unannotated transfrags, 634 (82.6%) yielded a set of 5′ and/or 3′ RACE products (table S2). Of these 768 regions, 438 (57.0%) yielded successful 5′ and 3′ RACE products on at least one genomic strand, and 467 (60.8%) show evidence of transcription on both strands. Thus, approximately 61% of surveyed loci show evidence of overlapping transcription on the positive and negative strands of the genome.

Among the 438 transfrags where 5′ and 3′ RACE were successful, 86 reside in intergenic regions, 145 reside in intronic regions, and 207 are adjacent to exons on either strand. To better understand the structure of putative novel transcripts found by RACE, 661 strand-specific RACE groups derived from the 438 index transfrags were analyzed against annotations of known genes and ESTs (Fig. 2) (*15*). Of these, 547 RACE groups contain transfrags that overlap annotations on the sense or antisense strand, 51 groups reside entirely within the intergenic regions, and 63 groups reside entirely within introns of known genes on the sense or antisense strand. Of the 547 RACE groups overlapping annotations, 118 groups contain transfrags that are nearly identical to annotated exons and, thus, are potential novel isoforms of known genes. The other 429 RACE groups contain transfrags that partially intersect annotations, representing transcripts that overlap with the well-characterized coding transcripts on either the sense or antisense strand. For RACE groups that overlap exons or reside within introns, approximately equivalent numbers were found to be sense or antisense to annotations. Overall, 44% of detected RACE-group transcripts are paired with at least one transcript present on the opposite strand. These results, combined with the entire RACE analysis, provide a consistent picture of overlapping transcription in the human genome (fig. S6).

Reverse transcription polymerase chain reaction (RT-PCR) was conducted on 250 (57%) of the 438 genomic loci that produced 5′ and 3′ RACE products from at least one genomic strand. A total of 217 (87%) regions yielded successful RT-PCR products, and 178 cDNA clones were isolated from 107 of the 217 regions. An example of one intergenic unannotated TUF is depicted in fig. S7.

The average length of the isolated transcripts is 680 nucleotides (range 173 to 4650 nucleotides), which are distributed over a range of 173 to 115,020 nucleotides in the genome. Of the 178 cloned transcripts, 114 (64%) are spliced, with an average of 3.2 exons per transcript and an average exon length of 238 nucleotides. Of the 178 characterized transcripts, 65% have a coding capacity of less than 100 amino acids.

Fifty-four percent of the spliced transcripts use canonical splice sites (GT/AG). Many of the noncanonical splice sites are previously characterized alternative splicing signals. However, a total of 26 (14.6%) of the spliced cDNAs were obtained from antisense transcripts, which are exact reverse complements of sense transcripts. Figure S8 illustrates three transcribed regions with mirror complementary sense and antisense transcripts pairings. Such

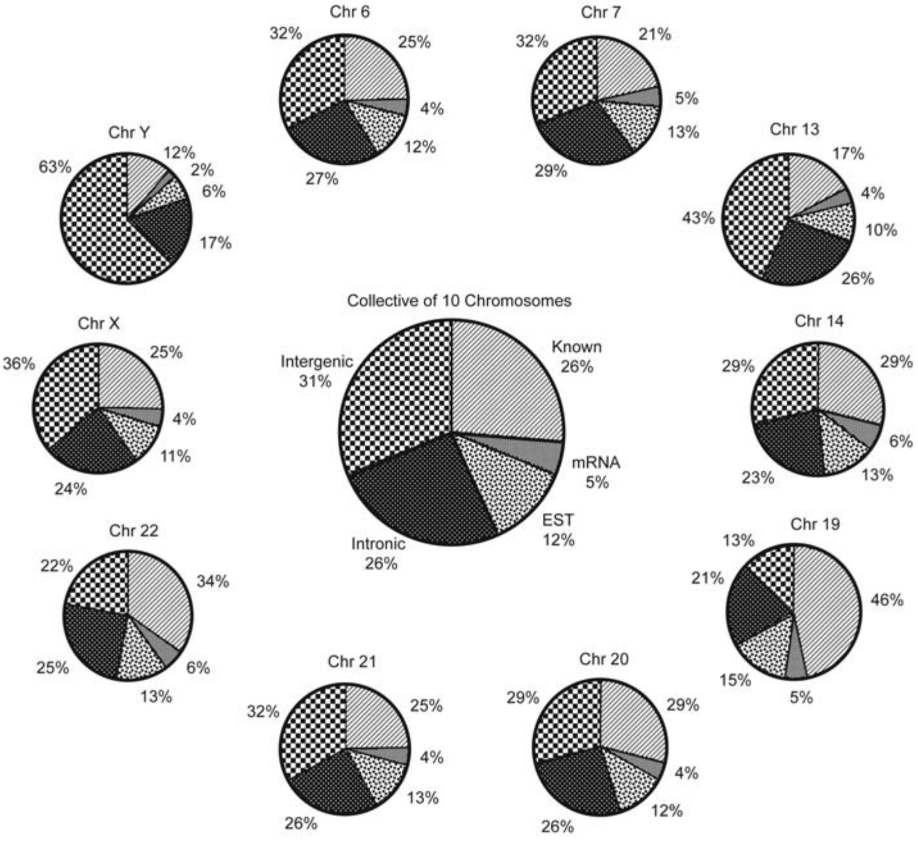

Fig. 1. The correlation of detected transcription in one of eight cell lines to annotations along each of the 10 chromosomes is shown for each chromosome individually and as a collective of all chromosomes. The detected transcription was determined using poly A+ cytosolic RNA from each of the eight cell lines. The annotations used in this correlation are defined in (*15*). The pattern code used in the central pie chart is used in all other pie charts.

complementary transcript pairs have been observed for well-characterized coding genes (*19*).

Poly A+ and A– transcripts and their distribution in the nucleus and cytosol. From the HepG2 cells, 58,874,113 (15.4%) nonredundant nucleotides were detected as transcribed and stable poly A+, poly A–, or bimorphic RNAs isolated from nuclear or cytosolic compartments. All analyses describing the proportions of poly A+, A–, and bimorphic transcribed sequences are based on 58,874,113 bp as a denominator, unless otherwise described. This percentage (15.4%) is nearly an order of magnitude greater than expected from the annotated exons and gene prediction. Fig. S9A illustrates the overlapping and nonoverlapping relationships among the four RNA samples. The comparisons of the four RNA samples result in 15 such relationships, of which 6 represent exclusive nuclear or cytoplasmic and poly A+ and poly A– groupings (Table 1A and fig. S9). The number of transcribed nucleotides and percentage of the total transcribed sequence of the nonrepeat sequences of 10 chromosomes is shown for each of the unique and overlapping poly A+ and A– categories of the relationships (Table 1A). Several of the overlapping relationships (2, 3, 6, 7, 8, 11, 12, 14, and 15) signify that the same detected sequences appear to be bimorphic with respect to the presence of poly A+ and A– sequences (Table 1B). The detected transcribed nucleotides present in poly A+ RNA samples (fig. S9B) and poly A– samples (fig. S9C) and the transfrag sequences found exclusively in the nucleus (fig. S9D) and cytosol (fig. S9E) reveal several characteristics of the composition and compartmentalization of the human transcriptome.

(i) Overall, there are about 2.2 times as many uniquely poly A– (43.7%) transcribed sequences as uniquely poly A+ (19.4%). Thus, 63.1% of the detected transcribed nucleotides are uniquely poly A+ or A– (Table 1B), with 36.9% comprising the bimorphic class of transcripts.

(ii) A large proportion of the sequences found in the nuclear and cytosolic compartments appears to be exclusive to these compartments. The amount of poly A+ sequences (9.7%) exclusively detected in the nucleus is less than one-third the amount of poly A– sequences (31.0%) (Table 1A). Bimorphic detected sequences found exclusively in the nucleus amount to 10.6%. Approximately 25% and 34% of poly A+ nuclear sequences (9.7%) are associated with well-annotated exons and introns, respectively (Fig. 3). The remaining 41% are associated with unannotated intergenic regions of the genome. In total, 75% of the exclusively nuclear-detected poly A+ sequences (9.7%) are unannotated. Similarly, 18% and 57% of poly A– exclusive nuclear sequences (31.0%) are associated with well-characterized exons and introns, respectively, whereas the remaining 25% are located in unannotated intergenic regions of the genome. These data indicate that 82.0% of the exclusive nuclear poly A– sequences are unannotated.

Poly A+ (3.1%) sequences exclusively detected in the cytosol are less than half as abundant as detected poly A– (6.5%) sequences (Table 1A). Bimorphic detected sequences found exclusively in the cytosol amount to 0.6%. About 43% and 22% of poly A+ cytosolic sequences (3.1%) are associated with well-annotated exons and introns, respectively (Fig. 3). The remaining 34% are associated with unannotated intergenic regions of the genome. In total, 56% of the exclusively cytosolic-detected poly A+ sequences (3.1%) are unannotated. We find that 16% and 36% of poly A– exclusive cytosolic sequences (6.5%) are associated with well-characterized exons and introns, respectively, whereas the remaining 48% are located in unannotated intergenic regions of the genome. A total of 84.0% of the exclusive cytosolic poly A– sequences are unannotated.

(iii) A comparison of exclusively nuclear or cytosolic transcribed nucleotides shows a five-

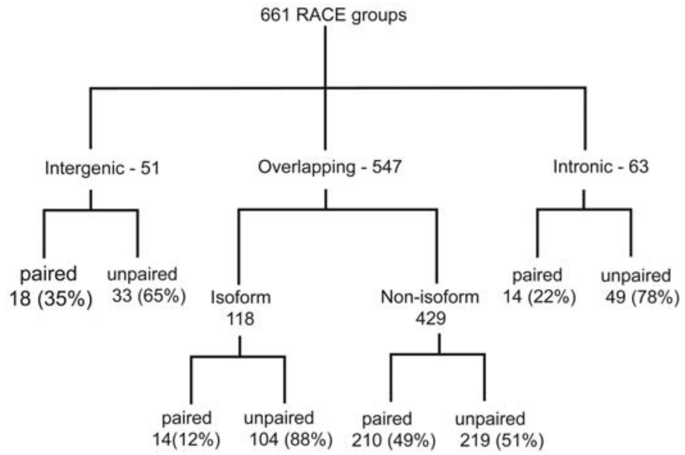

Fig. 2. A hierarchical tree describing the relationship among the RACE/array profiles derived from unannotated, array-detected regions and the annotations. A combined 5′ and 3′ RACE profile from each strand was treated as a separate RACE group. A RACE/array profile for each group is scored "intergenic" if it never overlaps the bounds of a known gene or any other annotation, including an EST; "overlapping" if it overlaps any annotation; or "intronic" if it is confined to the bounds of a known gene but does not overlap an annotation (*15*). "Paired" is defined as overlapping transcripts based on the RACE/array analysis. An overlapping group is further classified into "isoform" or "nonisoform" based on the precision of RACE/array alignment with the exon-intron boundaries of known genes. If a RACE/array profile resembles at least one annotated exon boundary, it is considered an isoform of a known gene.

Table 1. Number and percentage of transcribed nucleotides detected in nucleus and cytoplasm as poly A+ and A– RNA.

A.

Compartment	Description	Nucleotides	% Total
1	Unique to cytosolic A− (CyA−)	3,847,281	6.5%
9	Unique to nuclear A− (NuA−)	18,237,769	31.0%
13	Unique to cytosolic A+ (CyA+)	1,835,709	3.1%
5	Unique to nuclear A+ (NuA+)	5,706,194	9.7%
4	Unique to CyA− and NuA−	3,662,746	6.2%
15	Unique to CyA− and CyA+	349,320	0.6%
12	Unique to NuA− and CyA+	346,798	0.6%
10	Unique to CyA+ and NuA+	3,890,530	6.6%
14	Unique to NuA− and NuA+	6,263,761	10.6%
2	Unique to CyA− and NuA+	417,273	0.7%
8	Unique to CyA− and NuA− and CyA+	431,597	0.7%
3	Unique to CyA− and NuA− and NuA+	1,839,537	3.1%
11	Unique to NuA− and CyA+ and NuA+	3,219,788	5.5%
6	Unique to CyA− and CyA+ and NuA+	1,314,159	2.2%
7	Unique to Cy A− and NuA− and CyA+ and NuA+	7,511,651	12.8%
Grand total	All four compartments combined	58,874,113	100.0%

B.

Compartments	Description	% Total
1+13+15	Sequences detected only in cytoplasmic fraction	10.2%
5+9+14	Sequences detected only in nuclear fraction	51.3%
5+13+10	Sequences detected only in A+ RNA	19.4%
1+9+4	Sequences detected only in A− RNA	43.7%
15+12+14+2+8+3+11+6+7	Sequences detected in both A+ and A− RNA	36.9%

fold difference in sequence complexity detected in the nucleus (51.3%) compared with the cytoplasm (10.3%) of HepG2 cells (fig. S9, D and E). Such a difference is expected given the enrichment of transcribed intron sequences that appear to remain in the nucleus. The Werner (syndrome) helicase-interacting protein 1 (WHIP1) (*20*) on chromosome 6 illustrates how the intronic and exonic sequences of this gene are enriched in the nucleus and cytoplasm, respectively (fig. S10A). Transcribed intronic sequences, however, are not always found to be enriched in the nucleus. Serine (or cysteine) proteinase inhibitor, clade D (heparin cofactor) (*21*), member 1 on chromosome 22 is an example of a gene in which the intronic transcription detected in intron 4 is enriched in the cytosol, although other intron sequences for this gene are enriched in the nucleus (fig. S10B).

On a chromosomal scale, maps identifying the locations of transcription of poly A+ and A– transcripts found in the nucleus and cytosol provide a set of interesting contrasts (Fig. 4). Paired density plots were computed using a 60-kb sliding window for nuclear poly A+ and A– versus cytosolic poly A+ and A– transcribed regions, cytosolic and nuclear poly A+ versus cytosolic and nuclear poly A–, and well-characterized annotations. These maps provide two overriding impressions. First, at the resolution of 60,000 bp, the density of the synthesis sites of poly A+ and A– transcripts found in the nucleus and cytosol generally reflects the annotation density for each of the 10 chromosomes. However, several annotation-dense regions on chromosomes 6, 7, 13, and 21 appear to be more sparsely transcribed in the HepG2 cell line. Second, the annotation densities and detected transcribed regions differ in many positions along each chromosome, which indicates that additional regions of transcription are observed in annotation-dense locations. The chromosomal map positions for poly A+ and A– designated transfrags are shown in table S3.

(iv) The exclusively poly A– and a portion of the bimorphic transcripts found in the nucleus and cytosol would most likely not be identified with customary cDNA cloning approaches. Interestingly, almost half of the exclusive cytosolic poly A– detected transcripts (6.5% of the total detected) and a quarter of the exclusive nuclear poly A– transcripts (31.0% of the total detected) appear to be derived from intergenic regions of the genome. Thus, intergenic noncoding regions of the genome are a rich source of transcripts that are predominantly unannotated and underrepresented in our understanding of the composition of the transcriptome. Evaluation of the protein coding potential of poly A– transcribed sequences awaits efficient methods to copy and clone these types of transcripts.

Conclusions. Recent empirical experiments have provided consistent evidence that a larger percentage of the human, mouse, fly, and *Arabidopsis* genomes are being transcribed than can be accounted for by the current state of genome annotations. These observations were first described in tiling array-based studies that searched large parts or entire genomes for sites of transcription (*3*, *4*, *6*, *22–24*) and then by approaches aimed at isolation and characterization of full-length cDNAs (*25–29*) and of shorter cDNAs (ESTs and serial analysis of gene expression tags) (*19*, *30*, *31*). These studies used primary tissues and cell lines as RNA sources. Collectively, these empirical and computational observations point to several underappreciated characteristics of the human transcriptome.

(i) The human transcriptome is composed of an interlaced network of overlapping transcripts. The use of arrays in combination with 5′ and 3′ RACE reactions indicates that transcripts encoded on both strands often use the same sequences. Such overlapping transcription is observed in almost 50% of the investigated cases (Fig. 2 and figs. S6 and S8). We believe this estimate to be an underrepresentation. Striking examples of this class are pairs of complementary RNA molecules that appear to use both canonical (GT/AG) and complementary to canonical (CT/AC) signals at their splice junctions. The possibility that cDNA clones derived from the complementary transcripts came only from the sense strand was shown to be unlikely, because subsequent strand-specific RT-PCR reactions have produced cDNA products of expected lengths from the noncoding strand.

The existence of such complementary transcripts raises the question of how such transcripts are produced. One possibility is that the human cell-splicing machinery uses complementary sequences as alternative signals. This seems unlikely given the extent to which the consensus signals are missing from the same transcripts. A second possibility is that these transcripts are cRNA copies synthesized by an RNA-dependent RNA polymerase (RdRP). Such activity has been associated with the synthesis of small interfering RNAs that act as trans-acting regulatory molecules in *Arabidopsis* and *C. elegans* (*32*, *33*). Thus, this second possibility predicts that RdRP activities are likely to be found in human cells.

A second implication of the extensive transcription observed in unannotated genomic regions relates to the genotype-phenotype correlations. Such correlation experiments will require extensive analysis of the transcriptional activity of regions mapped as possible loci for genetic mutations.

(ii) Poly A– RNAs potentially comprise almost half of the human transcriptome. A variety of radiolabeling and sequence complexity studies have indicated that, in addition to histone mRNA transcripts, a large class of poly A– transcripts exists in human cells (*11–13*, *34*, *35*). The majority of studies using in vitro translation approaches, however, have not supported the idea of a separate set of protein products derived from the poly A– RNA fraction (*36*). Our results indicate that transcribed sequences exclusively associated with poly A– transcripts are twice as abundant as sequences transcribed exclusively as poly A+.

Of the exclusive poly A– sequences found in nuclear and cytosolic compartments (43.7%

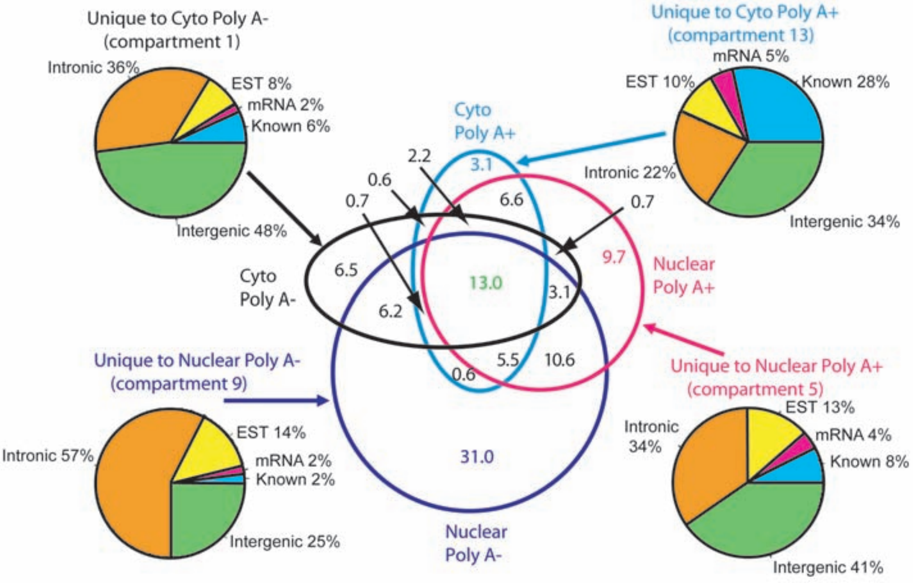

Fig. 3. Distribution of poly A+ and poly A– transcription in the nucleus and cytosol with respect to genome annotations. A four-circle Venn diagram represents proportions of transcribed base pairs in cytosolic poly A+ (cyan), cytosolic poly A– (black), nuclear poly A+ (red), and nuclear poly A– (dark blue). Numbers indicate percentage of total transcription detected in each unique compartment (fig. S9 and Table 1). Pie charts illustrate the distribution of transcribed base pairs detected in each indicated unique compartment among various classes of annotations. The annotations used in this correlation are described in (*15*).

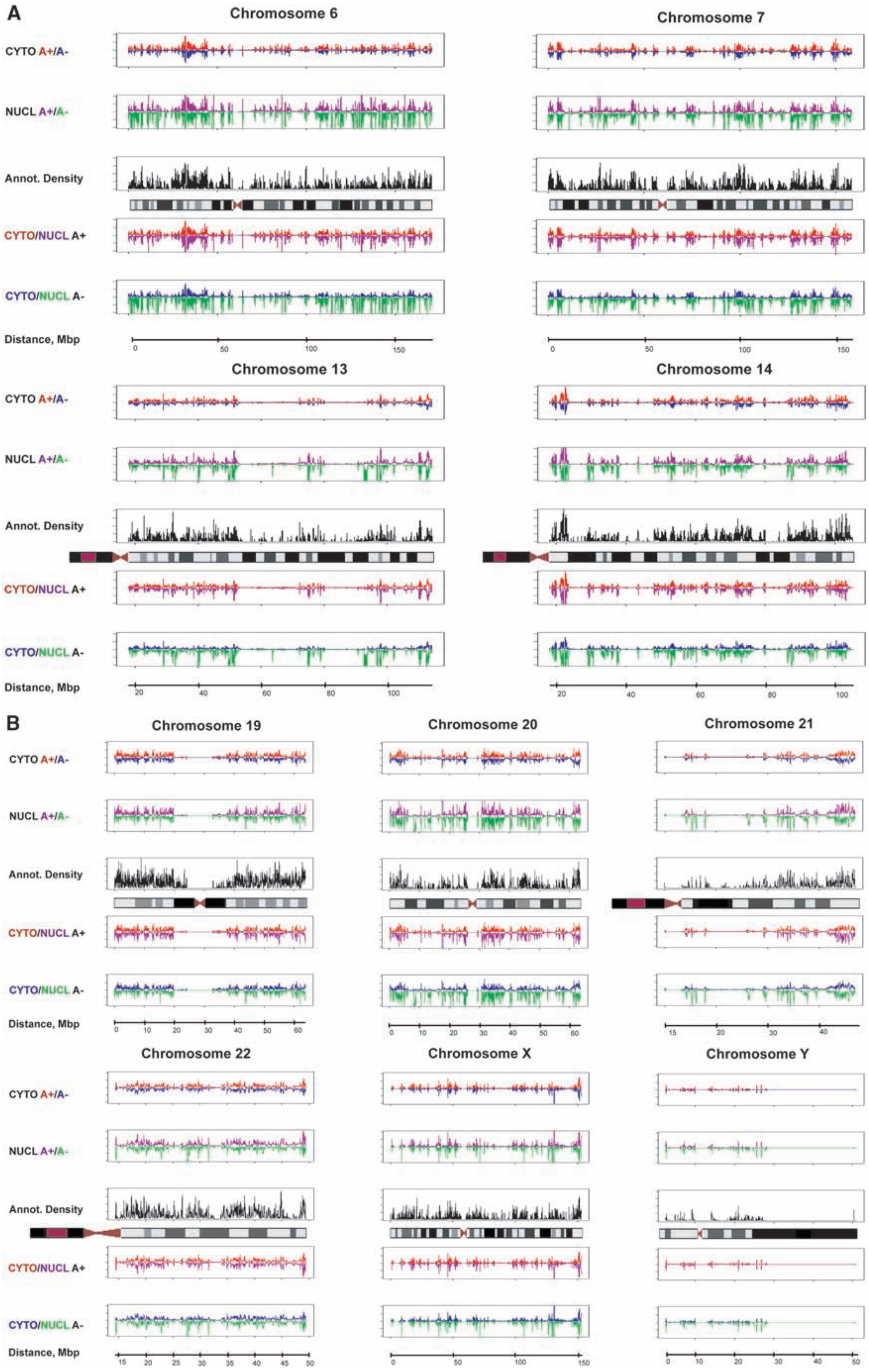

Fig. 4. Density distributions of transcription observed in cytosolic poly A+, poly A–, nuclear poly A+, and poly A– RNA fractions in the 10 human chromosomes. (**A**) Chromosomes 6, 7, 13, and 14. (**B**) Chromosomes 19, 20, 21, 22, X, and Y. The fraction of base pairs found in transfrags is calculated every 6 kb in an overlapping 60-kb window for cytosolic poly A+ (red), poly A– (blue), nuclear poly A+ (mauve), and poly A– RNA (green) and plotted for 10 human chromosomes alongside the base-pair density of exons (black) from Ref Seq, UCSC Known Genes, and GenBank mRNAs. The densities of cytosolic poly A+ versus cytosolic poly A– and nuclear poly A+ versus nuclear poly A– are compared in the top panels, and the densities of cytosolic poly A+ versus nuclear poly A+ and cytosolic poly A– versus nuclear poly A– are compared in the bottom panels for each chromosome.

of all transcription), more than half of nuclear poly A– sequences are derived from intronic regions (Table 1). Clearly, some of these poly A– sequences are introns of spliced coding gene transcripts and may or may not have further biological function once removed from the primary transcripts. However, in the cytosol, the amount of exclusively poly A– sequences is still twice as great as poly A+ sequences (Table 1A), which indicates that there are processed mature poly A– transcripts.

Finally, a total of 36.9% of transcribed sequences are detected as poly A– and poly A+ (Table 1B). These bimorphic sequences are distributed between the two subcellular compartments. It is important to note that detected bimorphic transcribed sequences may be two different transcripts, because transfrags do not identify the strand or specific full-length transcript. However, the presence of such a large proportion of bimorphic transcribed sequences suggests that novel regulatory mechanisms may be involved in the identification of transcripts whose polyadenylation states are altered as a means of regulation. Many of the detected bimorphic sequences are well-characterized coding genes found on the 10 analyzed chromosomes (table S3).

The observations derived from these studies provide some pause as to the state of our understanding concerning where and how the information from the human genome is organized. Many of these and other published observations indicate that our current understanding of the repertoire of transcripts made by the human genome is still evolving. A critical question that applies to both poly A– and poly A+ TUFs centers on the biological functions of these transcripts. Biochemical and genetic experimental approaches are currently being used to answer this question. Until these experiments are completed, systematic identification, mapping, and characterization of as many types of TUFs as possible will assist in understanding and appreciating the complexity of the human transcriptome.

References and Notes

1. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
4. D. Kampa *et al.*, *Genome Res.* **14**, 331 (2004).
5. S. Cawley *et al.*, *Cell* **116**, 499 (2004).
6. J. L. Rinn *et al.*, *Genes Dev.* **17**, 529 (2003).
7. R. Yelin *et al.*, *Nat. Biotechnol.* **21**, 379 (2003).
8. R. Martone *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12247 (2003).
9. The ENCODE Project Consortium, *Science* **306**, 636 (2004).
10. M. L. Birnstiel, M. Busslinger, K. Strub, *Cell* **41**, 349 (1985).
11. C. Milcarek, R. Price, S. Penman, *Cell* **3**, 1 (1974).
12. M. Salditt-Georgieff, M. M. Harpold, M. C. Wilsone, J. E. Darnell, *Mol. Cell. Biol.* **1**, 177 (1981).
13. P. K. Katinakis, A. Slater, R. H. Burdon, *FEBS Lett.* **116**, 1 (1980).
14. P. Bertone *et al.*, *Science* **306**, 2242 (2004).
15. Materials and methods are available as supporting material on *Science* Online.
16. D. Karolchik *et al.*, *Nucleic Acids Res.* **31**, 51 (2003).
17. W. J. Kent *et al.*, *Genome Res.* **12**, 996 (2002).
18. International Human Genome Sequencing Consortium, *Nature* **431**, 931 (2004).
19. J. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12257 (2002).
20. Y. Kawabe *et al.*, *J. Biol. Chem.* **276**, 20364 (2001).
21. R. C. Inhorn, D. M. Tollefsen, *Biochem. Biophys. Res. Commun.* **137**, 431 (1986).
22. D. D. Shoemaker *et al.*, *Nature* **409**, 922 (2001).
23. K. Yamada *et al.*, *Science* **302**, 842 (2003).
24. V. Stolc *et al.*, *Science* **306**, 655 (2004).
25. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
26. T. Ota *et al.*, *Nat. Genet.* **36**, 40 (2004).
27. The Mammalian Gene Collection Project Team, *Genome Res.* **14**, 2121 (2004).
28. T. Imanishi *et al.*, *PLoS Biol.* **2**, 856 (2004).
29. M. Seki *et al.*, *J. Exp. Bot.* **55**, 213 (2004).
30. S. Saha *et al.*, *Nat. Biotechnol.* **20**, 508 (2002).
31. H. Bono *et al.*, *Genome Res.* **13**, 1318 (2003).
32. A. Peragine, M. Yoshikawa, G. Wu, H. L. Albrecht, R. S. Poethig, *Genes Dev.* **18**, 2368 (2004).
33. F. Vazquez *et al.*, *Mol. Cell* **16**, 69 (2004).
34. M. Edmonds, M. G. Caramela, *J. Biol. Chem.* **244**, 1314 (1969).
35. B. J. Snider, M. Morrison-Bogorad, *Brain Res. Brain Res. Rev.* **17**, 263 (1992).
36. T. E. Geoghegan, G. E. Sonenshein, G. Brawerman, *Biochemistry* **17**, 4200 (1978).
37. The authors thank S. Cawley, C. Schaefer, and J. Manak for helpful discussions; M. Mittmann and D. Le for design of photolithographic masks; D. Bartell for software; R. Wheeler for assistance on the annotation database; H. Caley, H. Gorrell, and B. Wong for database support; J. Stevens for administrative support; and K. Kong for manuscript editing and management assistance. All sequenced transcripts have been submitted to GenBank (accession numbers: AY927468 to AY927642). The supplemental materials, feature intensity (CEL) files, graph file, transfrag, and RACE data are available at http://transcriptome.affymetrix.com/publication/transcriptome_10chromosomes and http://cgap.nci.nih.gov/Info/2005.1. Visual representations of the graph and transfrag data are available at http://genome.ucsc.edu/cgi-bin/hgTracks. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-C0-12400, and by Affymetrix, Inc. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1108625/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 and S2
References

13 December 2004; accepted 15 March 2005
Published online 24 March 2005;
10.1126/science.1108625
Include this information when citing this paper.

REPORTS

Wet Electrons at the $H_2O/TiO_2(110)$ Surface

Ken Onda,[1] Bin Li,[1] Jin Zhao,[1] Kenneth D. Jordan,[2]
Jinlong Yang,[3] Hrvoje Petek[1]*

At interfaces of metal oxide and water, partially hydrated or "wet-electron" states represent the lowest energy pathway for electron transfer. We studied the photoinduced electron transfer at the $H_2O/TiO_2(110)$ interface by means of time-resolved two-photon photoemission spectroscopy and electronic structure theory. At ~1-monolayer coverage of water on partially hydroxylated TiO_2 surfaces, we found an unoccupied electronic state 2.4 electron volts above the Fermi level. Density functional theory shows this to be a wet-electron state analogous to that reported in water clusters and which is distinct from hydrated electrons observed on water-covered metal surfaces. The decay of electrons from the wet-electron state to the conduction band of TiO_2 occurs in ≤15 femtoseconds.

The transport of charge through metal-oxide/hydrous phases is crucial to physical and chemical phenomena in many fields of science and technology, including geochemistry, electrochemistry, corrosion, photocatalysis, sensors, and electronic devices (*1*). When exposed to water vapor, metal oxides are partially hydroxylated and covered with up to several monolayers of H_2O. Interactions of the surface acidic metal and basic O ions, respectively, with the O and H atoms of water impose a two-dimensional (2D) order on the hydrated oxide interface (Fig. 1C). In a redox process, electrons must breach this unique 2D environment before they attain fully 3D hydrated Kevan structure proposed for liquid water, in which six tetrahedrally disposed water molecules point one of their H atoms into the excess electron cloud (*2*, *3*). Similar 2D environments, dubbed "wet-electron" states (*4*), in which the "dangling"—i.e., non–hydrogen bonded—H atoms bind and partially hydrate electrons on surfaces of small water clusters, have recently been predicted by theory and discovered in experiments (*5*–*9*). The electronic structure of wet-electron states at the metal-oxide/water interfaces and the dynamics of charge flow that they mediate have not been explored.