Mapping of Small RNAs in the Human ENCODE Regions

Christelle Borel,[1] Maryline Gagnebin,[1] Corinne Gehrig,[1] Evgenia V. Kriventseva,[1,4] Evgeny M. Zdobnov,[1,2,3] and Stylianos E. Antonarakis[1,*]

The elucidation of the largely unknown transcriptome of small RNAs is crucial for the understanding of genome and cellular function. We report here the results of the analysis of small RNAs (< 50 nt) in the ENCODE regions of the human genome. Size-fractionated RNAs from four different cell lines (HepG2, HelaS3, GM06990, SK-N-SH) were mapped with the forward and reverse ENCODE high-density resolution tiling arrays. The top 1% of hybridization signals are termed SmRfrags (Small RNA fragments). Eight percent of SmRfrags overlap the GENCODE genes (CDS), given that the majority map to intergenic regions (34%), intronic regions (53%), and untranslated regions (UTRs) (5%). In addition, 9.6% and 16.8% of SmRfrags in the 5′ UTR regions overlap significantly with His/Pol II/TAF250 binding sites and DNase I Hypersensitive sites, respectively (compared to the 5.3% and 9% expected). Interestingly, 17%–24% (depending on the cell line) of SmRfrags are sense-antisense strand pairs that show evidence of overlapping transcription. Only 3.4% and 7.2% of SmRfrags in intergenic regions overlap transcribed fragments (Txfrags) in HeLa and GM06990 cell lines, respectively. We hypothesized that a fraction of the identified SmRfrags corresponded to microRNAs. We tested by Northern blot a set of 15 high-likelihood predictions of microRNA candidates that overlap with smRfrags and validated three potential microRNAs (~20 nt length). Notably, most of the remaining candidates showed a larger hybridizing band (~100 nt) that could be a microRNA precursor. The small RNA transcriptome is emerging as an important and abundant component of the genome function.

Introduction

A functional annotation of the human genome by use of a combination of experimental and computational approaches is a high-priority effort in the post-sequencing era. The completion of the sequencing of the human and other genomes has enabled efforts to extensively annotate it with the use of a combination of computational and experimental approaches. Recent data from various high-resolution tiling-array approaches surprisingly suggest that the largest part of the human genome is indeed transcribed, and the function of this extensive transcriptional activity is unknown.[1,2] The vast majority of the newly identified transcribed nucleotides (90%) were outside of the annotated regions.

These findings, on RNA molecules with a size above 200 bp, suggest that the majority might not have significant coding capacity (non-protein-coding RNAs [ncRNAs]) and that a considerable fraction of nonpolyadenylated RNAs are not yet annotated.[1–4] The elucidation of the largely unknown transcriptome of small RNAs is of considerable interest, and the number of ncRNAs in the human genome is likely to be much higher and richer than had been anticipated. Recently, Kapranov et al. have presented an extensive genome-wide analysis of small RNAs below 200 nt from the nucleus or cytoplasm of different cell lines.[5] This study provided new insights into the "small transcriptome" and its potential biological role in gene regulation.

In our study, we focused our attention on smaller RNA molecules and investigated the transcription pattern of RNA molecules with a size below 50 bp. This subclass of small RNAs has been the subject of much interest because we know that antisense RNAs are implicated in many aspects of eukaryotic gene expression, including genomic imprinting,[6] RNA interference,[7] CpG island and chromatin remodeling,[8] alternative splicing,[9] X-inactivation,[10] and RNA editing.[11–13]

In this article, we describe hybridization results of the small RNA population (19–50 nt) generated with genomic tiling arrays of the ENCODE pilot regions.[14] We examined both strands of 44 regions representing 1% of the human genome and mapped small RNAs (19–50 nt) derived from four cell lines (HeLaS3, HepG2, GM006990, SK-N-SH with or without retinoic acid). The analysis and interpretation was substantially assisted by the extensive results generated in the pilot phase of the Encyclopaedia of DNA Elements (ENCODE) Project.[15]

Significant hybridization signals were termed SmRfrags (Small RNA Fragments); the data suggest a widespread transcriptional activity in annotated regions as well as outside current annotations. Interestingly, we highlight classes of SmRfrags with a specific genomic localization at the first exon of genes and at the 5′ gene boundaries. Additionally, we observed significant overlap of sense/antisense small transcripts. Moreover, our limited screen identified three new microRNA genes. The transcriptome landscape of small RNAs uncovers previously unknown genomic regions of functional potential and points to additional targets for pathogenic variation in genetic disorders and predisposition to common phenotypes.

[1]Department of Genetic Medicine and Development, University of Geneva Medical School and University Hospitals of Geneva, Geneva 1211, Switzerland; [2]Swiss Institute of Bioinformatics, Geneva 1211, Switzerland; [3]Division of Cell and Molecular Biology, Faculty of Natural Sciences, Imperial College London, London, UK; [4]Department of Structural Biology and Bioinformatics, University of Geneva Medical School, Geneva 1211, Switzerland
*Correspondence: stylianos.antonarakis@medecine.unige.ch
DOI 10.1016/j.ajhg.2008.02.016.

Material and Methods

Cell Culture

Human cell lines GM006690 (human lymphoblastoid, CEPH collection, Coriell Cell Repositories), HeLa S3 (human cervical epithelial carcinoma, ATCC No. CCL-2.2), SK-N-SH (neuroblastoma, ATCC No. HTB-11), and HepG2 (human hepatocellular carcinoma, ATCC No.HTB-8065) were grown in DMEM (HeLa S3, HepG2), RPMI1640 (GM06990), or MEM (SK-N-SH, 1.5 g/L sodium bicarbonate, 0.1 mM nonessential amino acids, and 1.0 mM sodium pyruvate) supplemented with 2 mM L-glutamine, 10% FBS at 37°C and 5% CO2. Differentiation of SK-N-SH cell line was induced by 6 μM all-trans-retinoic acid (Sigma) for 48 hr.

Microarray Hybridization

Total RNA was isolated with TRIZol (Invitrogen) according to the manufacturer's instructions. 100 ug of total RNA from HeLaS3, GM006690, SK-N-SH, HepG2, and SK-N-SH + retinoic acid was size-fractionated (19–40 nt) through a flashPAGE Fractionator (Ambion), precipitated, and concentrated. Small RNAs were prepared with the Mirvana miRNA labeling kit (Ambion). *E. coli* Poly(A) Polymerase and a mixture of unmodified and amine-modified nucleotides were used to add a 20–50 nucleotide tail to the 3′ end of each miRNA in the sample. The amine-modified small RNAs were then purified and coupled to NHS-biotin (Pierce). For each cell line, two ENCODE01 forward arrays and two ENCODE01 reverse arrays (genomic tiling array, Affymetrix, oligonucleotides of 25 mers, 22 bp resolution) were hybridized via Genechip CustomSeq resequencing-array protocol (Affymetrix) at 42°C overnight. The arrays were washed and stained by use of a streptavidine-phycoerythrin (SAPE) conjugate (Molecular Probes, Eugene, OR) according to the manufacturer's directions (DNA ARRAY – WS2-450, Affymetrix). The GeneChips were processed with a GeneArray Scanner (Agilent) by use of the current default settings. DAT image files of the microarrays were generated with Microarray Analysis Suite 5.0 (MAS; Affymetrix).

Data Analysis and Positive Signal Determination

Tiling-array raw data were quantile-normalized within replicate groups. The Affymetrix software GTRANS was used to analyze the intensity of each probe, with 20 bp and 1e-0.05 for the bandwidth and the threshold, respectively. We consider as positive signals oligonucleotides exhibiting fluorescence intensities above the top 99th intensity percentile. These genomic regions were termed SmRfrags (Small RNA fragments). We merged all positive overlapping intervals into single intervals and determined their length. A length equal to 25 nt corresponds to one oligonucleotide (25 nt), a length of 47 nt corresponds to two, a length of 69 nt corresponds to three, and a length below 91 nt corresponds to four consecutive oligonucleotides. The analysis was done independently with data obtained from reverse and forward arrays and with the different cell lines. All raw tiling-array data are available on the AnEuploidy website (see Web Resources).

SmRfrags and GENCODE Annotation

SmRfrags were compared to annotated ENCODE datasets from Galaxy, May 2004 assembly of the human genome sequence (NCBI build 35 or UCSC hg17). They were classified into seven categories via Galaxy (see Web Resources section): "coding sequence (CDS)," consisting of coding exons defined from the GENCODE experimentally verified coding set; "5′ UTR (untranslated region)"; "3′ UTR"; "distal intergenic," which are sequences between genes and greater than 5 kb away from an exon; "proximal intergenic," which denotes sequences between genes and no more than 5 kb away from an exon; "distal intronic," which denotes sequences greater than 5 kb away from an exon; and "proximal intronic," which denotes sequences no more than 5 kb away from an exon. SmRfrags distribution of each cell line and ENCODE array (forward-data array) is evaluated separately for seven categories by the intersection of elements with 1 nt minimum length of overlap.

SmRfrags GENCODE Exons

1211 genes annotated on the coding strand of ENCODE regions were derived from the UCSC public database (Gencode Genes October 2005 track, Gencode Ref/encodeGencodeGeneKnown-Oct05 files, hg17). On the basis of exonic annotation, we made six subgroups of exons depending on their position into the transcripts (first to sixth exons). SmRfrags distribution (forward data array) was evaluated separately for each exon type by intersecting elements of the two datasets with 1 nt minimum length of overlap. The same analysis has been performed with the whole oligonucleotide dataset of the ENCODE array.

Comparison of SmRfrags Distribution with Txfrags Annotation

We derived 4377 (GM06990) and 7254 (HeLaS3) Txfrags of ENCODE regions from the UCSC public database (Affy Transfrags track, EncodeAffyRnaGM06990 or EncodeAffyRnaHela files, hg17). First, HeLa S3-SmRfrags (forward and reverse) were intersected with HeLaS3-Txfrags (20 nt minimum length of overlap). Second, the same analysis was performed with (1) the GM06990-SmRfrags (forward and reverse) versus GM06990-Txfrags datasets and (2) all oligonucleotide datasets that constitute the ENCODE array versus HeLaS3-Txfrags and GM06990-Txfrags. Third, an analysis identical to the first one but with a restricted dataset of SmRfrags corresponding to intergenic distal SmRfrags (described above) was performed.

For the HepG2 cell line, we extracted four different Txfrags datasets: 1) Txfrags cytosolic/poly A(−), 2) Txfrags nuclear/poly A(−), 3) Txfrags cytosolic/poly A(+), 4) Txfrags nuclear/poly A(−) (Table S3). We conducted the same global analysis with HepG2-SmRfrags and intergenic distal HepG2-SmRfrags. Finally, the same global analysis was performed with the whole oligonucleotide dataset of the ENCODE array.

Comparison of SmRfrags Distribution with Functional Elements

The analyzed datasets comprise: 1) consensus set of ENCODE TSS (HPT-TSS 20060421 track); 2) the transcription-initiation sites of genes revealed by ChIP-on-chip experiments against Polymerase II + TAF 250 + histones modifications (HisPolTAF 20060421 track), of which the genome coordinates were downloaded from the Galaxy web page (hg17, May 2004); 3) the HeLaS3 proximal DNase I Hypersensitive sites (DHS within 2.5kb of a TSS); and 4) CpG islands (UCSC track, cpgIslandExt on ENCODE regions). All datasets are produced by the ENCODE working group of the ENCODE consortium.[15] We intersected forward SmRfrag coordinates obtained from different cell lines with datasets of TSS and His.Pol.TAF and DNase I Hypersensitive sites (1 nt minimum length of overlap). The same global analysis was also performed with the whole oligonucleotide dataset of the ENCODE array.

Correlation of Gene Expression with Presence of SmRfrags at TSSs

Files containing gene-expression data for 2480 GENCODE transcripts in GM06990 and HeLaS3 cell lines were uploaded from the UCSC public database (ENCODE transcripts level / Affy RNA signal tracks, AffyRnaGM06990 or AffyRnaHela files, hg17). TSS positions are available, as are the thresholds used to generate an ON/OFF scoring system. We considered as "expressed" those transcripts with an intensity signal superior to 8.375 (GM06990) or 4.5 (HeLaS3). We intersected the HeLaS3 and GM06990 SmRfrags database with TSS position ± 100 nt (20 nt minimum length of overlap), and we subdivided GENCODE transcripts according to the presence or absence of SmRfrags within their TSS.

Cell-Line Specificity

SmRfrags datasets (forward) from different cell lines were intersected with 20 nt minimum length of overlap in order to define the number of common SmRfrags among different cell lines. We used the Galaxy tools available on the web page (see Web Resources). The analysis was performed on the forward data, but there were no identifiable differences observed between SmRfrags distributions on the two strands (data not shown).

Retinoic-Acid Differentiation and SmRfrags in SK-N-SH Cell Line

We compared the signal intensity of SmRfrags in the SK-N-SH cell line before and after retinoic-acid treatment. The ratio of signal intensity for each positive signal between the two datasets was determined. We termed as newly expressed SmRfrags all the signals that were absent in the undifferentiated state and appeared after retinoic-acid differentiation. The procedure was applied for both the forward and reverse datasets.

Sense-Antisense SmRfrags Overlapping

We overlapped forward and reverse datasets of SmRfrags specific for each cell line, with 20 nt minimum length of overlap. We determined the cell-line specificity as explained above (see "Cell-Line Specificity").

MicroRNA Prediction

The prediction approach consisted of the following steps: (1) the human ENCODE sequences that have no gaps in the Mlagan[16] multispecies alignments were scanned with a sliding window of 100 nt for regions capable of folding into a stable stem-loop structure with MFE < -20, total stem length > 25 n, terminal loop < 20 nt, and internal loops < 4 n, with the use of the Lfold procedure from the Vienna RNA package;[17] (2) we filtered the resulting candidates for evidence of evolutionary conservation by overlapping them with binCons[18] regions of the ENCODE multiple alignments; (3) we evaluated the candidates for evidence of RNA secondary-structure conservation in the orthologous sequences of different species extracted from the corresponding regions of the Mlagan ENCODE multiple alignments (considering as representative of closely related primate species, excluding too-divergent sequences of frog and fish species, and discarding any sequences that contain $>$ three consecutive gaps), requiring p value > 0.5 and z score < -2.5 of the RNAz procedure;[19] (4) the remaining candidates were scored for A) having higher-folding free energy than the randomized sequences with 1000 iterations of the Randfold procedure[20] and B) having characteristic conservation profiles similar to Berezikov et al.,[21] and they were identified on the basis of C) the BayesMiRNAfind gene prediction Web Server v1.3[22] and D) overlap with tiling-array expression. The scores were binned to allow the final multifeature sorting on the basis of (4.B), (4.C), number of conserved orthologs in step 3, RNAz p value of step 3, randfold p value of step (4.A), and the covariation score of the RNA secondary-structure conservation. There are only four known microRNA genes in the regions, and three of them were recovered by this approach in the top 40 predictions. The miR-196b, which was predicted on the basis of sequence homology to miR-196a, failed the filter for RNA secondary-structure conservation among the orthologous sequences. We ended up with a list of 95 microRNA predictions. We selected the top 15 for Northern-blot analysis, on the basis of their prediction-score rankings and their tiling-array expressions.

Statistical Analysis

A hypergeometric test was applied for estimation of the statistical significance of SmRfrags distribution along genomic regions in various tissues (Figures 1 and 2). A chi-square test was applied for estimation of the statistical significance of the clustering of SmRfrags along the genome and the correlation of SmRfrags presence with gene expression in various tissues (Figure 3). We used R statistical Language implementation of the tests.

Northern-Blot Analysis

Total RNA was isolated from different cell lines with TRIZol (Invitrogen). From each sample, 20 ug of total RNA was run on 15% polyacrylamide-urea gels, transferred to Genescreen Plus membranes (Perkin Elmer), UV crosslinked, and incubated at 80°C for 1 hr. LNA-oligonucleotides of 25 nt were end-labeled with [γ-^{32}P]ATP and T4 kinase (Ambion). Blots were prehybridized in hybridization buffer (ULTRahyb- oligo buffer, Ambion) for 1 hr at 42°C and hybridized overnight in hybridization buffer containing labeled probe at 42°C. After stringent washes (one wash 30 min at 42°C in 2X SSC 0.5% SDS and two washes 30 min at 42°C in 0.5X SSC 0.5% SDS), membranes were exposed by autoradiography. The sequences of the Northern-blot probes are listed in Table S1.

Results

In order to map small RNAs to precise regions of the human genome, we fractionated the population of small RNAs (19–50 nt) from total RNA of different human cell lines (HeLaS3, GM006690, HepG2, SK-N-SH with or without retinoic acid) (see Material and Methods). These small RNAs were 3′ labeled in duplicate and hybridized to forward and reverse ENCODE tiling arrays; this allows identification of transcribed small RNAs from the forward or reverse strands. The pilot ENCODE sequences represent 1% of the human genome (30 Mbp) included in 44 selected genomic regions.[14] We used the top 1% of oligonucleotides exhibiting fluorescence intensities as a positive hybridization signal, and we conducted the entire analysis with individual hybridization probes. The positive signals are detected at 98% in one to two consecutive oligonucleotides, corresponding to transcripts with a length equal or inferior to 50 nt (see Table S1). We termed the transcripts "identified SmRfrags" (Small RNA fragments). On the basis

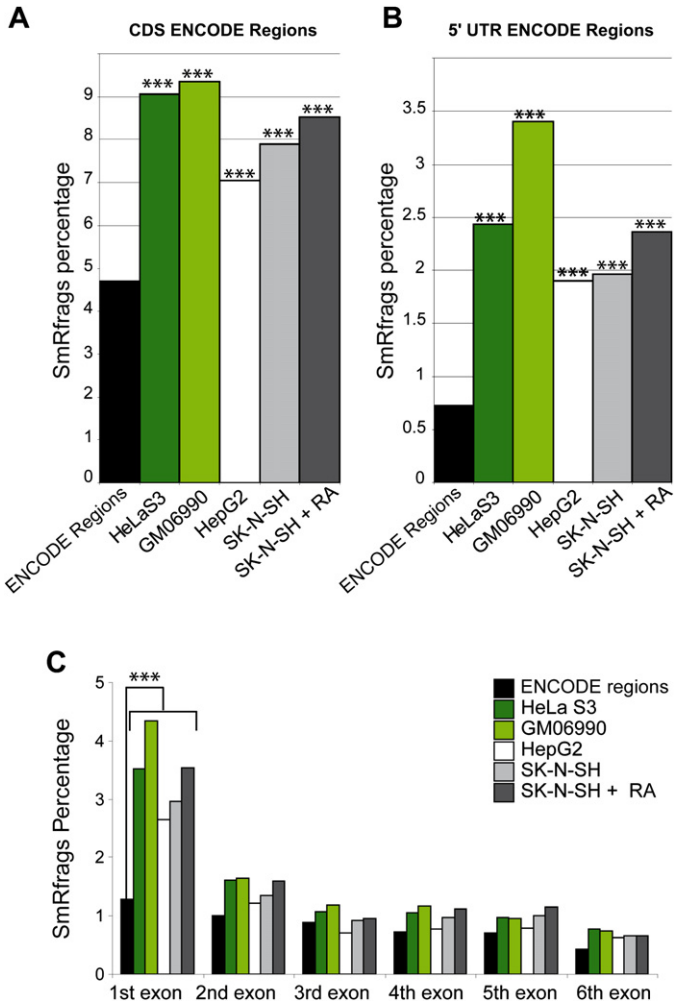

Figure 1. SmRfrags Map in CDS ENCODE Regions, in 5′ UTR ENCODE Regions, and across the First Six Exons of ENCODE Genes in Different Cell Lines

All panels show the percentage of total SmRfrags (CDS ENCODE regions [A], 5′ UTR ENCODE regions [B], across the first six exons of ENCODE genes [C]). "Encode Regions" indicates the ENCODE-array content in each category. p value was determined by Hypergeometric testing. Statistical significance is labeled by *** for p values < 0.005. The analysis was performed on the forward data, but there were no identifiable differences between SmRfrags distributions on the two strands (data not shown). "RA" indicates retinoic-acid treatment (6 μM, 48 hr).

of the top 1% signal threshold, there are 7224 to 8702 positive signals in the different cell lines. This is admittedly arbitrary, and the choice of alternative thresholds reads to different numbers of SmRfrags.

SmRfrags and Annotated Genes

Many SmRfrags (34%) map in nonannotated regions, both in proximal (9.9%) and in distal intergenic regions (24.1%) (Table S2). Particularly notable is the significant enrichment of SmRfrags in CDS sequences (7.1%–9.3%, p value = 3.32^{-21} and 6.31^{-75}, respectively) and 5′ UTRs (1.9%–3.4%, p value = 4.59^{-25} and 8.23^{-103}, respectively) of ENCODE regions, compared with the fraction of those regions in the arrays (4.7% and 0.7%, respectively) (Figures 1A and 1B; Table S2). Moreover, SmRfrags are specifically enriched in first exons. For example, 3.52% of SmRfrags in HelaS3 cells correspond to first exons (p value = 2.97^{-15}), compared to the expected frequency of first exons (1.28%) or other exons (second to sixth exons) (Figure 1C, Table S2).

SmRfrags located outside of known annotations are likely to represent regions for novel noncoding transcripts. In fact, a significant fraction of SmRfrags (21.57% – 25.68% in the various cell lines) localized in distal intergenic regions. To investigate whether these SmRfrags coincide with genes not yet defined, we examined the colocalization of intergenic SmRfrags with Transcribed Fragments (Txfrags), regions identified by unbiased tiling arrays,[3,14,15] Txfrags are transcription sites of poly A+ cytosolic RNA (> 200 nt) derived from several cell lines, two of which are common with our dataset (HeLaS3, GM06990). For one cell line (HepG2), maps were constructed for cytosolic and nuclear poly A(−) and poly A(+) transcripts. Approximately 3.4% (p value = 2.28^{-29}) and 7.2% (p value = 1.99^{-70}) of SmRfrags in intergenic distal regions overlap Txfrags in HeLaS3 and GM06990 cell lines, respectively (compared to the 2.1% [HeLaS3] and 1.1% [GM06990] expected; see Table S3). Thus, SmRfrags partially overlap with Txfrags, providing a validation of additional transcribed regions. From HepG2 cells, 4.59% (p value = 3.06^{-4}) of intergenic distal SmRfrags map to poly A(−) Txfrags exclusively detected in the nucleus (see Table S3). These data suggest a potential biological function, in which long nuclear transcripts could serve as precursors for smaller RNA.[5]

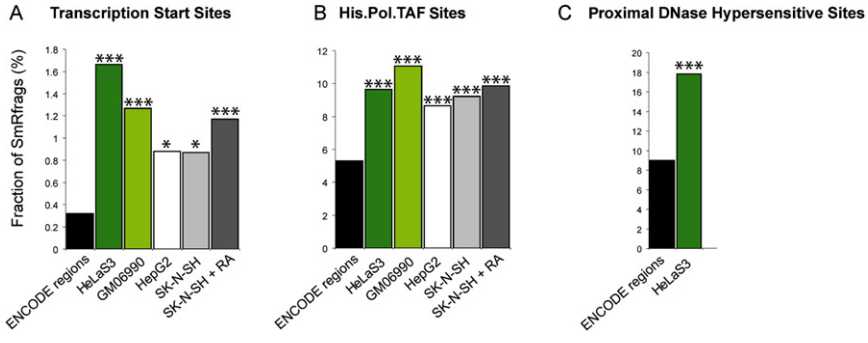

Figure 2. SmRfrags Localization Relative to TSSs, His.Pol.TAF Sites, and Proximal DNase Hypersensitive Sites

The data are shown for SmRfrags mapping in the 5′ UTRs of annotated genes (TSSs [A], His.Pol.TAF sites [B], proximal DNase Hypersensitive sites [C]). Statistical significance is labeled by * for p values < 0.05, by ** for p values < 0.01, and by *** for p values < 0.005, via Hypergeometric test. "A" indicates retinoic-acid treatment (6 μM, 48 hr).

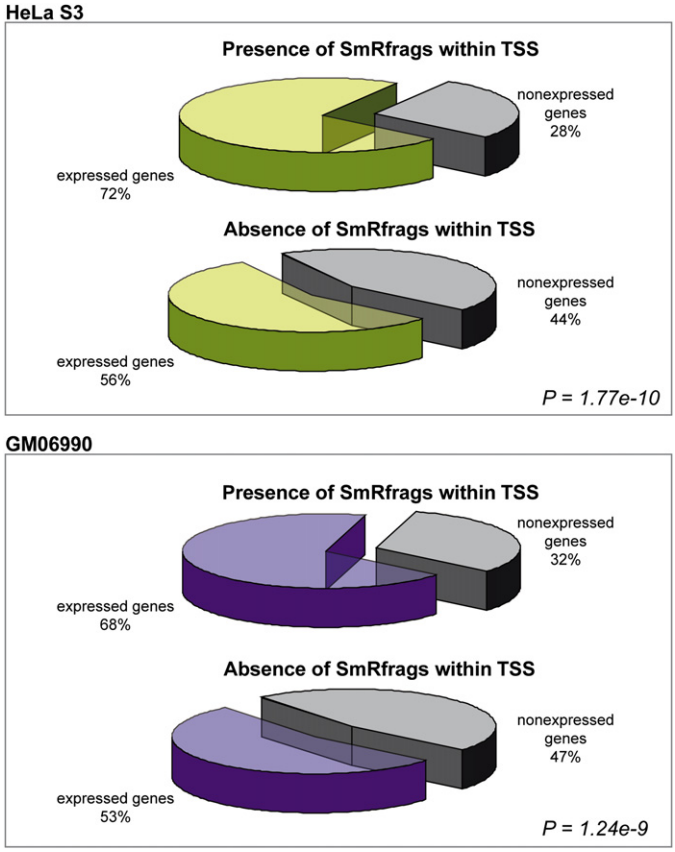

Figure 3. Correlation of Gene Expression and SmRfrags in TSSs

Colors in the pie charts indicate the proportions of expressed genes (green [HeLa S3] and purple [GM06990]) and nonexpressed genes (gray) annotated in ENCODE regions. p value was determined with a Chi-square test.

SmRfrags and Transcription-Initiation Signals

In order to further characterize SmRfrags enriched in the 5′ ends of annotated genes, we compared the mapping position of SmRfrags to transcription start sites (TSS), open-chromatin sites, and other features of functional sequences from the ENCODE study. We overlapped the forward-array-detected signals located in 5′ UTR regions with 1) a consensus set of TSS as established by the ENCODE Project Consortium, 2) the transcription-initiation sites of genes revealed by ChIP-on-chip experiments with Polymerase II and TAF 250 antibodies (250 kDa TATA-box-binding protein [TBP]-associated factor 1), and specific histone modifications (His.Pol.TAF,[15]), 3) the proximal HeLaS3 DNase I Hypersensitive sites (DHS within 2.5kb of a TSS[14]) and 4) CpG islands. Figure 2 shows that on average, 9.69% ($p < 0.05$) and 1.17% ($p < 10^{-7}$) of SmRfrags in the 5′ UTR region overlap significantly with His.Pol.TAF sites and TSS, respectively (compared to the 5.30% and 0.32% expected; see Table S4). These results raise the possibility that SmRfrags located in 5′ UTRs are produced by the initiation of genes in the process of transcription. DNase I hypersensitivity sites mark an altered chromatin structure and are usually associated with functional and regulatory genomic features such as promoters, enhancers, and TSS. Figure 2 shows that on average, 17.8% (p value $= 9.12^{-9}$) of SmRfrags in 5′ UTRs overlap significantly with proximal DHS sites (compared to the 9% expected). Moreover, we found no enrichment of SmRfrags in CpG islands (12.57%, versus 12.44% expected), suggesting that our analysis is not biased to GC-rich probes or regions.

We next analyzed the correlation of SmRfrags in TSS with their transcriptional activity of the respective genes. Data generated and analyzed by the laboratories of Tom Gingeras at Affymetrix and Kevin Struhl at Harvard Medical School are available on the UCSC genome browser for two cell lines, GM06690 and HeLaS3 (see Material and Methods). Among genes that contain SmRfrags within their TSS, we detected 72% (p value $= 1.77^{-10}$, HeLaS3) and 68% (p value $= 1.24^{-9}$, GM06990) that are actively transcribed in the corresponding cell lines (Figure 3). It is evident that the presence of SmRfrags in TSS substantially increases the likelihood of transcription of the respective gene.

Differential Expression of SmRfrags

Figure 4 shows the degree of cell-line-specific transcription of SmRfrags on the basis of the top 1% signal threshold applied. A fraction of SmRfrags is cell-line specific (25.9%–42.8% of total SmRfrags, depending on the cell line) (Figure 4). This observation is consistent with the fact that the cell types used in this study originate from different developmental origins and therefore might have unique expression profiles. Another important fraction of SmRfrags (30%–33% of total SmRfrags) is ubiquitously expressed in all four cell lines (Figure 4). Obviously, this fraction of ubiquitously expressed SmRfrags could diminish with the study of additional cell lines.

We have attempted to determine whether SmRfrag expression might be associated with a biological function. The expression profiles of SmRfrags were monitored during the response of the neuronal cell line SK-N-SH after retinoic-acid treatment (6 μM, 48 hr). The retinoid signal is mediated by the binding of retinoid ligand to the nuclear retinoid receptor (RXRα, β, and γ) protein dimers, which then leads to altered transcriptional activity of target genes.[23] This treatment on SK-N-SH cells promotes significant neuritic outgrowth.[24] Remarkably, approximately 35% of SmRfrags expressed on the forward strand and 25.8% SmRfrags expressed on the reverse strand are induced greater than five-fold in response to retinoic acid. In addition, approximately 32.3% (forward strand) and 28% (reverse strand) of smRfrags are downregulated more than five-fold in response to retinoic acid. We also observed 51.74% (forward strand) and 35.87% (reverse strand) newly expressed SmRfrags in SK-N-SH cells after retinoic-acid treatment. Thus, a considerable number of SmRfrags are regulated in response to retinoic acid, and these dynamic changes suggest their involvement in biological responses.

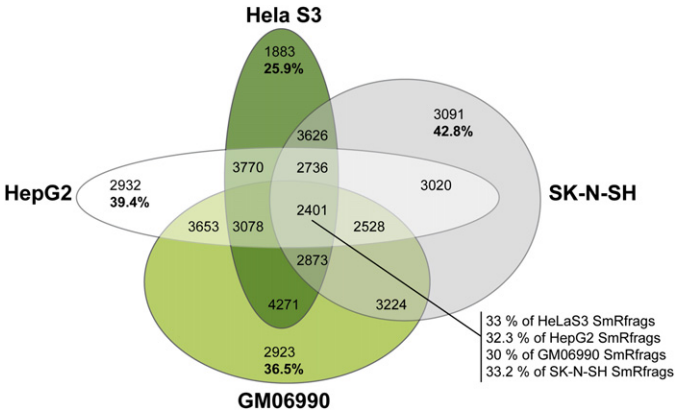

Figure 4. SmRfrags in Different Cell Lines

The fraction of SmRfrags detected in the indicated cell lines is shown. Percentage is expressed as the percentage of total SmRfrags.

Natural Antisense SmRfrags

A growing number of endogenous antisense transcripts have been reported during the last several years in a variety of eukaryotic organisms.[25] They are composed of pairs of RNAs that are transcribed from the opposite strands of DNA at the same genomic locus (*cis*-NATs; *cis*-Natural transcripts) or from a different genomic locus of the sense RNA (*trans*-NATs; *trans*-Natural transcripts). In this report, we analyzed only small *cis*-NATs, and we refered to these loci as sense-antisense pairs.

We set out to analyze the extent of sense-antisense SmRfrags in the human ENCODE regions. To identify transcripts that originate from the same genomic locus, we intersected the forward and reverse array-detected signals from the same cell line. Overlapping of expressed sequences resulted in a total of 1249–1758 sense and antisense pairs (17%–24% in the various cell lines), with 812 (46%–65% of total sense-antisense SmRfrags) common hits among all the cell lines (Figure 5).

This large fraction of small NATs (46%–65%) that are ubiquitously expressed suggests that small NATs are preferentially more involved in "housekeeping" functions or are more necessary for ensuring the basic structural and metabolic requirements of living cells. Interestingly, sense-antisense SmRfrags are not correlated with overlapping annotated natural sense-antisense transcripts (data not shown, based on Refseq ENCODE database).

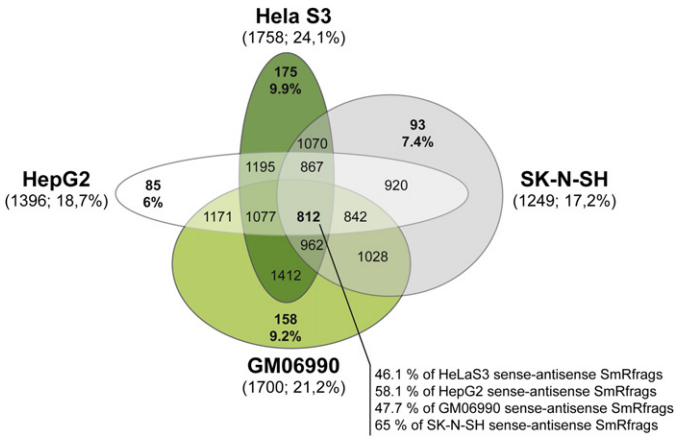

Figure 5. Sense-Antisense SmRfrags, in Different Cell Lines

In brackets, the numbers of sense and antisense pairs are indicated for each cell line. Percentage of total SmRfrags is shown in bold.

Prediction and Verification of Novel MicroRNA Genes

MicroRNAs (miRNAs) ranging in size from 20 to 25 nts represent an important family of ncRNAs that are processed from hairpin precursor transcripts by Dicer.[26,27] The miRNA database (MiRBase) contains 4922 mature miRNA products, among which are 798 human miRNAs (August 2007, release 10.0.). Four known microRNAs map to the ENCODE genomic regions: hsa-miR-192,[28] hsa-miR-194-2,[29] hsa-miR-196b identified by similarity with hsa-miR-196a-1 but not experimentally validated, and hsa-miR-483.[30] They are all detectable with a lower cutoff of SmRfrags (42% lower): miR-196b is expressed in HeLaS3, miR-194-2 in SK-N-SH, miR-192 in SK-N-SH and HeLaS3, and miR-483 exclusively in SK-N-SH after treatment with retinoic acid (2 days, 6 μM) (Figure S1). Interestingly, miR-483 is an intron-derived microRNA of the *IGF2* gene, known to be an effective regulator of cell proliferation. Several reports described the specific upregulation of *IGF2* transcript after treatment with retinoic acid on SKNSH cell lines and more generally on neuroblastoma cell lines.[31–33] These results confirm that our microarray-detection methodology can detect additional putative microRNAs by the use of a lower cutoff of SmRfrags.

To test the hypothesis that some of the SmRfrags are microRNAs, we employed computational analysis to predict candidate microRNAs and then overlapped the predictions with evidence of expression from the tiling-array experiments. We reasoned that this strategy, as opposed to the testing of expressed regions for their likelihood to be microRNA genes, should result in lower rate of false-positive predictions, given that current computational approaches suffer from low specificity.

We tested 15 candidate miRNA predictions in both strands by Northern blot. For three predictions, we detected RNA bands of sizes compatible with precursor (~60–100 nt) and mature miRNAs (~23 nt; Figure 6). The newly identified putative microRNAs have the following Northern-blot probe coordinates (from hg17) and pattern of expression: (1) chr15:41598067-41598092, expressed in GM06990, HeLaS3, and SK-N-SH; (2) chr11:116106598-116106623, expressed in SK-N-SH differentiated with retinoic acid and in HepG2; (3) chr11:64150987-64151012, expressed in HeLaS3 (Figure 6).

Discussion

This study provides an overview of the small transcriptome and estimates the abundance of small RNA molecules (< 50 nt) for 1% of the human genome included in the 30 Mb ENCODE pilot regions from four different cell lines. For each cell line studied, one third of SmRfrags map in intergenic regions. A fraction of SmRfrags in intergenic regions overlap with Txfrags, suggesting that SmRfrags located in nonannotated regions might correspond to novel ncRNAs.

We observed three classes of SmRfrags significantly enriched in preferred genomic loci. First, we observed SmRfrags that cluster in the 5′ UTR of genes and overlap with TSSs, His.Pol.TAF sites, and DNase I hypersensitive sites. Second, we observed SmRfrags that are enriched in first exons of genes. Third, we detected a significant fraction of sense-antisense overlapping expressed SmRfrags.

The biological functions of these SmRfrags could be diverse. SmRfrags in the 5′ UTR of annotated genes are associated with initiation sites of transcription (TSSs, His.Pol.TAF sites, and DNase I hypersensitive sites) and show a strong correlation with the expression of adjacent genes. Transcription initiation is a multistep process and entails two sequential stages delimited by the initiation-elongation transition. Before committing to productive elongation, small transcripts (a length of 14–15 nucleotides) can be produced abortively at the site of initiation of transcription by RNA polymerases.[34] This early stage of transcription is referred to as promoter escape, during which a considerable fraction of the elongating polymerases can eject the nascent chain and recycle to the initiation site. Moreover, this phenomenon is linked to an effective expression of the adjacent genes.[35] Therefore, we presume that the enrichment of SmRfrags upstream of annotated genes is correlated with initiation of transcription. It appears from our analysis that around 70% of annotated genes carrying SmRfrags in TSSs are significantly expressed.

Studies of TSS usage with genome-scale approaches have indicated different classes of promoters that could corroborate our observations.[36,37] Some promoters are described with one distinct TSS located at one specific genomic position (classical TSS, TATA box promoter), whereas the majority consist of closely located TSSs at a distance of around 50–100 bp within the promoter region (TATA-independent transcription).[36,38] Recently, a new class of promoters has been identified within exons.[36,39–41] This last class of exonic promoters could be related to a possible slow-down or pausing of RNApolII elongation within exons and thus could serve to recruit the entire gene to the transcription factories,[36] or it could be related to exonic splicing enhancers.[42] These data are concordant with our observations that SmRfrags are enriched in the 5′ UTR and in the first exons of genes. This widespread alternative TSS usage could contribute to the regulation of gene expression (spatial and temporal) and also to mammalian proteomic complexity.

A recent study described the maps of RNA species less than 200 nt in eight cell lines.[5] This small-RNA-mapping study has detected patterns of enrichment in the 5′ UTR of genes, similar to our study. This class of small RNAs was called PASRs (Promoter-Associated Small RNAs), with lengths of around 26–50 nt, and showed an expression correlated with that of adjacent genes. This study also revealed another class of small RNAs, called TASRs (Termini-Associated sRNAs), at the 3′ boundaries of genes. Interestingly, we did not detect a significant enrichment of SmRfrags in 3′ of annotated genes. This could be explained by the size of the small-RNA population used in the two different studies. P. Kapranov et al. hybridized a small-RNA population with a length less than 200 nt,[5] whereas we analyzed small RNAs with a length less than 50 nt. We presume that TASRs are small RNAs with a length around 50–200 nt. In light of this study, we also analyzed the position of SmRfrags relative to exon junction, and we found no significant pattern of enrichment.

A particularly interesting aspect of the present study is the identification of a novel class of small RNAs representing 17 to 24% (depending of the cell line) of SmRfrags that show evidence of overlapping transcription on both strands. This class was not previously described, because other studies used single strand tiling arrays.[5] Evidence that antisense transcription is a common feature of eukaryotic genomes initially came from the analysis of reverse complementarities between all available human mRNA sequences.[43–46] All of these studies identified human overlapping transcripts called *cis*-Natural Antisense Transcripts. In mammals, the percentage of transcriptional units involved in an overlap ranges from 5% to 29% (based on annotated full-length cDNAs and expressed sequence tags [ESTs]; for review in [12]). The total number might be even greater, considering that information on the complexity of mammalian RNA transcription constantly increases with the introduction of tiling arrays. Interestingly, we also observed *cis*-NATs in the small transcriptome, given that we detected between 1249–1758 sense-antisense pairs of SmRfrags in the ENCODE regions (17%–24%, depending on the cell line). Transcription by RNA polymerase involves both large protein complexes and the unwinding of duplex DNA; it is thus unlikely that

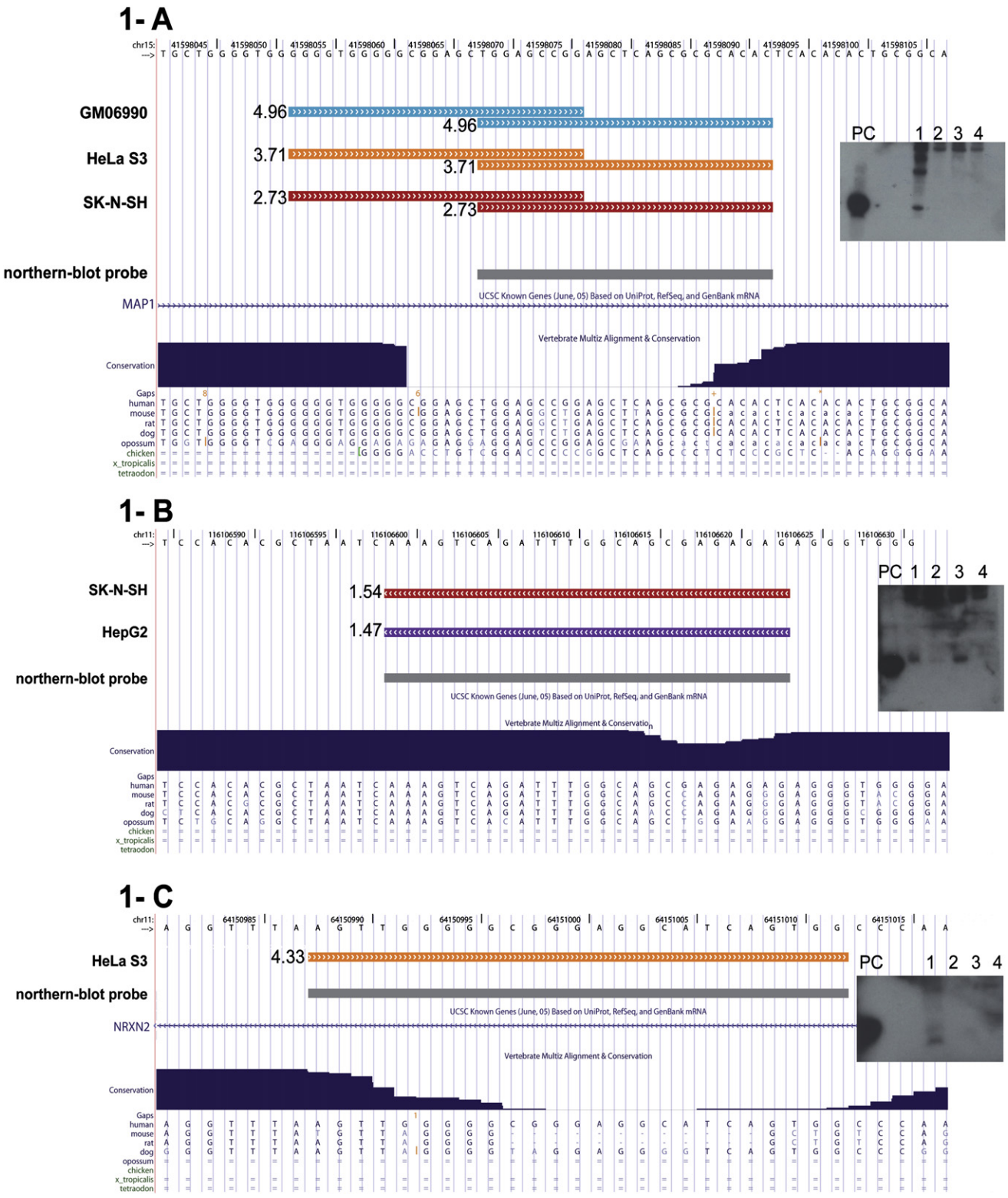

Figure 6. Three Potential New MicroRNAs in the ENCODE Regions

Genomic localization of SmRfrags and Northern-blot analysis: Color bars depict the position of SmRfrags with their respective signal intensities (log2). The cutoff for the top1% positive signals ranges from 2.5 (log2) to 3.5 (log2), depending on the cell lines. Arrows indicate the strand direction. Grey bar represents the Northern-blot probe. The conservation pattern ("conservation" track) is based on the UCSC phastCons scores. This track shows evolutionary conservation in 17 vertebrates, including mammalian, amphibian, bird, and fish species, on the basis of phastCons, a phylogenetic hidden Markov model.[68] Multiz alignments of the assemblies were used to generate this track (generated with UCSC genome browser). The conservation is visualized by a blue scale density gradient and sequence annotation specific for each species.

two overlapping transcriptional units could be transcribed concomitantly.[12,47] The mechanisms responsible for generation of this class of SmRfrags are as yet unknown. We found no evidence for double-stranded or hairpin RNA precursors that could represent intermediates in the biogenesis of such natural antisense SmRfrags.

Several models could be proposed for the regulation of gene expression involving sense-antisense SmRfrags. They might mask *cis*-regulatory elements within transcripts, thereby inhibiting the binding of *trans*-regulatory factors. This steric hindrance could affect any step in gene expression involving protein–RNA interactions. This RNA-masking phenomenon has already been described for *cis*-NATs. For example, alternative splicing of the Rev-ErbAα transcript in B cell lines is inhibited by a short antisense RNA,[48,49] whereas a similar mechanism regulates the human *HFE* gene, which is involved in hereditary hemochromatosis (HH [MIM 235200]).[50]

Furthermore, silencing of *Drosophila* stellate repeats by small sense–antisense RNAs has been well documented.[51,52] as well as a new class of repeat-associated siRNAs (rasiRNAs, 24–29 nt) in the *Drosophila* germline.[53–55] A similar class of germline-specific small RNAs in mammalian cells has been identified on the basis of their specific interaction with mammalian germline-specific Piwi partners piRNAs (26–31 nt).[56–60] However, the biogenesis of piRNAs and their cellular functions remain hypothetical; studies suggested that piRNAs could repress transposition of retrotransposons or be implicated in meiosis.[61]

One of the most well-characterized emerging classes of small ncRNAs are the microRNAs. They were identified over a decade ago in *C. elegans* and are now recognized as a large conserved family of regulatory RNAs 20–25 nucleotides long, which cause posttranscriptional gene repression by base pairing to the mRNAs of protein-coding genes.[27,62] They are implicated in gene-expression regulation in several ways, such as controlling of mRNA stability or translation, promotion of mRNA degradation and turnover, and targeting of epigenetic modifications to specific regions of the genome.[63] To date, thousands of miRNA genes have been identified in animals species, and this number is expected to increase.[64–66] Our microarray approach enabled us to detect all the known microRNAs in ENCODE regions by reducing the stringency of the SmRfrags-detection threshold. In addition, we used Northern-blot analysis to test a set of 15 high-likelihood microRNA predictions that overlap with SmRfrags. This Northern-blot analysis identified three potential miRNA genes that give rise to processed 21–25 nucleotide RNAs; yet most of the remaining candidates show in Northern blot a larger band (~100 nt) that could be a microRNA precursor. A possible explanation would be that the design of the probe is not adequate for detection of a shorter band. Most of the known miRNAs are highly conserved, with > 90% sequence identity between human and mouse.[67] Only two of the newly discovered miRNA genes are conserved to this extent (88% and 92% sequence identity, human-mouse comparison). The third is conserved with only 36% sequence identity between human and mouse. In addition, further study is needed for characterization of the exact sequence of the mature form of the microRNAs. Remaining challenges include identification of the targets of these putative miRNAs and determination of the function of small RNAs.

Our study contributes to the identification of the small-RNA transcriptome and emphasizes the emerging view that the complexity of transcripts is much larger than anticipated. A major challenge for the future will be the elucidation of all the functions and enormous transcription potential of the genome. The newly identified transcripts may harbor pathogenic variation for monogenic and complex genetic phenotypes. Thus, searches for pathogenic genetic variants need to consider these short transcripts as candidate regions.

Supplemental Data

Two figures and five tables can be found with this paper online at http://www.ajhg.org/.

Acknowledgments

We thank P. Descombes and the members of the genomics platform for their assistance. We thank Marc Friedli for comments on the manuscript. This work was funded by the Swiss National Science Foundation, "Frontiers in Genetics" of the National Center for Competence in Research (NCCR), the ChildCare Foundation, the European Union AnEUploidy project, and Fondation Jérôme Lejeune.

Received: November 16, 2007
Revised: January 28, 2008
Accepted: February 26, 2008
Published online: April 3, 2008

Web Resources

The URLs for data presented herein are as follows:

Encyclopedia of DNA Elements (ENCODE) Project, http://www.genome.gov/10005107, http://genome.ucsc.edu/ENCODE/
GALAXY, http://main.g2.bx.psu.edu/
MicroRNA Database (MiRBase), http://microrna.sanger.ac.uk/sequences/index.shtml
UCSC Genome Browser, http://genome.ucsc.edu/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
AnEUploidy, http://www.aneuploidy.eu/ (raw tiling array data)

Northern-blot validation of microRNAs: Probes are 25 mer LNA oligonucleotides (see Material and Methods). Each blot contains a positive control lane (PC), which is a 25-mer oligonucleotide with the complementary sequence used for the probe. Lanes (1), (2), (3), and (4) correspond to total RNA from HeLa S3, GM06990, SK-N-SH, and HepG2, respectively. Two specific bands of 25 and 70 nt were detected, corresponding to the mature and the precursor forms of the putative microRNA, respectively.

References

1. 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*, 1149–1154.
2. 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*, 987–997.
3. 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*, 916–919.
4. 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*, 1559–1563.
5. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.
6. Moore, T., Constancia, M., Zubair, M., Bailleul, B., Feil, R., Sasaki, H., and Reik, W. (1997). Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. Proc. Natl. Acad. Sci. USA *94*, 12509–12514.
7. Caudy, A.A., Ketting, R.F., Hammond, S.M., Denli, A.M., Bathoorn, A.M., Tops, B.B., Silva, J.M., Myers, M.M., Hannon, G.J., and Plasterk, R.H. (2003). A micrococcal nuclease homologue in RNAi effector complexes. Nature *425*, 411–414.
8. Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs, D.R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat. Genet. *34*, 157–165.
9. Munroe, S.H., and Lazar, M.A. (1991). Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. J. Biol. Chem. *266*, 22083–22086.
10. Lee, J.T., Davidow, L.S., and Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. Nat. Genet. *21*, 400–404.
11. Kumar, M., and Carmichael, G.G. (1997). Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. Proc. Natl. Acad. Sci. USA *94*, 3542–3547.
12. Lapidot, M., and Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep. *7*, 1216–1222.
13. Prasanth, K.V., and Spector, D.L. (2007). Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev. *21*, 11–42.
14. Consortium. EP (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–40.
15. Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.
16. Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. *13*, 721–731.
17. 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*, 1383–1390.
18. Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. (2003). Identification and characterization of multi-species conserved sequences. Genome Res. *13*, 2507–2518.
19. Washietl, S., Hofacker, I.L., and Stadler, P.F. (2005). Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA *102*, 2454–2459.
20. Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics *20*, 2911–2917.
21. Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. Cell *120*, 21–24.
22. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C., and Showe, M.K. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics *22*, 1325–1334.
23. Chambon, P. (1996). A decade of molecular biology of retinoic acid receptors. FASEB J. *10*, 940–954.
24. Vu, D., Marin, P., Walzer, C., Cathieni, M.M., Bianchi, E.N., Saidji, F., Leuba, G., Bouras, C., and Savioz, A. (2003). Transcription regulator LMO4 interferes with neuritogenesis in human SH-SY5Y neuroblastoma cells. Brain Res. Mol. Brain Res. *115*, 93–103.
25. Yin, Y., Zhao, Y., Wang, J., Liu, C., Chen, S., Chen, R., and Zhao, H. (2007). antiCODE: a natural sense-antisense transcripts database. BMC Bioinformatics *8*, 319.
26. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
27. Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.
28. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003). Vertebrate microRNA genes. Science *299*, 1540.
29. Michael, M.Z., van Holst, S.M.O.C., Pellekaan, N.G., Young, G.P., and James, R.J. (2003). Reduced accumulation of specific microRNAs in colorectal neoplasia. Mol. Cancer Res. *1*, 882–891.
30. Fu, H., Tie, Y., Xu, C., Zhang, Z., Zhu, J., Shi, Y., Jiang, H., Sun, Z., and Zheng, X. (2005). Identification of human fetal liver miRNAs by a novel method. FEBS Lett. *579*, 3849–3854.
31. Babajko, S., and Binoux, M. (1996). Modulation by retinoic acid of insulin-like growth factor (IGF) and IGF binding protein expression in human SK-N-SH neuroblastoma cells. Eur. J. Endocrinol. *134*, 474–480.
32. Matsumoto, K., Gaetano, C., Daughaday, W.H., and Thiele, C.J. (1992). Retinoic acid regulates insulin-like growth factor II expression in a neuroblastoma cell line. Endocrinology *130*, 3669–3676.
33. Ueno, T., Suita, S., and Zaizen, Y. (1993). Retinoic acid induces insulin-like growth factor II expression in a neuroblastoma cell line. Cancer Lett. *71*, 177–182.
34. Dvir, A. (2002). Promoter escape by RNA polymerase II. Biochim. Biophys. Acta *1577*, 208–223.
35. Wang, X., Spangler, L., and Dvir, A. (2003). Promoter escape by RNA polymerase II. Downstream promoter DNA is required during multiple steps of early transcription. J. Biol. Chem. *278*, 10250–10256.

36. 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*, 626–635.
37. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. *8*, 424–436.
38. Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). Dynamic usage of transcription start sites within core promoters. Genome Biol. *7*, R118.
39. Brodsky, A.S., Meyer, C.A., Swinburne, I.A., Hall, G., Keenan, B.J., Liu, X.S., Fox, E.A., and Silver, P.A. (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. Genome Biol. *6*, R64.
40. Bai, L., Santangelo, T.J., and Wang, M.D. (2006). Single-molecule analysis of RNA polymerase transcription. Annu. Rev. Biophys. Biomol. Struct. *35*, 343–360.
41. Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon tethering in transcription by RNA polymerase II. Mol. Cell *21*, 849–859.
42. Wu, Y., Zhang, Y., and Zhang, J. (2005). Distribution of exonic splicing enhancer elements in human genes. Genomics *86*, 329–336.
43. Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. (2003). Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. *21*, 379–386.
44. Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R.Z., and Rowley, J.D. (2004). Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. *32*, 4812–4820.
45. Sun, M., Hurst, L.D., Carmichael, G.G., and Chen, J. (2006). Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity. Genome Res. *16*, 922–933.
46. Lehner, B., Williams, G., Campbell, R.D., and Sanderson, C.M. (2002). Antisense transcripts in the human genome. Trends Genet. *18*, 63–65.
47. Prescott, E.M., and Proudfoot, N.J. (2002). Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. USA *99*, 8796–8801.
48. Hastings, M.L., Milcarek, C., Martincic, K., Peterson, M.L., and Munroe, S.H. (1997). Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. Nucleic Acids Res. *25*, 4296–4300.
49. Hastings, M.L., Ingle, H.A., Lazar, M.A., and Munroe, S.H. (2000). Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA. J. Biol. Chem. *275*, 11507–11513.
50. Thenie, A.C., Gicquel, I.M., Hardy, S., Ferran, H., Fergelot, P., Le Gall, J.Y., and Mosser, J. (2001). Identification of an endogenous RNA transcribed from the antisense strand of the HFE gene. Hum. Mol. Genet. *10*, 1859–1866.
51. Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr. Biol. *11*, 1017–1027.
52. Aravin, A.A., Klenov, M.S., Vagin, V.V., Bantignies, F., Cavalli, G., and Gvozdev, V.A. (2004). Dissection of a natural RNA silencing process in the Drosophila melanogaster germ line. Mol. Cell. Biol. *24*, 6742–6750.
53. Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev. Cell *5*, 337–350.
54. Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320–324.
55. Saito, K., Nishida, K.M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and Siomi, M.C. (2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. *20*, 2214–2222.
56. Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. *20*, 1732–1743.
57. Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207.
58. Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.
59. Grivna, S.T., Pyhtila, B., and Lin, H. (2006). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc. Natl. Acad. Sci. USA *103*, 13415–13420.
60. Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.
61. Ambros, V., and Chen, X. (2007). The regulation of genes and genomes by small RNAs. Development *134*, 1635–1641.
62. Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.
63. Kloosterman, W.P., and Plasterk, R.H. (2006). The diverse functions of microRNAs in animal development and disease. Dev. Cell *11*, 441–450.
64. Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S., et al. (2006). Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. *16*, 1289–1298.
65. Berezikov, E., Cuppen, E., and Plasterk, R.H. (2006). Approaches to microRNA discovery. Nat. Genet. Suppl. *38*, S2–S7.
66. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203–1217.
67. Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. *22*, 1–5.
68. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.