doi:10.1038/nature11233

SUPPLEMENTAL MATERIALS & METHODS:

TABLE OF CONTENTS

I. PRIMARY DATA GENERATION

Cell Culture and Cell Lines Assayed

Data presented in the manuscript are derived from Tier I cell lines: K562, GM12878, H1-hES cells (H1-hESC); Tier II cell lines: HUVEC, HepG2, Hela-S3; and Tier III cell lines: NHEK, MCF7, AGO4450, SK-N-SH + Retinoic Acid, A549, HSMM, NHLF, HMEC, and BJ. Growth protocols and cell line source information can be found at: http://genome.ucsc.edu/ENCODE/cellTypes.html. Cell culture and RNA isolation for K562, GM12878, HepG2, Hela-S3, HUVEC and NHEK were done in the Gingeras Lab (CSHL). H1-hESC were cultured at Cellular Dynamic and distributed to the entire ENCODE Consortia for analysis. MCF7, AGO4450, SK-N-SH (+ Retinoic Acid), A549, HSMM, NHLF, HMEC, and BJ cell culture and RNA isolation were done by collaborating labs within the ENCODE project. The RNA extracts were distributed amongst the transcriptome groups and interrogated with independent assays, that is, long and small RNA-seq at CSHL, CAGE at RIKEN and RNA-PET at the GIS. Most assays were conducted on identical biological replicates for the majority of samples. Additional sets of long RNA-seq data (polyA+) for independent biological replicates were generated by the Wold Lab (CalTech) from K652, GM12878, H1-hESC, Hela-S3, HepG2, HSMM, HUVEC, NHEK, MCF7, and NHLF cells. All assays, except CalTech long RNA-seq, are directional (stranded).

Sub-cellular Fractionation and RNA Isolation (CSHL)

- *Whole Cell RNA*: RNA from whole cells was isolated using the Qiazol (Qiagen) kit as per the manufacturer's instructions.

- *Nuclear and Cytoplasmic Fractions*: Cells were lysed in RLN buffer (Qiagen). The lysate then was spun at 3,200 rpm for 10 min at 4C to pellet the nuclei. The supernate (cytoplasm) was removed from the nuclear pellet. RLT buffer (Qiagen) was added to the washed nuclear pellet and the sample was homogenized by passing it through an 18½ gauge needle. Six and seven volumes of Qiazol (Qiagen) solution were added to the nuclear and cytoplasmic fractions, respectively. The RNA isolation was subsequently performed according to the manufacturers' directions.

- *Nuclear Subcompartments*: The nuclei extracted from the K562 cell line were further subfractionated into chromatin, nucleoplasm and nucleoli extracts using the protocol outlined in[1].

- *DNAse Digestion*: One-Phor-All buffer (to 1X), 80 Units of RNAse-Free DNAase (Roche) and 80 units of RNAseIn (Promega) were added to 100 ug (or less) of total RNA and incubated at 37C for 30 minutes.

- *Long and Small RNA fractions*: RNAs of >200 nt length are selected using the RNeasy MiniElute Cleanup kit (Qiagen) as per the manufacturers' instructions following DNAse digestions (see above). Simultaneously, RNAs of <200 nt length are selected from the flow-through in the above RNease MiniElute Cleanup (Qiagen). The flow-through containing the small RNAs was combined with 450 ul of 100% ethanol and bond to a fresh MiniElute column. Subsequent purification and elution followed the manufacturers' instructions.

- *Polyadenylated/ Non-polyadenylated RNA:* To fractionate total RNA into polyA+ and polyA- we used the Oligotex (Qiagen) kits as per the manufacturers' instructions. Here the fraction bound on to the column represents the polyA+ sample, the unbound the polyA- sample. To further enrich for non-polyadenylated RNAs the polyA- fraction was passed through another round of Oligotex.

- *Ribosomal RNA Depletion*: To remove ribosomal RNAs (rRNAs) prior to library construction all RNA fractions (long, small, polyA+ and polyA-) were pre-treated with the Ribominus Eukaryotic Kit for RNA-seq (Invitrogen)

- *Subfractionation and RNA Quality Controls*: The RNA extracts were routinely assessed on the BioAnalyzer. We checked to see that the rRNA peaks were intact. The rRNA peaks were also used to assess the efficiencies of cellular subfractionation by looking for the presence of the rRNA precursor in the nuclear samples and its absence in the cytoplasmic compartment.

More detailed methods and quality control figures for these steps are available for download at: http://genome.ucsc.edu/ENCODE/downloads.html.

CSHL Long (>200) RNA-seq Library Preparation

We generated directional (stranded) libraries for Paired End (PE) sequencing on the Illumina platform as described in Parkhomchuck et. al[2].

Briefly, 100 ng of Ribominus, (Invitrogen) treated polyA+ or polyA- RNA with length > 200 nt were mixed with 2 ng of exogenous RNA spike-in pool 14[3]. A mixture of random hexamers and oligo-dT_{21} were used to prime the reverse transcriptase reaction. Entry sites for second strand synthesis catalyzed by *E. coli* DNA Polymerase are generated by means of RNAse H nicks of the DNA:RNA duplex. dTTP is replaced with dUTP during the second strand synthesis. The (ds)cDNA is then sheared through sonification (Covaris). Staggered ends generated during shearing, are repaired and adenylated to prime them for adapter ligation with Illumina Y-adapters. The second strand containing dUTP is eliminated using UNG digestion. The resulting (ss)cDNA is run on an agarose gel and bands with the desired insert sizes of ~ 200 nt are cut out. Cluster compatible sequences are appended in an 18-cycle PCR reaction and the final library is gel purified. We found that libraries from polyA- samples treated with Ribominus (invitrogen) still show between 40-60% of the reads mapping to ribosomal RNAs (data not shown). In order to increase the signal of reads mapping to non-ribosomal RNAs, we applied the DSN Normalization Protocol[4] to all polyA- libraries. All libraries from biological replicates were prepared in parallel to minimize day-to-day variation in the experimental procedure. Finally, the libraries were sequenced on the Illumina GAIIx platform to an average depth of ~100 million mate pairs per sample. In total, we gathered 25,213,363,142 human 76 nt long paired end RNA-seq reads.

CalTech Long RNA-seq Library Preparation

GM12878, K652, HepG2 and HUVEC cells were grown according to the approved ENCODE cell culture protocols. At collection, two independent biological replicates of GM12878 cells (2 x 10^7 or 4 x 10^7), K562 (2 x 10^7 or 4 x 10^7), HepG2 (8 x 10^7) or HUVEC (2 x 10^7) cells were lysed in either 4ml (GM12878 and K562), 12ml (HepG2) or 4ml (HUVEC) of RLT buffer (Qiagen RNEasy kit), and processed on RNEasy midi columns according to the manufacturer's protocol, with the inclusion of the "on-column" DNAse digestion step to remove residual genomic DNA. The two independent replicates of HUVEC cells (2 x 10^7 cells) were frozen in liquid nitrogen as cell pellets and stored at -80°C until lysis with RLT buffer, whereas the GM12878, K562 and HepG2 cells were lysed prior to freezing.

Hela-S3 (6 x 10^7 cells), NHEK (2.8 x 10^7), MCF7 (3.4 x 10^7 and 3.2 x 10^7), and HCT116 cells (2 x 10^7 and 6 x 10^7 cells) were grown according to the approved ENCODE cell culture protocols. At collection, two independent biological replicates of the cells were spun down into compact pellets, flash frozen in liquid nitrogen, and stored at -80°C until lysis. Total RNA was extracted from the cell pellets using the Ambion mirVana mRNA isolation kit (catalog # AM1560), according to the manufacturer's protocol. Residual genomic DNA was removed from the total RNA fraction using the Ambion Turbo DNA-free kit (catalog # AM1907), according to the manufacturer's protocol.

H1-hESC cells were grown according to the approved ENCODE cell culture protocols by an outside vendor (CellµLar Dynamics). At harvest, two independent biological replicates of 5 x 10^6 cells were collected as cell pellets, frozen in liquid

nitrogen and stored at -80°C until lysis. The pellets were split for total RNA isolation with either the Qiagen RNEasy kit or the Ambion mirVana mRNA isolation kit. Residual genomic DNA was removed from the total RNA fraction using either on column DNAse digestion (RNEasy kit) or the Ambion Turbo DNA-free kit.

HSMM and NHLF cells (Lonza) were grown according to the approved ENCODE cell culture protocols. Two independent biological replicates (1 x 10^6 cells each) were processed to extract total RNA using the Qiagen ALLPrep DNA/RNA/Protein Mini Kit (catalog # 80004). No DNAse digestion was performed.

NHEK cells were grown according to the approved ENCODE cell culture protocols, and processed for total RNA extraction using the Qiagen ALLPrep DNA/RNA/Protein Mini Kit and no DNAse digestion (1 x 10^6 cells), or using the Ambion mirVana mRNA isolation kit and Ambion Turbo DNA-free kit (1.4 x 10^7 cells).

mRNA isolation, fragmentation and reverse transcription: Total RNA was selected twice with oligo-dT beads (Dynal) according to the manufacturer's protocol, to isolate mRNA from each of the preparations. After fluorometric quantitation, 100ng of mRNA was processed into double stranded cDNA following the procedure described in Mortazavi et al.[5]. Prior to fragmentation, a 7 µL aliquot (~ 500 pg total mass) containing known concentrations of 7 spiked-in control transcripts from *A. thaliana* and the lambda phage genome were added to a 100 ng aliquot of mRNA from each sample. This mixture was then fragmented to an average length of 200 nt by metal ion/heat catalyzed hydrolysis. The hydrolysis was performed in a 25 µL volume at 94°C for 90 seconds. The 5X hydrolysis buffer components are: 200 mM Tris acetate, pH 8.2, 500 mM potassium acetate and 150 mM magnesium acetate. After removal of hydrolysis ions by G50 Sephadex filtration (USA Scientific catalog # 1415-1602), the fragmented mRNA was random primed with hexamers and reverse-transcribed using the Super Script II cDNA synthesis kit (Invitrogen catalog # 11917010). After second strand synthesis, the cDNA went through end-repair and ligation reactions according to the Illumina ChIP-Seq genomic DNA preparation kit protocol (Illumina catalog # IP102-1001), using the paired end adapters and amplification primers (Illumina Catalog # PE102-1004). Ligation of the adapters adds 94 bases to the length of the cDNA molecules.

Size selection: The cDNA library was size-fractionated on a 2% TAE low melt agarose gel (Lonza catalog # 50080), with a 100bp ladder (Roche catalog # 14703220) run in adjacent lanes. Prior to loading on the gel, the ligated cDNA library was taken over a G50 Sephadex column to remove excess salts that interfere with loading the sample in the wells. After post-staining the gel in ethidium bromide or SYBR Gold (Invitrogen catalog # S11494), a narrow slice (~2 mm) of the cDNA lane centered at either the 300 bp or 500 bp marker was cut. The slice was extracted using the QiaEx II kit (Qiagen catalog # 20021), and the extract was cleaned with either SPRI beads (Beckman Coulter Genomics, catalog #A63881) or Amicon YM-100 filters (Millipore catalog # 42413) to remove DNA fragments shorter than 100bp.

Amplification: One-sixth of the cleaned sample volume was used as template for 15 cycles of amplification using the paired-end primers and amplification reagents supplied with the Illumina ChIP-seq genomic DNA prep kit. The amplified product was then cleaned up over a Qiaquick PCR column (Qiagen catalog # 28104), and then either the SPRI beads or Amicon YM-100 cleanup was repeated, to remove both amplification primers and amplification products shorter than 100 bp. A final pass over a G50 Sephadex column was performed, and the library was quantified using the Qubit fluorometer and PicoGreen quantification reagents (Invitrogen catalog # Q32853). The library was then used to build clusters on the Illumina flow cell according to protocol, and sequenced on the GAII Genome Analyzer (Illumina) as paired end (2 x 76 bp) reads.

Small (<200) RNA-seq Library Preparation

The RNA (<200 nt) was treated with the Ribominus kit (Invitrogen) to eliminate the 5S and 5.8S rRNAs. The RNA was then denatured at 85°C for 2 minutes and placed on ice. A Tobacco Acid Pyrophosphatase reaction was carried out in 1X TAP reaction buffer (Epicenter), 40U of Anti-RNAse (Ambion) and 5U of TAP (Epcenter) for 2 hours at 37°C. Following P/C/I extraction and ETOH precipitation, the RNA was dentured at 85°C for 2 minutes and a 3-prime polyA tailing reaction was carried out in 1X PolyA buffer (Ambion), 250 mM $MnCl_2$, 5mg/ml BSA, 200mM ATP (Roche), 20U Anti-RNAse (Ambion) and 10 Units of E_PAP PolyA Polymerase (Ambion) for 30 minutes at 37°C. Following P/C/I extraction and ETOH precipitation,

the RNA was denatured at 85°C for 2 minutes and the 5-prime RNA linker was ligated on using 20U Anti-RNase (Ambion), 1X T4 ligase buffer (Ambion), 0.5 mg/ml BSA, 7.5U T4 RNA Ligase (Ambion) and 400 uM of adapter: 5'-rArCrArCrUrCrUrUrUrCrCrCrUrArCrArCrGrArCrGrCrUrCrUrUrCrCrGrArUrCrU-3'. The ligation reaction was carried out at 4°C overnight. An additional 5U of T4 RNA ligase was added in the morning and the reaction was further carried out at 25C for 1 hour. Following P/C/I extraction and ETOH precipitation, first strand synthesis was carried out in 20U of Anti-RNAse (Ambion), 15 mM dNTPs (Roche), 1X first strand buffer (Invitrogen), 150 mM DTT, and 200 U Superscript RT III (Invitrogen) (RT Primer: 5'-TCTCGGCATTCCTGCTGAACCGCTCTTC CGATCTTTTTTTTTTTTVN). In order to append cluster compatible adapters a PCR reaction in 1X Phusion (NEB) mix is carried out with 50uM each of the following oligos: PCR 1: 5'-AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATC, PCR 2: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCA TTCCTGCTGAACCGCTCTTC. The libraries are purified on an agarose gel and sequenced on the Illumina GAIIx platform to a depth of 30 million reads in single-end 36 base format.
The reads were mapped using STAR[6].

CAGE Library Preparation

The preparation of the CAGE libraries were adapted from Valen et al.[7] and Takahashi et al.[8] and modified to work with Illumina GA, GAII and GAIIx sequencers.

The cDNA synthesis was performed using 5 to 50 mg of polyA- RNA or 1 to 50 mg of polyA+ RNA and 12 mg RT random primer (RT-N15-EcoP 5'- AAGGTCTATCAGCAGNNNNNNNNNNNNNNC-3') or combination of random primer and oligo-dT primer (5'- AAGGTCTATCAGCAGTTTTTTTTTTTTTTTVN-3') in a 4:1 ratio. The reverse transcription was performed with either M-MLV Reverse Transcriptase RNase H Minus, Point Mutant (Promega) or PrimeScript Reverse Transcriptase (Takara) in presence of 0.132 M trehalose and 0.66 M sorbitol[9] as follow: 30 sec at 25°C, 30 min at 42°C, 10 min at 50°C, 10 min at 56°C, 10 min at 60°C. The M-MLV reaction was stopped with EDTA and proteinase K and subsequently purified with GFX-CTAB[10]. The PrimeScript reaction was directly purified with Agencourt RNAClean XP kit (Beckman).
Subsequently, cDNA reaching the 5' ends (cap sites) was selected with biotinylated cap-trapper method[11]. Briefly, the diol residue of RNA was oxidized using 250 mM NaIO4 for 45 min on ice in the dark. The reaction was stopped with 40% glycerol and 1 M of Tris-HCl pH 8.5. Then, cDNA/RNA hybrids were either purified with GTX-CTAB followed by Microcon YM-100 or Agencourt RNAClean XP kit (Beckman). The capped RNA was biotinylated using 10 mM of biotin dissolved in water in the presence of 1 M sodium citrate pH 6.0 for 12 to 14 hours at room temperature. Then, the biotinylated products were treated with 1 M Tris-HCL pH 8.5, 0.5 M EDTA and RNase ONE Ribonuclease (Promega) for 30 min at 37°C and 5 min at 65°C. The reaction was stopped with 10% SDS, 0.5 M EDTA and proteinase K followed by GTX-CTAB purification and Microcon YM-100 or directly purified using Agencourt RNAClean XP kit. Before cap-trapping, streptavidin sepharose or MPG beads (Takara) were blocked with *E. coli* tRNA for 30 min at room temperature. Sepharose beads were separated by collecting them by centrifugation (12,000 rpm, 10 sec); beads were washed and resuspended in Binding and Wash (B&W) buffer (0.5 M NaCl and 50 mM EDTA). Then, cDNA/RNA hybrids, 5 M NaCl and 0.5 M EDTA were added to the washed beads in a column and incubated with mild agitation at room temperature for 15 min. The cap/bead complexes were collected by centrifugation and sequentially washed with: 1 wash with B&W buffer, 1 wash with B&W buffer containing 0.05% SDS, 2 washes with B&W buffer containing 0.1% Tween20 and 5 to 10 washes with B&W buffer.
The capture full-length cDNAs were eluted from the beads using 4 washes of 50 mM NaOH and incubated at 37°C for 10 minutes. Alternatively, the MPG beads were separated from buffer using a magnetic plate, washed and resuspended in 5 M NaCl and 0.5 M EDTA buffer. The cDNA/RNA hybrids were added to the beads and incubated for 30 min. The beads were sequentially washed with buffers 1 to 4[8]. For releasing the cDNA, the beads were resuspended in 50 mM NaOH and incubated for 10 min at room temperature followed by magnetic separation of beads from the eluted cDNA. The eluted cDNAs were transferred to a tube containing 1M Tris-HCl (pH 7.0) and purified using either GTX-CTAB followed by Microcon YM-100 or Agencourt RNAClean XP kit. The purified single stranded cDNA was then ligated with 5´linkers that contain 3 bp index tag and class III restriction enzyme *Eco*P15I. The 5' linker upper oligonucleotide (N6: 5'-CCACCGACAGGTTCAGAGTTCTACAGXXXCAGCAGNNNNNN Phos -3', GN5: 5'- CCACCGACAGGTTCAGAG

TTCTACAGXXXCAGCAGGNNNNN Phos -3') was mixed in a 1:1 ratio with lower oligonucleotide (5'-Phos CTGCTG XXXCTGTAGAACTCTGAACCTGTCGGTGG-3') for both N6 and GN5 linkers. For some libraries, 10 ng of N6 linker was added to the single stranded cDNA and for other libraries 200 ng of the mixture of N6:GN5 linker with ratio of 1:4 was used. The ligation was performed using Ligation kit (Takara) at 16 oC during overnight incubation. The ligated cDNA was purified with using either GTX-CTAB with 0.55M NaCl followed by Microcon YM-100 or Agencourt AMPure XP kit.
The second strand cDNA was synthesized using 1.25U to 5 La Taq (Takara), 2.5mM MgCl2, 0.4mM of each dNTPs and 10 or 200ng of second-strand primer (5'- Bio CCACCGACAGGTTCAGAGTTCTACAG-3') under following conditions: 3 min at 94°C, 5 min at 42°C, 20 min at 68°C and 2 min at 62°C. The cDNA was purified with using either GTX-CTAB with 0.55M NaCl followed by Microcon YM-100 or Agencourt AMPure XP kit. The double-stranded linker:cDNA complexes were cleaved with 0.1U of *Eco*P15I (NEB) in the presence of 100 mM sinefungin at 37°C for 3 hr. Then 10 mM of $MgCl_2$ was added and reaction was further incubated at 65°C for 20 min. The 3´ linker ligation was performed using 10 to 100ng of 3´ linker (Upper: 5'-Phos NNTCGTATGCCGTCTTCTGCTTG-3', Lower: 5'-CAAGCAGAAGACGGC ATACGA-3') and 1200U of T4 DNA ligase (NEB) at 16 oC during overnight incubation. Then, 5M NaCl and 0.5M EDTA were added and the ligated products were purified with streptavidin sepharose beads previously blocked with tRNA. The beads were collected, washed and mixed with biotynilated CAGE tags at room temperature for 1 hour. The tag/bead complexes were washed as described in cap-trapping step. Alternatively, after 3'ligation, the solution was directly mixed with washed tRNA coated MPG beads and tag/beads were washed as described in cap-trapping step. The CAGE tag/bead complexes were then washed and resuspended in water.
DNA fragments were amplified by bulk PCR (6 to 15 tubes) using 1U of Phusion polymerase (NEB), 0.2 mM dNTPs, 0.5 to 1 mM PCR Forward primer (5'- AATGATACGGCGACCACCGACAGGTTCAGAGTTC-3') and 0.5 to 1 mM PCR Reverse primer (5'-CAAGCAGAAGACGGCATACGA-3') under the following conditions: 30 sec at 98°C and 14 to 20 cycles of 10 sec at 98°C, 10 sec at 60 °C. The optimal cycle number for each sample was previously determined with small scale PCR reactions to investigate how many cycles were necessary to observe amplification. The resulting PCR products were treated with proteinase K, purified by ethanol precipitation and finally resuspended in TE buffer. The PCR products were purified on a 12% PAGE for 3hr by 170V. The appropriate 96-bp band was cut out from the gel, crushed and eluted with 0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA pH 8.0, 0.1% SDS buffer room temperature during overnight. The tags were filtrated with MicroSpin empty columns (Amersham Biosciences) and the flow-through volume was reduced by Microcon YM-10 (Millipore) at 14000 x g for 30 min. The resulting extract was then purified, the DNA was phenol-chloroform extracted, ethanol precipitated and finally resuspended in water. Alternatively, after bulk PCR amplification, the resulting products were treated with 40U Exonuclease I for 1h at 37°C. The CAGE tags were purified with MinElute PCR purification kit (Qiagen). The concentration of final products was measured using Agilent 2100 Bioanalyzer DNA 1000 kit and adjusted to 10 nM.
The libraries were sequenced with Illumina GA, GAII or GAIIx using specific sequencing primer (5'- CGGCGACCACC GACAGGTTCAGAGTTCTACAG -3').

nanoCAGE Library Preparation

For the some of the libraries (K562 polysomal polyA-, K562 nucleus total, K562 nucleoplasm total, K562 chromatin total) we used nanoCAGE modified from the published protocol[12] (see reference for more details).

We first mixed 500 ng of RNA with 2 µl of 0.66 M D-threalose, 3.3 M D-sorbitol**,** 100 µM of template switching oligonucleotide (5'- TAGTCGAACTGAAGGTC TCCAGCArGrGrG), 10 µM of random reverse-transcription primer with a random pentadecamer tail (5'- GTACCAGCAGTAGTCGAACTGAAGGTCTCCTCTN15) and reduced the solution volume to 2 µl in a centrifugal evaporator at room temperature. We then heat-denatured the mixture at 65°C for 10 min and transferred it quickly on an ice-water mix. Reverse transcription was accomplished in a volume of 10 µl with the following components: 1.25× first-strand buffer, 650 µM dNTPs, 1.3 mM DTT, 925 mM betain and 200 units of SuperScript II, and the reaction was incubated at 22°C for 10 min, 50°C for 30 min, 75 °C for 15 min. Finally, we immediately transferred the tube on an ice-water mix.
We performed a small-scale semisuppressive PCR. Every two cycles we collected 10 µl aliquots and analyzed them on 1% agarose gel. For each cDNA preparation, we amplified 2 µl of first strand cDNA in reactions of 100 µl using a mixture containing 100 nM forward PCR primer (5'- TAGTCGAACTGAAGGTCTCCAGC), 100 nM reverse PCR primer (5'-GTACCAGCAGTAGTCGAACTGAAGGTCTCCTCT) and 5 units of ExTaq with the following program: 5 min 95 °C, *25* × (10 s at

95 °C, 15 s at 65 °C, 2 min at 68 °C) and using hot start. We then amplified the remaining volumes of cDNAs (6 µl each) in 4 reactions of 100 µl with the chosen number of cycles. We finally pooled and cleaned the PCR products using CTAB and GE Healthcare GFX purification columns.

We digested 45 µl of cDNA in 100 µl using 100 units of EcoP15I, 100 µM sinefungin, and incubated it at 37 °C for 4 h. We mixed the reactions with 130 µl of TE buffer and purified the low-molecular-weight cleavage products using Microcon YM-100 membranes centrifuged for 15 min at 500*g*. We then refilled the columns with 130 µl of TE buffer and centrifuged them for 15 min at 500*g*. We concentrated the flow-through using Microcon YM-10 centrifuged for 20 min at 12,000*g*. We recovered the low-molecular-weight cleavage products by an additional centrifugation for 3 min at 1,000*g* after flipping the cartridges. We mixed equimolar amounts of the two ligation adaptors oligonucleotides ([up- oligonucleotide: 5'-NN**GTC**CTGTAGAACTCTGAACCTGT'] and [down oligonucleotide: 5'- ACAGGTTCAGAGTTCTACAG**GAC**] were mixed 1:1.), heated them at 95°C and left them to cool to room temperature. We ligated 1 pmol of adaptors to 10 µl of the EcoP15I cleavage products in 20 µl of 0.5x Mighty DNA ligation mixture overnight at 16 °C. We determined the optimal number of cycles (8 cycles) for the ligation product to be amplified by PCR with 5 µM of forward PCR primer (5'-CAAGCAGAAGACGGCATACGATAGTCGAACTGAAGGTCTCCAG), 5 µM of reverse PCR primer (5'- AATGATA CGGCGACCACCGACAGGTTCAGAGTTCTACAG), and 5 units of ExTaq. The program was 5 min 95 °C, *8* × (10 sec at 95°C, 10 sec at 68°C). We performed three PCRs in 100 µl for each sample. We digested the excess of the primers with 5 units of exonuclease I at 37°C for 30 min and then heat inactivated the enzyme at 55°C for 15 min.

We purified the PCR products by electrophoresis on 8% polyacrylamide gel, cut the band corresponding to the expected size (114 bp) and passed it through a syringe to break the structure of polyacrylamide. We extracted the DNA at room temperature by rotation with 800 µl of 1× TE buffer overnight. We centrigfuged the tubes at maximum speed for 10 min and collected the supernatant. We added 600 µl of 1× TE buffer to the microtube and let rotate at room temperature for 1 h. We repeat this step a second time. Then, we pooled all the collected fractions for each sample and eliminated residual traces of polyacrylamide using a Microspin filter. We concentrated 2 ml of filtrate to 100 µl using a Microcon YM-10 column. We estimated the purity and the concentration of the sample using Agilent 2100 Bioanalyzer DNA 1000 kit and adjusted to 10nM. nanoCAGE libraries were sequenced by Illumina Genome Analyzer with the nanoCAGE sequence primer (5'- CGGCGACCACCGACAGGTTCAGAGTTCTACAG).

RNA-PET Library Preparation

RNA-PETs were prepared according to an updated, cloning-free method based on the GIS-PET method Ng. et al.[13]. Briefly, polyA+ RNA was isolated from total RNA.The cap-trapper approach[14] was combined with GsuI-poly(T) oligonucleotides to prepare full-length cDNAs, after which full length cDNAs were methylated to prevent EcoP15I cleavage. The cDNAs were then ligated to special linkers containing flanking EcoP15I sites and circularized in a dilute volume (~0.1 ng DNA/ml). After removal of uncircularized products, the cDNAs were digested with EcoP15I to release Paired-End Tags (PETs) with 27 bp tags. The PETs were then ligated to sequencing adaptors (either ABI SoLiD or Illumina), PCR amplified, and approximately 20-30 million PETs were sequenced by ABI SoLiD or Illumina per library. Signature sequences corresponding to the linkers were identified; PETs without such sequences were discarded. After this, PETs were mapped to the reference genome. The PETs were then analyzed to define transcriptional units, and possible fusion or trans-splicing transcripts.

Full-length cDNA preparation: We isolated total RNA using Trizol (Invitrogen). PolyA+ RNA was isolated from total RNA using a Miltenyi Biotec µMACs mRNA Isolation Kit. PolyA+ RNA was then mixed with a GsuI-poly(T) oligonucleotide (5'-GAGCTAGTTCTGGAGTTTTTTTTTTTTTTTTVN-3') and reverse transcriptases (Superscript II and Superscript III, Invitrogen). Next, we added warm trehalose (Sigma) to the mixture and performed reverse transcription, and then digested all enzymes with Proteinase K (Ambion) and used phenol-chloroform (Ambion) with isopropanol (Sigma) precipitation to purify the RNA/DNA heteroduplexes. Next, we oxidized the diol structures using fresh Sodium Periodate (NaIO4) (Sigma) and biotinylated the ends using fresh biotin hydrazide (Vector Laboratories). After this, we removed free RNAs using RNAseONE (Promega) and bound full-length biotinylated DNA/RNA heteroduplexes to M-280 Streptavidin Dynabeads (Invitrogen). Next, we performed alkaline hydrolytic degradation of bound RNA to release the full-length cDNA strand, and synthesized the second strand of the cDNA by ligating (Takara) Cap-Trapper 5' linkers and performing primer

extension with ExTaq polymerase (Takara). Following this, we performed GsuI digestion (Fermentas) to remove poly A tails and thereby produce 3' terminal ends. We then isolated full-length cDNA by size fractionation, which also functioned to remove excess linkers (Invitrogen). The full-length DNA was then quantified by a Quant-iT Picogreen fluorimetry kit (Invitrogen).

PET extraction: The cDNA was methylated by the EcoP15I enzyme (NEB) to protect the EcoP15I sites from subsequent cleavage by EcoP15I. Next, biotinylated linkers with flanking EcoP15I sites were ligated to the cDNAs with T4 DNA ligase (NEB), following which, T4 DNA polynucleotide kinase (NEB) was used to phosphorylate the cDNAs. After this, linker-ligated cDNAs were circularized in a 5 ml volume with T4 DNA ligase. The DNA was nick-repaired with *E. coli* DNA ligase (NEB) and *E. coli* DNA polymerase I (NEB). Plasmid-Safe DNAse (Epicentre/Illumina) was then used to remove any uncircularized cDNA, and the cDNA was digested with EcoP15I to release 27 bp 5' and 3' tags flanking a linker (Paired-End Tags; PETs). The PETs were then bound to M-280 Streptavidin magnetic beads and ligated to sequencing adaptors followed by nick repair and 20 cycles of PCR amplification with Phusion polymerase (Finnzymes/Thermo Fisher Scientific). The PCR band was then gel-excised from a 6% Tris-Borate-EDTA polyacrylamide gel (Invitrogen) and purified using a gel-crush method. The libraries were then sequenced by either the Illumina Genome Analyzer (one lane of 2x36 bases) or ABI SoLid (1/8th of a slide, at 2x35 bases) using paired read sequencing.

Splice Junction Validation By RT-PCR and 454 Sequencing

The short read data generated on the Illumina platform contains many reads that map in a split manner and appear to flank canonical GU/AG introns not yet annotated but which are reproducibly detected, with IDR values < 0.1. These reads serve as excellent anchoring points from which to derive longer range sequence information to better determine the transcript structure in this region. We sought to gather longer-range 454 data generated from targeted RT-PCR products spanning 3,000 unannotated introns in intergenic space from the HepG2, H1-hESC and HUVEC cell lines. PolyA+ whole cell RNA from HepG2, H1-hESC and HUVEC were separately used as input in oligo-dT primed reverse transcription reactions. PCR primers spanning novel candidate introns were used to amplify up the cDNA in independent 96-well PCRs. Each PCR was run on an agarose gel to check for the presence of a band smaller than the genomic DNA distance, suggestive of splicing. The PCR products derived from HepG2 cDNA were pooled together. The same was done for HUVEC and H1-hESC. The pooled products were run on an agarose gel and a region 150-700 bp in size was excised and gel purified. The purified products were used to make a 454 library, one per cell line. The 454 sequencing was done on the Roche Genome Sequencer FLX. Each of the 3 runs generated ~1 million reads.
More detailed methods, primer sequences, targeted junction sequences and the FASTA file for the 454 reads can be downloaded from the Gene Expression Omnibus (GEO) website under accession number GSE38886 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38886). .

II. PRIMARY DATA PROCESSING, ELEMENT GENERATION AND ASSESSMENT OF REPRODUCIBILITY.

Long RNA-seq Processing and Elements

CSHL

Assessing Technical Reproducibility: Each library was sequenced on 3 lanes of the GAIIx machine to achieve a depth of ~100 million read-pairs. To avoid mixing data from different samples, each lane was mapped independently against the human genome (hg19) using STAR [GRCP024]. Simple exon quantifications were produced using SAMtools and normalized as Fragments per Kilobase per Milion Read-pairs Mapped (FPKM). Spearman correlation was computed for all 3 lanes of data derived from the same library. Only lanes with rho > 0.9 were then merged into a final mapped dataset and

formatted for downstream analysis (.BAM and .bigwig format). Reads mapping to spike-in sequences were co-mapped and parsed into separate .BAM files.

Assessing Sequencing Performance: To verify consistent sequencing performance of the samples, we assess a variety of metrics. (1) To look at overall mapping rates, we compute the proportion of mapped read-pairs, including all that map uniquely and to multiple locations (2-10 loci). We find on average 88.74% of all reads derived from polyA+ libraries and 76.13% of all reads derived from polyA- libraries mapped. The lower mapping rates for polyA- samples presumably reflect their enrichment for rRNAs and other repeat RNAs that may not have been eliminated with Ribominus (Invitrogen) or DSN normalization and map more than 10 times and hence are not recorded. (2) We assess the overall quality of the sequencing data by computing the number of reads that mapped with and without mismatches per library. On average, 70.7% of the mapped reads map without any mismatches. The remaining proportion has on average 2.11 mismatches per mapped read. (3) Prior to mapping, the mapping algorithm STAR, may trim off bases with low quality scores to improve the number of mapped reads. In order to ensure that the majority of the sequences we map are near full-length we compute the average mapped lengths of the reads per library. We find that the average length for mapped reads is 73.46 nt,indicating that the underlying base calls are of high quality throughout the read lengths with little decline at the ends due to sequencing chemistry. (4) Our libraries also contain a set of 96 exogenous RNA spike-ins of defined uninterrupted sequence3. We use these to assess the insert sizes for each library by computing the distance between mates of a pair in reads mapped to spike-ins (0.15-18.13% of the total mapped reads depending on the library). We find the average insert sizes to be 212.31nt (+/- 48.75 nt between the libraries)**.** Non-directional inserts may be generated when dUTP is not incorporated during second strand synthesis or when UNG fails to digest the second strand leading to non-stranded inserts. (5) To assess the overall proportion of non-stranded inserts in the data we compare the proportion of reads that map in a sense vs. antisense orientation to spike-ins for each library. We find that on average 98.78% map to the correct strand

Splice Junctions: Splice junctions were identified by collapsing spliced STAR alignments from the merged biological replicates. The total number of alignments, as well as the number of alignments with non-identical 5' and/or 3' ("staggered alignments") crossing a junction were used as the measure of abundance (expression level). The number of staggered reads per junction was used as signal in the non-parametric IDR calculation and computed on a per replicate basis. Generation of junctions is agnostic to annotations. A Gencode v7 junction is a junction belonging to a Gencode v7 transcript.

Contigs: Contigs represent regions of directional RNA-seq coverage. They are called from merged biological replicates but each contig is scored against individual replicates to facilitate IDR analysis. Gaps in coverage up to 25 bases are allowed. Contigs are strand-specific, but contigs with more than 9 times more antisense then sense signal are filtered as possible artifacts of strand specific library construction. Each contig contains associated values: (1) BPKM, "Bases per Kilobase per Million mapped bases", averaged between the replicates. (2) A non-parametric irreproducible discovery score (npIDR) between the replicates. (3) The total number of mapped bases in the contig in both replicates (sum of wiggle track signal). The IDR method is based on this score. Only contigs with score >0 in both replicates are reported. Generation of contigs is agnostic to annotations.

De novo transcripts: Cufflinks 1.0.315 was used to assemble the transcripts from STAR alignments. Only uniquely mapping non-duplicated alignments crossing GU/AG junctions were utilized. The alignments from two bio-replicates were merged before Cufflinks assembly, and all samples were assembled separately by Cufflinks using the default parameters. The assemblies from poly-A+, poly-A- and total RNA samples were merged separately with Cuffmerge15 into three large transcript super-sets.

CalTech

Read mapping: The last base pair of each read was removed. The resulting 2 x 75bp reads were mapped using TopHat (version 1.0.14) in *de novo* mode against the hg19 genome. The same procedure was applied to polyadenylated RNA-seq data from 16 tissues generated using Illumina HiSeq 2000 as part of the Human Body Map 2 project. The *de novo* discovered splice junctions from all cell lines and tissues (except for LHCN, GM12891, GM12892, and HCT116 which were sequenced later) were combined with the set of splice junctions in the GENCODE v4 annotation to derive an extended set

of junctions. Reads were mapped again using TopHat (version 1.0.14) against the male or female version of the hg19 version of the human genome with the extended set of junctions supplied while keeping the de novo junction discovery option turned on. All subsequent analysis was done on the resulting alignments.

Splice Junctions: For each sample, splice junctions identified by TopHat were given a score corresponding to "staggered fragment" coverage in order to filter out potential PCR duplicates, where the number of staggered fragments is defined as the collapsed set of different 5' and 3' ends (for both reads in the pair) spanning the junction.

De novo transcript models. Cufflinks 1.0.1 was used to assemble transcripts in *de novo* mode from the TopHat alignments**.** Each sample was processed individually**.** The assemblies from all the samples were merged together with Cuffmerge into a large transcript super-set.

CSHL/ CalTech

De novo transcript models: In order to obtain a most sensitive set of Cufflinks transcript models we mergerd the CSHL polyA+ superset with the CalTech superset (since the Caltech protocoll on selected for polyadenylated RNAs). The CSHL polyA- and total RNA supersets were not modified. We then defined transcripts of these super-sets that were either intergenic or antisense to long Gencode v7 transcripts as novel and kept only those for downstream analysis.

Gencode v7 and novel intergenic / antisense element quantification: For each long paired-end (PE) RNA-seq experiment, performed either at CSHL or at CalTech, annotated transcripts and novel intergenic / antisense transcripts were quantified using Flux Capacitor. The program allows individual transcript quantification by attributing the reads according to exon length and splice site support, so called 'read deconvolution' (see http://flux.sammeth.net for more details). To produce high quality quantifications we applied stringent criteria on the input data. Firstly, only reads for which the two mates mapped were taken into account. Secondly only those reads in which the two mates were located on the same chromosome, and for stranded experiments, on the same strand and pointing towards each other, were counted. For each bio-replicate transcript quantification was done individually and normalized by transcript length and sequencing depth, ultimately measured as ´Reads Mapped per Kilobase per Million Mapped Reads´ (RPKM). For downstream analysis we only considered experiments with two bio-replicates. For those transcript quantifications found in each bio-replicate were further assessed for reproducibility using npIDR (see below). Only transcripts expressed in both bio-replicates and with a npIDR value lower than 0.1 were regarded as expressed, those not passing the threshold as not expressed. From the transcript quantifications we derived quantifications for other elements. Exon expression was calculated as the sum of the expression of those transcripts that shared the respective exon. Gene expression was calculated as the sum of the expression of all transcripts belonging to the gene. As for transcripts, exon and gene quantifications were further filtered for reproducibility using an npIDR threshold of 0.1.

Ascertainment of Reproducibility: Non-parametric IDR (npIDR) ascertains reproducibility of the detection of genomic elements (such as splice junctions, exons, transcripts, etc) in RNA-seq experiments with biological replicates, referred to as 1 and 2 below. The elements in each bio-replicate are binned according to their signal, and for all bins the $npIDR_{1in2}$ is calculated as the proportion of elements in each bin in replicate 1 that have exactly zero signal (i.e. not detected) in replicate 2. Similarly, the $npIDR_{2in1}$ is calculated as the proportion of elements in each bin in replicate 2 that have exactly zero signal (i.e. not detected) in replicate 1. The final npIDR value for each bin is defined as the mean of $npIDR_{1in2}$ and $npIDR_{2in1}$.

Small RNA Processing and Elements

Contigs: We generated contig files representing regions of directional small RNA-seq coverage. Prior to contig generation the mapped data were filtered to exclude: reads with mapped lengths less than 16 bases, reads mapping to more than one locus, and mapped reads that contain 5 or more consecutive A's. The contigs are called from merged biological replicates but each contig is scored (reads per million of mapped reads (RPM)) against individual replicates to facilitate IDR analysis. Also, contigs with only a single read are discarded. NpIDR is performed against the RPM values. The npIDR-

read number curve is smoothed by taking the logarithm on both values and then performing polynomial fitting as follows. Smoothed npIDR=exp (log (a·r2+b·r+c)). Only contigs with score > 0 in both replicates are reported. Generation of contigs is agnostic to annotations.

Gencode Exon Quantitation: For every exon annotated in Gencode v7 we computed RPM values per library using simple overlap.

CAGE Processing and Elements

Mapping: Raw CAGE reads were mapped to the genome (hg19) using Delve (personal communication), a probabilistic mapper. In brief, Delve uses a pair hidden Markov model to iteratively map reads to the genome and estimate position dependent error probabilities. After all error probabilities are estimated, individual reads are placed to a single position on the genome where the alignment has the highest probability to be true according to the pHMM model. Phred scaled mapping qualities, reflecting the likelihood of the alignment at a given genome position, are also reported. Reads mapping with a quality of less than 10 (< 90% chance of true) were discarded.

Aggregation of CAGE reads into CAGE clusters: Mapped CAGE reads were clustered using the program paraclu[16]. The output is a hierarchal organization of overlapping clusters delineating very broad regions and sub-clusters focusing on increasingly denser regions of CAGE expression. Given that the length of a nucleosome is approximately 150bp we selected all clusters shorter than 200bp from the set of generated clusters for downstream analysis. We created sets of clusters for each cell line individually and a global set of clusters merging data across all samples. The former is used as a basis for our analysis while the latter was used only to derive a sequence based model for transcriptional start sites (TSS).

TSS predictor: The TSS predictor is a non-supervised classifier based on modeling sequences surrounding CAGE regions via hidden Markov models (HMMs). Two models are trained on all sequences surrounding CAGE clusters. The model architecture is designed to capture sequence motifs of length 2-8 present at a certain distances from the middle of each cluster. During training, the main model uses the number of raw reads observed in each cluster to proportionally weight the corresponding sequence while the background model assumes equal weight for all sequences. The posterior probability of each cluster fitting to the main model is calculated using Bayes' rule. Essentially we are asking whether the sequence contains features that give rise to many CAGE reads within the region.

Normalization and IDR: For each sample we normalized the raw number of reads found in each cluster by the library size and multiplied by one million (tags per million - tpm normalization). The IDR method as described by Qunhua Li et. al.[17] was applied to the normalized expression data.

RNA-PET Processing and Elements

From the PETs, We identified signature sequences corresponding to the adaptors that distinguish the 3' tag from the 5' tag, and define the orientation of the PET. The PETs that did not have signature sequences were processed separately from those that had such sequences. The PETs were mapped to the reference genome (hg19), and multiple-mappings were filtered away. Mapping was performed using Bowtie. PETs where both 5' and 3' tags mapped to the same chromosome, same strand, in correct orientation, and within 1Mbp of each other were defined as concordant PETs, whereas those that did not map in such a fashion were called discordant PETs. Clustering was performed on these PETs, and recurrent discordant PETs were taken to suggest fusion genes or trans-splicing variants.

III. COMPUTATIONAL ANALYSIS

Gencode (v7) and Novel Element Statistics (Figures 1, S4-5 and Tables 1, S3 and S5)

- Tables 1, S3, and S5

As a first overview to the transcriptomic diversity in different RNA fractions and cellular compartments we computed summary statistics for detected (expressed) annotated and novel intergenic / antisense elements after quantification (see above). We computed this separately for each RNA fraction (Tables S3, S5) and combined for the polyA+ and polyA- fractions (Table 1), i.e. counting all elements being present in either one or both fractions. To produce these counts in a non-redundant fashion, we regarded two elements as identical, if they: 1) shared the same coordinates (chromosome, start, end, strand) or 2) in the case of transcripts, had the same identifier. Since the intergenic / antisense elements in polyA+ and polyA- fractions are derived from different supersets (see section II), we assessed identity based on their intron structure (see below).

Subsets:
First, we divided the 493,918 annotated distinct exons, 318,693 splice junctions, 161,999 transcripts and 51,706 genes into three sub-categories: (1) long non-coding (lncRNA), (2) protein coding, and (3) other (i.e. not contained in the first two classes). As defined in18, there are in total 14,880 annotated lncRNA transcripts and 9,277 lncRNA genes. Protein coding transcripts were defined as those protein coding transcripts (N=71,006) belonging to protein coding genes (N=20,687). LncRNA and protein coding distinct exons and splice junctions were defined as distinct exons and splice junctions of lncRNA and protein coding transcripts, respectively. Exon coverage is defined as the proportion of nucleotides in expressed exons covered by long RNA-seq contigs.
Secondly, we subdivided the novel (intergenic / antisense) exons, splice junctions, transcripts and genes into two categories: mono-exonic and multi-exonic, depending on whether they stem from a transcript consisting of one or more than one exon, respectively.

Merging intergenic / antisense polyA+ and polyA- objects: transcripts are defined by their intron structure, i.e. list of intron coordinates. When merging polyA+ and polyA- transcripts, we make the union of the intron structures present in the two fractions, and take as most 5'/3' bp of the resulting transcript, the most 5'/3' of the 5' /3' bp of all individual transcripts included in the merged transcript. PolyA+ and polyA- mono-exonic transcripts, genes and exons that stem from mono-exonic transcripts are merged for each strand separately; multi-exonic exons are exons of multi-exonic transcripts. PolyA+ and polyA- detected genes are merged (separately for each strand) if they overlap by at least 1bp. If one of the initial genes is multi-exonic the resulting gene object is called multi-exonic as well, otherwise it is considered mono-exonic.

- Figure 1.

As for the above tables, we computed summary statistics for detected (expressed) annotated and novel intergenic / antisense elements after quantification in a non-redundant fashion. The cumulative number of elements of both RNA fractions (polyA+ and polyA-), represents all elements detected as expressed in either one or both extracts. For the compartments, the cumulative distribution of elements detected represents all elements found in either one, two or all three extracts.

- Figure S4.

Finally, using contigs, we have analyzed genomic coverage with transcribed elements on the one hand, and the transcriptional activity across the genome stemming from their quantity, on the other hand. As such, the genome was partitioned into four different genomic regions based on the Gencode v7 annotation: exonic, intron, intergenic, gaps. This was done in an unstranded fashion and priority was given to features in a hierarchical order: gaps > exons > introns >

intergenic regions. We analyzed only CSHL and CalTech paired-end long RNA-seq experiments that had 2 bio-replicates for a given RNA fraction, cell line, cell compartment, or combination of those.

- Figure S5.

All CSHL and CalTech paired-end long RNA-seq experiments with 2 bio-replicates represent a total of 14 cell lines. The number of nucleotides contributed by each cell line is computed as the number of distinct *unstranded* (strand information collapsed) nucleotides present in the contigs of all experiments performed in this cell line and with a npIDR value lower than 0.1 (cell line contribution). The nucleotides that are specific to a given cell line are the ones present in the contigs of this cell line but not in the contigs of the 13 others (cell line specific nucleotides). If cell lines are ordered by decreasing number of contributed nucleotides, the cumulative number of nucleotides of a given cell line is computed as the number of distinct unstranded nucleotides present in the set of cell lines seen until this point (cumulative total).

Evidence of Protein Expression in Detected Transcripts (Figure S7, Table S4)

We produced proteomic data for two ENCODE Tier1 cell lines, K562 and GM12878, using state-of-the-art mass spectrometry and generated 998,570 tandem mass (MS/MS) spectra. The details of the cell culture and spectral generation are described in Khatun et al.19. We performed proteo-genomic mapping with 263,171 sequences from novel intergenic / antisense transcript models using our in-house software Peppy (www.peppyresearch.com).

We performed a 3-frame translation of those novel transcript models and constructed a protein database, where each stop codon indicated the end of one protein and the beginning of another. We then enzymatically digested those proteins *in silico* and scored the resulting peptides for each MS/MS spectrum. We used our newly developed scoring algorithm IMP, which is embedded in Peppy, to score each peptide/spectrum match.

E-value thresholds were calculated for a specified false discovery rate (FDR) from a search using a decoy database. The decoy database was generated by reversing the amino acid sequences of the same constructed protein database, a method originally suggested by Elias et al.20. To minimize false positive identifications, we analyzed only proteo-genomic mapping results at a 1% FDR. In addition, since many of the exons of the novel isoforms overlapped with annotated exons, we focused on only the proteomic hits in the novel exons of the novel models. We counted the number of peptide and spectral matches falling inside novel exons for each transcript, then plotted this distribution as a histogram.

Supplementary file for Figure S7 and Table S4:
ftp://genome.crg.es/pub/Encode/Transcriptome_paper/FigureS7_RawInput.txt

K562 Nuclear Subcompartment (Tables S5 and S6)

Gencode
In order to define elements specific to any of the K562 nuclear sub-compartments (nucleolus, chromatin, nucleoplasm), we analyzed the CSHL experiments performed for those with respect to the experiments performed for the polyA+ fraction in the three main cell compartments: cell, nucleus and cytosol (of K562). We excluded the polyA- fraction from the analysis, since the polyA+ RNA fractions of the three main compartments were sequenced roughly to the same depth as the total RNA fractions of the nuclear sub-compartments.

We computed the number of Gencode v7 and novel intergenic / antisense detected features in each K562 nuclear sub-compartment, as well as the subset of those that were unique to this compartment. A given feature (exon, splice junction, transcript or gene) is considered to be unique to a given compartment if it is detected in this compartment but not in any of the 5 others. We created sets of features present in each of the 5 compartments and compared those sets with the set

of features present in the remaining compartments. Two Gencode features are considered identical if they share the same coordinates (chromosome, start, end, strand) or in the case of transcripts, the same identifier. Two novel intergenic / antisense elements are considered identical if they overlap on the same strand or, in the case of multi-exonic transcripts, share the same intron list.

We then performed GO-term enrichment analysis (biological process) on those genes specific to the nucleolus, the chromatin and the nucleoplasm using an in-house program21. Seven GO-terms were found to be enriched in the nucleolus set, 4 in the nucleoplasm set and none for the chromatin set (Bonferroni correction, $p<0.05$) (data not shown).

Cell line specific genes (Figure S10, Table S7)

To identify cell line specific genes, we have counted all expressed protein-coding, non-coding and novel genes found in the whole cell extract of 14 cell lines that pass npIDR ≤ 0.1. Hence, a cell line specific gene is a gene detected in one cell line but not in any of the 13 others.

We then performed GO-analysis on the annotated protein coding cell line specific genes - 14 gene sets in total, were analysed, comprising between 6 and 308 protein coding genes per set per cell line. Table S3 reports the number of enriched or depleted GO terms (biological process) found for each cell line. In general genes specific to a given cell line are enriched in functions related to the known biology of the cell line, such as muscle related functions for HSMM, erythrocyte development and differentiation for K562, and keratinization and epidermal related functions for NHEK (data not shown).

Alternative splicing (Figure 4, S11-12)

NOTE: This part of the analysis was performed without any IDR filtering.

- Isoform usage (Figure 4)

We have computed the transcript usage based on the relative frequencies of the gene's annotated isoforms in a given cell line using the CSHL long RNA-seq data from polyA+ whole cell extract. Although there are genes with up to 65 annotated isoforms we have chosen genes with up to 25 isoforms as representative set.

- Splice site usage (Figure S11)

For each protein coding gene in each cell line, we have computed the number of detected splice junctions, relative expression of the most frequently used splice junction and the Shannon's diversity index on the relative usage of gene's annotated splice junctions. This is done based on the relative frequencies of the gene's annotated isoforms in a given cell line. Let g be a gene with n annotated isoforms with relative frequences p1, … ,pn in a given cell line, the entropy of g

$$H(g) = -\sum_{i=1}^{n} p_i \ln p_i$$

H(g) is computed as . We find that H(g) growths with the number of annotated isoforms and with the eveness of their frequencies. H(g) is zero when there is only one expressed isoform, and it is maximum when all isoforms are equally expressed.

The box plots in Figure S11 display, separately for genes with different number of splice junctions, up to 40 splice junctions per gene (N= 87 genes), the distributions pooled together for all cell lines. As with the expressed transcript isoforms (see above, main document), the distributions for the detected splice sites seem to plateau with increasing annotations per gene. The relative usage of the dominant splice junction reaches a plateau around 10% relative expression but the average entropy is slightly higher when computed on splice site usage than on isoform usage (though also over half the theoretical maximum). Note however that the two distributions are not comparable. In general, the

proportion of detected transcripts will be larger than the proportion of expressed splice forms. Indeed, the most common alternative splice event when comparing two alternative splice junctions is exon skipping[22]. Therefore, restricting the analysis only to the alternative splice event, the usage of only one of the two compared isoforms reduces necessarily isoform usage to ½, but usage of splice junctions is reduced at the most to ½ (if the isoform used is the one skipping the exon).

- **Usage of major isoform across cell lines (Figure S12)**

Here we have calculated how often a transcript appeared as the isoform with the highest relative expression (= major isoform) per gene across all cell lines using CSHL polyA+ whole cell data. We find that there are about 3,300 multi-transcript genes with only one major isoform across all cell lines. Of these, about 1,200 genes have two annotated isoforms and about 600 have three isoforms. The trend seems to roughly follow the function 1/n.

Transcription Start and Termination Sites (Table S9, Figure S13)

- Agreement between PET and RNA-seq (Table S9).

We have investigated the overlap of PET and long RNA-seq data in 4 cell lines based on Gencode annotated and novel intergenic / antisense transcripts. For the long RNA-seq data we have restricted the comparison to the polyA+ fraction (all available cell compartments), since the PET data is derived from to polyadenylated RNA only (in theory). We have calculated the concordance of the two data types using three matrices: 1) PolyA+ expressed Gencode v7 transcripts overlapped by a PET end, 2) PolyA+ expressed novel IA transcripts overlapped by a PET end, and 3) PET clusters representing potentially novel transcripts.

The two first categories of transcripts are then further partitioned into 4 classes for each PET experiment and cumulatively, in this order in case transcripts belong to several categories:

1. Both PET 5' and 3' ends map within 1 million bases of each other on the same strand and that they both map within 150bp of an expressed transcript 5' and 3' end respectively,
2. Only the PET 5' end maps within 150 bp of an expressed transcript 5' end,
3. Only the PET 3' end maps within 150 bp of an expressed transcript 3' end,
4. None of the PET ends map within 150 of an expressed transcript 5' and 3' end, but one of them maps within an expressed transcript.

The total number of transcripts in these 4 classes is also provided together with the percentage it represents over the total number of polyA+ expressed Gencode / novel intergenic / antisense transcripts. A cumulative column is also provided for all 4 PET experiments. PET clusters that do not overlap any polyA+ expressed Gencode or IA transcript are called a potential novel PET transcript. Numbers of potential novel PET transcripts are provided for each PET experiment as well as for all 4 PET experiments. The number of potential novel PET transcripts across the 4 PET experiments was obtained by projecting the PET clusters that did not overlap any polyA+ expressed Gencode or intergenic / antisense transcript onto the genome in a stranded way and allowing a 150bp matching window on both 5' and 3' ends.

Supplementary data files for PET can be found here with a README file:

ftp://genome.crg.es/pub/Encode/Transcriptome_paper/cPET_with_expr_tr_info.tar.gz

- Detection of Transcription Start Sites (TSS) with CAGE, PET and RNA-seq (Figure S13).

In order to compare CAGE and PET defined TSS to RNA-seq TSS in a global way, CAGE clusters and PET 5' ends from all 23 CAGE and 4 PET experiments were projected onto the genome (separately for each strand) which resulted in 82,783 CAGE clusters and 63,864 PET 5' end clusters. The clusters were then separately compared to the 97,778 annotated TSS expressed in the polyA+ long RNA-seq experiment. For calculating our statistic, we selected all CAGE and PET 5' end clusters that fell within a 50bp interval of the respective TSS and computed their distance to the closest TSS.

- Identification of polyadenylation sites.

To find sites of polyadenylation we screened the unmappable reads for trailing As or leading Ts: if 5 or more As or Ts were found, or if 6 out of 7 nucleotides at the end were A or T, we trimmed the As or Ts and remapped the read. If the trimmed read mapped uniquely with up to 2 mismatches, we designated the mapping-site a putative polyA site, and we designated the read itself a polyA read. The putative polyA sites for all datasets were clustered together as in Fu et. al[23]. We accepted as a polyA site those clusters that were supported by two or more polyA reads.

PolyA reads were excluded from the study if stretches of As or Ts were found in the genomic region corresponding to the trimmed part of the read. We also excluded those polyA clusters that mapped to known or novel splice sites. The strand from which the polyA read was transcribed was recovered from whether the polyA read contained leading Ts or trailing As and from what strand it mapped to. If As were trimmed from the polyA read, we assumed the original strand was the same as the strand the read mapped to; if Ts were trimmed from the polyA read, we assumed it was transcribed from the opposite strand, since this read fragment had been reverse-transcribed. In this manner we were able to recover the strand of the polyA reads and therefore also the strand of the polyA sites.

For each polyA site we searched 40 bp downstream for one of the canonical polyadenylation signals (AATAAA or ATTAAA), or one of their top 11 variants as found in Beaudoing et. al.[24]. We considered polyA sites associated with PET or novel Intergenic / antisense transcripts if they clustered within 100pb of PET clusters or the 3' ends of intergenic / antisense transcript models.

- Alternative 3' polyadenylation usage.

In this section we set to investigate the differences in the relative usage of the 3' UTR forms in the different cell lines from where cytosol and nuclear polyA+ RNA-seq data is available. We followed an approach similar to the one used in Sandberg et al, and Ji and Tian[25]. For genes with Tandem Alternative Polyadenation, i.e. where alternative 3' UTR forms are found contiguously in the same 3' UTR exon, we compared the usage of a proximal (closest to the stop codon) form with relation to a distal (extended) form. The proximal segment is defined by the region from the stop codon to the proximal polyadenylation site and the distal segment from this point to the longest annotated 3' UTR of that gene. For each segment we calculate the respective RPKM, and then obtain the normalized expression as the ratio of the sum of the expression values in the two forms:

$$NE_{proximal} = \frac{RPKM_{proximal}}{RPKM_{proximal} + RPKM_{distal}}$$

where $NE_{proxinal} + NE_{distal} = 1$

We selected only 3' UTR exons that do not overlap with any other annotated region, with proximal and distal segments that are at least 250bp long. The proximal form is defined by clusters of polyA reads previously defined (see above section); the distal form corresponds to the segment defined from the end of the proximal form to the longest Gencode annotated form.

This yielded a set of 1938 3'UTR exons. A comparison of the 3'UTRs with minimum expression (RPKM >=1) shows a change (difference in the normalized expression larger than 10%) between the cytoplasm and the nucleus in 7.5% to 45.6% of the genes in the seven cell lines. H1-hESC is the cell line with the largest number of changing genes (885). With the exception of HEPG2 and H1-hESC, the direction of change in the other five cell lines is towards a shortening in the cytoplasm. The

percentage of changing genes with larger usage of a proximal polyadenylation site in the cytoplasm is 60.2% (out of 214) for GM12878, 57.2% (out of 217) for K562, 55.0% (out of 261) for Hela-S3, 69.9% (out of 143) for HUVEC, 45.0% (out of 294) for HepG2, 37.9% (out of 885) for H1-hESC and 82.1% (out of 157) for NHEK. These numbers indicate that alternative polyadenylation is an important and active mechanism inducing differential polyadenylation site usage between the nucleus and the cytoplasm.

Annotated Short RNAs (Table 2A, Figures S15, S17 and S18)

- Summary statistics (Table 2A).

  For each main category of Gencode v7 small genes, meaning miRNA, snoRNA, snRNA, tRNA, other small non pseudogene RNA and the total set of those, we calculated the following statistics (Tab. 2A):

  1. Gencode total: number of Gencode genes belonging to this category,
  2. Detected genes (% detected): number and proportion of Gencode genes belonging to this category that is actually detected by a short RNA-seq experiment. Here a gene is called detected if there is at least one short RNA-seq experiment with two bio-replicates for which this gene has a npIDR value smaller than 0.1 (using npIDR on RPM),
  3. Number of genes expressed in only 1 cell line (% detected): number (and proportion over number of detected genes) of genes belonging to this category that is detected in only 1 cell line,
  4. Number of genes expressed in 12 cell lines (% detected): number (and proportion over number of detected genes) of genes belonging to this category that is detected in 12 cell lines (which is the total number of cell lines represented by all short RNA-seq experiments with 2 bio-replicates),
  5. miRNA guide fragment: for the miRNA gene class is provided the cumulative number of annotated expressed guides for all detected miRNAs (here virtually all annotated guides of detected miRNAs are expressed),
  6. miRNA passenger fragment: for the miRNA gene class is provided the cumulative number of annotated expressed passengers for all detected miRNAs (here virtually all annotated passengers of detected miRNAs are expressed),
  7. Internal fragments of annotated short RNA (average per detected gene): for each gene class is provided the average number of internal fragments per detected gene. An internal fragment of a detected gene is defined as a short RNA-seq mapping which 5' end lies 5 bp after the start and 5 bp before the end of a detected gene.

- Nucleotide coverage of small annotated RNAs over annotated long RNAs (Figure S18)

  Features (gene, transcript, exon, intron, UTR, CDS) of long annotated and novel intergenic / antisense transcripts were projected onto the genome (separately for each strand). Small annotated RNAs were intersected with these projected features and nucleotide coverage was calculated for both the long and the small transcripts. Gencode v7 lists 45 different transcript types. We merged biotypes with similar features into the following 'superfamilies':

  Long transcripts:

Superfamily	Gencode v7 transcript biotype
Long non coding:	non coding, processed transcript, antisense, retained intron, ncrna host, lincRNA
protein coding:	ambiguous orf, protein coding, nonsense mediated decay
processed pseudogene:	processed pseudogene, pseudogene, transcribed processed pseudogene, retrotransposed
unprocessed pseudogene:	processed pseudogene, pseudogene, transcribed processed pseudogene, retrotransposed
other pseudogene:	IG C pseudogene, IG J pseudogene, IG V pseudogene, rRNA pseudogene, TR V pseudogene
rRNA:	rRNA, Mt rRNA

IG gene:	IG C gene, IG D gene, IG J gene, IG V gene
TEC:	TEC
TR gene:	TR C gene, TR J gene, TR V gene
novel transcript models:	novel transcript models for all cell lines with IDR <=0.

Small transcripts:

Superfamily	Gencode v7 transcript biotype
small ncRNA:	miRNA, misc RNA, Mt tRNA, snoRNA, snRNA, tRNAscan (tRNA predictions by tRNAscan)
small ncRNA pseudogene:	miRNA pseudogene, misc RNA pseudogene, Mt tRNA pseudogene , scRNA pseudogene , snoRNA pseudogene, snRNA pseudogene, tRNA pseudogene
Short RNA contigs:	short RNA contigs for all cell lines with npIDR ≤ 0.1

Unannotated Short RNAs (Table 2B, Figure S16)

- Summary statistics (Table 2B).

To describe previously unannotated short RNAs we pool contigs from all short RNA-seq experiments with 2 bio-replicates (cell, nucleus and cytosol compartments) that have a npIDR ≤ 0.1 and do not overlap any small Gencode annotated gene (including pseudogene) on the same strand. (Note that actually the vast majority of short RNA-seq contigs that pass npIDR are un-annotated by this definition). We collapse the short RNA-seq contigs across experiments into a non-redundant set for a given compartment, considering two contigs with exactly the same genomic coordinates (i.e. chromosome, start, end, strand) as the same contig.
We then asses the abundance of these unannotated short RNAs across genomic regions for each of the cellular compartment separately as well as for all compartments combined (see Table 2B). This is done using cell compartment specific stranded partitions of the genome derived from polyA+ expressed Gencode v7 and novel intergenic / antisense exons (or a global partition when assessing the combined compartments). In addition, we investigated an overlap of these short RNAs with the intron-exon structure and TSS/TTS of long transcripts. For exon, intron, gene and intergenic region numbers, we required a total and stranded inclusion of the short un-annotated contig. For the exon-intron and gene-intergene boundaries, we simply calculated the difference with the total number of genes and total number of genes + intergenic elements, respectively.

- Figure S16.

We calculated the coverage of various types of annotated small RNAs with short RNA-seq/CAGE reads. For each of the Gencode v7 small transcript category (miRNA, snoRNA, snRNA, tRNA), and for each short RNA-seq/CAGE experiment done in the nucleus or in the cytosol (6 and 7 cell lines each respectively), we applied the following procedure:

1. All Gencode transcripts belonging to this category were divided into 10 distance bins,
2. For each short RNA-seq/CAGE read mapping with its 5' end into a Gencode transcript of this category, the distance from its 5' end to the Gencode transcript 5' end was computed.
3. For each Gencode transcript and each bin we counted the presence or absence of the short RNA-seq/CAGE 5' end.

The frequencies obtained for a given compartment and technology were further normalized by the total number of experiments for this compartment and technology.

Origins of Short RNAs (Figures S15 and S19)

The long RNA contigs overlapping detected short RNA contigs implicitly provides 'host' or 'precursor' information of short RNA. The heatmaps on fractionated cellular samples help to understand their distribution and abundance in different cellular compartments. The rpkm value of long RNA contig overlapping different known short RNA classes are then plotted on these heatmaps. Meanwhile, the distribution of short RNA contig is also illustrated using scatter plot of log-ratio of rpm value of cytoplasm over nucleus. Putting the scatter plot together with the heatmap illustrates the distribution of both short and long RNA in different cellular fractions and therefore helps to understand the genealogy of detected short RNA or their relationship to long RNA.

Allele specific expression (Table S10)

In order to assess the amount of allele-specific expression (ASE) present in the GM12878 RNA-seq datasets we used the AlleleSeq pipeline[26]. RNA-seq reads were independently mapped using Bowtie against both maternal and paternal haplotype sequences constructed for the NA12878 genome using phased variant calls (SNPs, indels and deletions) from the pilot phase of the 1000 Genomes Project Consortium (2010). Heterozygous SNPs that are in sufficiently highly transcribed regions can be used to distinguish those regions that exhibit ASE (regions that are preferentially expressed from only one haplotype) from those that are not, by counting reads mapping to each allele. A false discovery rate of 0.01 was selected using an explicit computational simulation given a null model of equal expression from both haplotypes.

Using the AlleleSeq pipeline we have analyzed both the long polyA+ and polyA- GM12878 RNASeq data for all three cellular fractions (whole cell, cytoplasm and nucleus) as well as for pooled sets of these RNASeq datasets (see Table S10). We find that the proportion of protein-coding Gencode v7 genes that show ASE is the same for all three cellular fractions. In order to maximize our statistical power we pooled the long polyA+ RNA from all three cellular fractions, revealing that 375 out of 2,153 genes (17%) that are assessable for ASE (i.e. sequenced at sufficient depth to potentially allow determination of ASE as well as containing a heterozygous SNP) exhibit the behavior. We performed a similar analysis for the annotated Gencode v7 long non-coding RNAs, using the pooled the reads from all cellular fraction from both the long PolyA+ and PolyA- RNA-seq datasets we found that 147 out 816 ASE assessable long non-coding RNAs (18%) show ASE behavior. Thus the fraction of protein coding genes and long non-coding RNAs that exhibit ASE are quite consistent.

Moreover, we found that 88% (68 of 77) of protein-coding genes that contain two or more ASE SNPs are completely consistent in terms of the haplotype being expressed. In order to assess the effect of potential enrichment of ASE SNPs in the vicinity of exon-intron splice junctions we trimmed exons by 5nts on either side and found a similar number of genes showing ASE (the number of ASE genes changed by less than 6% using the shortened exons).

Repeat region transcription (Figure S21)

- Mapping and new technologies

  CAGE reads were mapped to the genome using Delve a probabilistic aligner. For each mapping location Delve assigns a mapping quality or the probability that the mapping is correct (see Heng Li et al.[27] for a detailed description of how mapping qualities are derived). While it is true that repetitive regions almost by default obtain lower quality scores, in the present study we set a mapping threshold to 10 or 90% confidence that the read originated from the reported location. Hence all regions highlighted by CAGE, irrespective of their annotation, represent starting sites of transcription. Lowering the thresholds may reveal many additional repeat loci currently missed due the stringent quality thresholds applied to the data. Longer read technologies in future may help to further disambiguate transcription from repetitive elements.

- Hierarchical annotation
  Repetitive elements often occur around or within known gene structures including promoters. To avoid mis-labeling a CAGE cluster as a repeat cluster, while it probably is more likely to correspond to an overlapping classical promoter, we annotate clusters in a hierarchal manner. All CAGE clusters are first scanned against promoters, and only the non-promoter CAGE clusters compared agains exons and so forth. The full order of annotations used was TSS > TTS > EXON > INTRON > SINE > LINE > LTR > other repeats > intergenic. This scheme ensures that CAGE clusters annotated as repeats are not overlapping other genomic annotations relevant to transcriptional initiation.

- Extension of work by Faulker et. al[28]
  The present analysis on repeat element description extends the work by Faulker et. al.[28] in several ways. We were able to demonstrate that repeats exhibit cell, rather than tissue, specific expression patterns, confirm that transcripts originating from within repeats remain predominantly in the nucleus and that repeats not overlapping other gene annotations are important.

Enhancer RNAs (Figures 5 and S22)

Predicted enhancer loci were taken from the ENCODE elements group[29] for GM12878 cells. These predictions are based on chromatin evidence of transcription factor binding to non-promoter regions. The analysis was focused on enhancer predictions more than 10kb distal to any Gencode annotation. RNA elements (long RNA-seq contigs from polyA+ and polyA- RNA and CAGE tag clusters from pooled sub-cellular fractions) within a 5kb window around the enhancer prediction were listed (accessible at supplementary data file below). DNAse I Hypersensitive (HS) sites[29] overlapping the enhancer predictions were used as the central position to align loci for Figure 5A. The plots show the number of RNA elements of each type at any distance relative to these loci divided by the total number of predictions with DNAse I HS sites.

The relative long RNA-seq signal at intergenic enhancer predictions (RPKM) was computed in each RNA fraction in a 1kb window. For enhancer with at least 0.1 RPKM pooled from all fractions, the ratios of polyA+ to pooled polyA+/polyA- signal and of nuclear to pooled nuclear/cytoplasmic signal was determined.

For comparison the same ratios were calculated at annotated Gencode promoters and at predicted novel intergenic promoters from the ENCODE elements group[29]. Chromatin signals were taken from the ENCODE chromatin tracks (Broad Histone ChIP for GM12878[30]). The signal in a 5kb window around each enhancer was averaged (in terms of fold change over the genome-wide average). The enhancers were then grouped into those showing evidence of transcription initiation based on CAGE tags (over 0.1 tags per million aligned tags) in the enhancer locus and those without.

Supplementary data file for figure 5A:
ftp://genome.crg.es/pub/Encode/Transcriptome_paper/10_DnaseoverlapYipPaperbef_Gm12878_Pooled_output_matches.txt.gz

Genome Coverage (Figure S23, Table S11 and Figure 6)

- Genome covered by different kinds of elements (Figure S23 and Table S11).

The percentage of whole genome / encode regions (excluding gaps) covered by the following kinds of elements was computed for different number of reads x supporting junctions and contigs :

- contigs (introns) : contigs (junctions) from all PE RNA-seq experiments with more than x supporting reads were projected onto the genome before computation of coverage,
- intersection introns+contigs: only nucleotides present in both a junction and a contig with more than x supporting reads for a given PE RNA-seq experiment were considered,
- union introns+contigs: nucleotides present either in a junction or in a contig with more than x supporting reads for a given PE RNA-seq experiment were considered,
- contigs+Gencode exons: nucleotides present either in contigs or in Gencode exons were considered,
- contigs+introns+Gencode genes: nucleotides present either in contigs+introns or in Gencode genes (exons+introns) were considered.

- Novel transcripts and intergenic space reduction (Figure 6).

Intergenic space was calculated relative to the following gene data sets:
1. expressed Gencode v7 genes (gen7 expressed),
2. expressed Gencode v7 genes and cufflinks *de novo* gene predictions (gen7 expressed and novel ig/as).

Genes of the two data sets were projected onto the unstranded genome and assembly gaps (as annotated by UCSC) were removed subsequently. The resulting intervals were binned (bin size: 10kb) and used for further analysis.

IV. SUPPLEMENTARY FIGURE AND TABLE LEGENDS

Figure legends

Figure S1: Sample Flowchart.
The ENCODE transcriptome data are obtained from several cell lines which have been cultured in replicates. They were either left intact (whole cell) and/or fractionated into cytoplasm and nucleus prior to RNA isolation. Total RNA was then isolated and partitioned into RNA > 200bp (long) and < 200bp (short). The long RNA was further partitioned over an oligo-dT column into polyA+ and polyA- fractions. The K562 cell line also underwent additional fractionation into nucleoli, nucleoplasm and chromatin, but no further partion into polyA+ and polyA-. RNA-seq was conducted on polyA+, polyA- and total (K562) RNA samples. CAGE was conducted primarily on polyA+ and total RNA but also some polyA- samples. RNA-PET was conducted on PolyA+ samples only (not shown here are RNA-seq experiments performed at CalTech on polyA+ whole cell RNA extracts).

Figure S2: Data Processing.
This figure shows an overview of the transcriptome data processing pipeline. Raw read data (FASTQs) from each biological replicate is independently mapped against the hg19 ENCODE sex-specific assemblies according to the gender of the sample to generate .BAM (alignment) and .wig (signal) files. Each data type is processed though a custom pipeline developed by the production lab and subsequently distilled into elements, like splice junctions, contigs, de novo assembled genes and transcripts, CAGE and PET clusters. Though this is usually done by pooling biological replicates each element is independently quantified per replicate to allow for a statistical assessment of reproducibility. The mapped data is also used to quantify Gencode annotated features, such as genes, transcripts and exons. All elements and features are assessed for their reproducibility using npIDR or IDR (see Supplement Methods). Finally, all data is sent to the Data Co-ordination Centre (DCC) at UCSC.

Figure S3: Gencode annotation growth over time.
This figure shows the number of Gencode (A) genes and (B) transcripts over time. Ensembl-only: found by the Ensembl pipeline only; Havana & merged: found by Havana manual annotation or confirmed by both Havana and Ensembl.

Figure S4: Nucleotide coverage and relative expression of Gencode exonic, intronic and intergenic regions.
Using RNA-seq data (RNA-seq contigs with npIDR ≤ 0.1), we estimated the relative coverage of the whole genome (unstranded, excluding gaps) (black), the relative coverage of the 3 main genomic domains (exons, introns, intergenic

regions) (red), and the distribution of RNA-seq nucleotides (nt) in the 3 main genomic domains (blue). The box plots are generated from values across all cell lines illustrating the variation across cell lines. The largest point shows the cumulative value, the smaller points depict outliers. D: set of nt of a given genomic domain; R: set of nt in one or several RNA-seq experiments; ∑D: set of nt in the genome.

Figure S5: Cell line contribution to the detected transcripts.
All CSHL and Caltech PE long RNA-seq experiments with 2 bio-replicates are considered, representing a total of 14 cell lines. Cell lines are ordered by number of nucleotides contributed, i.e. by number of distinct unstranded nucleotides present in the npIDR'ed contigs of this cell line (cell line contribution). Nucleotides present in a given cell line but not in the 13 others are considered specific to this cell line (cell line specific nucleotides). The cumulative number of nucleotide of a given cell line is the number of distinct unstranded nucleotides present in the set of cell lines seen until this point (cumulative total). At a given cell line, the cumulative novel nucleotide is the subset of the cumulative total nucleotides that are not Gencode v7 exonic nucleotides (using an unstranded Gencode v7 partition of the genome).

Figure S6: RT-PCR Validation of Novel Splice Junctions.
(A) A set of 3,000 GT/AG splice junctions identified from the Illumina RNA-seq data that are not annotated in Gencode and which map to intergenic and antisense regions of H1-ESC, HepG2 and Hela-S3 were selected for further validation using targeted RT-PCR. The unspliced reads from 2 mate pairs which share a common targeted junction were used to guide primer selection. These primers were used to separately amplify cDNA from the relevant cell line. The products were run on an agarose gel (data not shown) and pooled for sequencing on the Roche FLX 454.
(B) The percentage of candidate junctions validated by Roche 454 sequencing as a function of the number of supporting RNA-seq reads, for the H1-hESC validation experiment. Different lines (1-5) correspond to the minimum number of Roche 454 reads per junction required for validation.

Figure S7: Distribution of spectral and peptide identifications in novel exons.
The height of each bar represents the number of model sequences for which there were peptide matches in novel exons. Red, blue, and green bars show that these model sequences were identified from 5 or fewer, 10 or fewer, or more than 10 spectra, respectively. The numbers at the bottom of each bar show how many distinct peptides were identified for these models.

Figure S8: Cumulative expression of (A) genes and (B) transcripts.
Shown is the Empirical Cumulative Distribution Function (ECDF) of gene and transcript expression for all genes and transcripts of a particular cell line within a given RNA fraction and compartment. To include non-expressed genes and transcripts in the graph, we adjusted those elements with expression levels of 0 RPKM to an artificial value of 10^{-6} RPKM, so that the onset of each graph represents the fraction of non-expressed genes and transcripts. Only features with npIDR ≤ 0.1 are shown.

Figure S9: Abundance of genes types in cellular compartments.
(A) Shown are 2D Kernel density plots for (1) long non-coding, (2) protein coding, (3) small non-coding and (4) novel intergenic / antisense (cufflinks) genes, representing the nuclear/cytosolic enrichment of those genes vs their abundance in the whole cell extract. Only those genes present in all three RNA extracts are displayed. The actual values of the estimated Kernel density are indicated by the color shades. (N* = average number of genes per cell line, 7 cell lines total). Not filtered by npIDR.
(B) The box plots represent the nuclear/cytosolic abundance of various gene biotypes in different cell lines. The larger the ratio, the more nuclear enriched a biotype is (npIDR ≤ 0.1).

Figure S10: Cell line specific genes.
Number of genes detected in multiple cell lines. Only protein-coding, non-coding and novel intergenic/antisense genes with npIDR ≤ 0.1 were counted as expressed

Figure S11: Splice junction usage.
For each protein coding gene in each cell line, we compute (A) the number of detected splice junctions, (B) the relative expression of the most frequently used splice junction, and (C) the Shannon's diversity index on the relative usage of the

gene's annotated splice junctions. The box plots display, separately for genes with different number of isoforms, the distributions pooled together for all cell lines. Shannon's entropy: The red dot is the maximum entropy given a number of isoforms. The blue dots are half this value.

Figure S12: Usage of major isoform across all cell lines.
Here we have calculated how often a transcript appeared as the isoform with the highest relative expression (= major isoform) per gene across all cell lines. (A) Number of multi-transcript genes that use one, two, etc, major isoforms across all cell lines. Within each bar, the different colors correspond to the different number of annotated isoforms. The black line is the function 1/n, where n is the number of annotated isoforms.

Figure S13: Distance between CAGE and PET TSS to the closest RNA-seq expressed Gencode TSS.
The 82,783 CAGE and the 63,864 PET 5' end clusters / TSS obtained from all CAGE and PET experiments were compared to the 97,778 polyA+ RNA-seq expressed Gencode TSS. Plotted here is for each such CAGE and PET TSS the distance to the closest expressed Gencode Transcription Start Site (TSS).

Figure 14: Transcription Start Sites (TSS).
Heatmap showing the presence (red) or absence (yellow) of various features at putative transcription start sites (5' ends of RNA-seq transcript models expressed in at least one cell line). Each line represents one putative TSS in one cell line. The 'Transcript' column indicates if an RNA-seq transcript model from this TSS is expressed in this sample. 'Cage' shows the presence of a Cage cluster, 'CageHMM' a Cage cluster filtered by the HMM TSS filter. The other columns show DNAse Hypersensitivity sites and ChIP-seq peaks for various histone modifications and DNA binding proteins associated with promoter regions.

Figure S15: Compartmentalization of annotated small RNAs.
Annotated small RNAs (miRNA, snoRNA, snRNA, tRNA) show sub-cellular localization patterns according to their functions. (A) Nuclear/cytosolic enrichment versus whole cell expression. (B) The abundance of each annotated small RNA class in a cell compartment is represented as the sum over all RPMs of individual transcripts. (C) Shows the prevalence of a specific class within the repertoire of small RNAs detected in a sub-cellular compartment. A: all cell lines; B,C: only K562.

Figure S16: Fragments of short RNA-seq and CAGE in annotated small RNAs.
Shown is the coverage of annotated small RNAs (miRNA, snoRNA, snRNA, tRNA) by short RNA-seq read / CAGE tag 5' ends in the nucleus and the cytosol. The coverage is calculated as the number of short RNA-seq read / CAGE tag most 5' ends which fall at a given distance from the annotated small transcript 5' end (shown on the x-axis). The distance (i.e. transcript length) has been normalized using bins. The counts have been derived per individual cell line (see Supplementary methods for more details)

Figure S17: Proportion of annotated elements in different genomic domains that overlap different classes of small RNAs.

Figure S18: Nucleotide coverage of small RNAs over long RNAs.
Gencode v7 annotated small ncRNAs (miRNA, snoRNA, snRNA, tRNA) in elements (CDS, introns, UTRs, exons and intergenic regions) of (A) protein coding and long noncoding transcripts and (B) novel intergenic / antisense transcript models.

Figure S19: Expression of long RNA contigs corresponding to detected small RNA.
The expression of the long RNA contigs which corresponding to detected small RNA are color-coded in the heatmaps. Blue indicates no expression yellow indicates high expression. The log-ratio of detected small RNA expression in cytoplasm over nucleus is shown in the scatter plot on the left side. Cytoplasm enriched small RNA contigs are distributed to the right side of zero while nucleus enriched small RNA contigs to the left. (A) for detected miRNA contigs; (B) for snoRNA; (C) for snRNA; (D) for tRNA and (E) for total unannotated small RNA.

Figure S20: Profile of RNA editing in ENCODE whole-cell datasets and compartments.

(A) The profile of RNA-detected single nucleotide variants (SNV) in GM12878 that are detected independently in both the Caltech whole-cell polyA-selected non-stranded dataset and the CSHL stranded dataset, with 65% of the detected SNVs match entries in dbSNP 132, and showing a balanced distribution of A->G and G->A substitutions. More than 80% of the additional 7067 (13%) RNA SNVs that are not within 5bp of an intron boundary are A->G substitutions, with G->A corresponding to less than 4%. (B) SNV substitution frequency in the same samples as B. While A->G SNVs are always the most prevalent RNA-based SNV, they represent less than 60% of the total in six of the 10 samples.

Figure S21: Cell specific expression of repeat elements.
(A) Shannon Entropy of CAGE cluster expression profiles across all experiments separated by broad annotation categories. Intergenic LINE, SINE and LTR repeats are noticeably more narrowly expressed than other categories. Heatmaps of LINE (B), SINE (C) and LTR (D) repeat expression across cell lines and compartments. Each column represents an individual repeat copy expressed at least 1 tag per million in any experiment. Expressed repeats predominantly cluster with other repeats in the same cell line rather than across compartments.

Figure S22: Diverse features of transcription at predicted enhancer loci, and eRNA cell type specificity.
Density plots of the relative RNA-seq signal in the (A) polyA+ and (B) nuclear RNA fractions compared to total signal pooled from all fractions. The majority of transcripts at enhancers are depleted in the polyA+ fraction and enriched in the nuclear fraction, but considerable diversity exists in both dimensions.

Figure S23: Genome Coverage.
The percentage of whole genome and pilot ENCODE regions coverage by RNA-seq contigs and introns as a function of the number of supporting RNA-seq reads per element. All Gencode v7 exons and introns are also included in the coverage calculation. The genomic gaps (as annotated by UCSC) are excluded from the calculation.

Table Legends

Table S1: ENCODE Cell Lines.
Cell lines profiled in this manuscript and by the ENCODE consortia. Tier 1 cell lines: K562, GM12878, H1-hESC. Tier 2: HepG2, HUVEC, Hela-S3, A549, SK-N-SH + Retinoic Acid, AG04450, MCF7. Tier 3 cell lines: BJ, NHEK, NHLF, HMEC, HSMM.

Table S2: RNA data and processing software.

Table S3: (A) Polyadenylated and (B) non-polyadenylated RNAs.

Table S4: Number of peptide identifications from proteogenomic mapping.
This table shows the number of total peptide identifications as well as the number of peptide identifications in novel exons only (noted in parenthesis). The results presented here are at 1% FDR. A total of 998,570 MS/MS spectra and 263,171 novel transcript sequences were used for this search.

Table S5: K562 nuclear subcompartments (total RNA).

Table S6: Elements specific to K562 nuclear subcompartments.
This table shows the total and unique number of elements (exons, splice junctions, transcripts and genes) detected in the K562 nuclear subcompartments. (1) Annotated elements and (2) Novel intergenic / antisense elements

Table S7: Cell line apecific Gencodegenes
Annotated protein coding genes expressed in the different cells lines.

Table S8: Reliable polyA+ Transcriptional Start Sites identified by CAGE.
This table shows the number of CAGE peaks (clusters), raw and filtered for reproducibility, in different genomic regions.

Table S9: Transcripts detected by RNA-PET experiments.
This table shows the intersection of the RNA-PET cluster, of various classes. MM indicates that both the PET 5' and 3' ends map within 1 million bases of each other on the same strand and that they both map within 150bp of an expressed transcript 5' and 3' end respectively, MX indicates that only the PET 5' end maps within 150 bp of an expressed transcript 5' end, XM indicates that only the PET 3' end maps within 150 bp of an expressed transcript 3' end, XX indicates that none of the PET ends map within 150 of an expressed transcript 5' and 3' end, but one of the PET end maps within an expressed transcript. They are shown with respect to the Gencode annotations (Gencode transcripts) and intergenic / antisense (cufflinks) transcripts.

Table S10: Allele specific expression.
Counts of Gencode v7 coding genes and long non-coding RNAs that exhibit allele-specific expression (ASE) for the various RNA-seq data sets for different cellular fractions as well as pooled datasets. This was done using the AlleleSeq pipeline[26]. For each RNA-seq dataset analyzed we display in the third column the counts of genes or long non-coding RNAs that are expressed at sufficient sequencing depth in order to assess allele-specific behavior and contain a heterozygous SNP. The last column shows the percentage of assessable genes that exhibit allele-specific behavior.

Table S11: Genome Coverage. Companion data for figure S23.

V. GEO ACCESSIONS

Data is available at GEO under the following accessions: GSE26284 (CSHL, Long RNA), GSE33480 (Caltech, A+ RNA-seq) GSE24565 (CSHL, Small RNA), GSE33600 (GIS, RNA-PET), GSE34448 (RIKEN, CAGE), GSE38886 (CSHL, splice junction validation by RT-PCR and 454 sequencing).

VII. SUPPLEMENTAL REFERENCES

1 Bhorjee, J. S. & Pederson, T. Rapid, preparative-scale purification of chromatin proteins. *Biochim Biophys Acta* **418**, 154-159 (1976).

2 Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**, e123, doi:gkp596 [pii]

10.1093/nar/gkp596 (2009).

3 Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**, 1543-1551, doi:gr.121095.111 [pii]

10.1101/gr.121095.111 (2011).

4 Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res* **32**, e37, doi:10.1093/nar/gnh031

32/3/e37 [pii] (2004).

5 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:nmeth.1226 [pii]

10.1038/nmeth.1226 (2008).

6 Dobin, A. *et al.* STAR: mapping non-contiguous RNA-seq data (GRCP024). *Genome research (submitted)* **XXX** (2012).

7 Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* **19**, 255-265, doi:gr.084541.108 [pii]

10.1101/gr.084541.108 (2009).

8 Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods in Molecular Biology Gene Regulatory Network* **786**, 181-200 (2012).

9 Carninci, P. *et al.* Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 520-524 (1998).

10 Salimullah, M. *et al.* Tunable fractionation of nucleic acids. *BioTechniques* **47**, 1041-1043 (2009).

11 Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37** (1996).

12 Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat Methods* **7**, 528-534, doi:nmeth.1470 [pii]

10.1038/nmeth.1470 (2010).

13 Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2**, 105-111, doi:nmeth733 [pii]

10.1038/nmeth733 (2005).

14 Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol* **303**, 19-44 (1999).

15 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:nbt.1621 [pii]

10.1038/nbt.1621 (2010).
16 Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res* **18**, 1-12, doi:gr.6831208 [pii]

10.1101/gr.6831208 (2008).
17 Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics* **5**, 1752-1779 (2011).
18 Derrien, T. *et al.* The GENCODE v7 catalogue of human long non-coding RNAs: Analysis of their structure, evolution and expression (GRCP002). *Genome research (submitted)*.
19 Khatun, J. *et al.* Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *Genome research* **XXX** (2012).
20 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214, doi:nmeth1019 [pii]

10.1038/nmeth1019 (2007).
21 Martin, D. *et al.* GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome biology* **5** (2004).
22 Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *Plos computational biology* **4** (2008).
23 Fu, Y. *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**, 741-747, doi:gr.115295.110 [pii]

10.1101/gr.115295.110 (2011).
24 Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome research* **10** (2000).
25 Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Site. *Science* **320** (2008).
26 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:msb201154 [pii]

10.1038/msb.2011.54 (2011).
27 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851–1858 (2008).
28 Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature genetics* **41**, 563-571, doi:ng.368 [pii]

10.1038/ng.368 (2009).
29 consortium, T. E. p. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **XXX** (2012).

30 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046, doi:10.1371/journal.pbio.1001046 (2011).