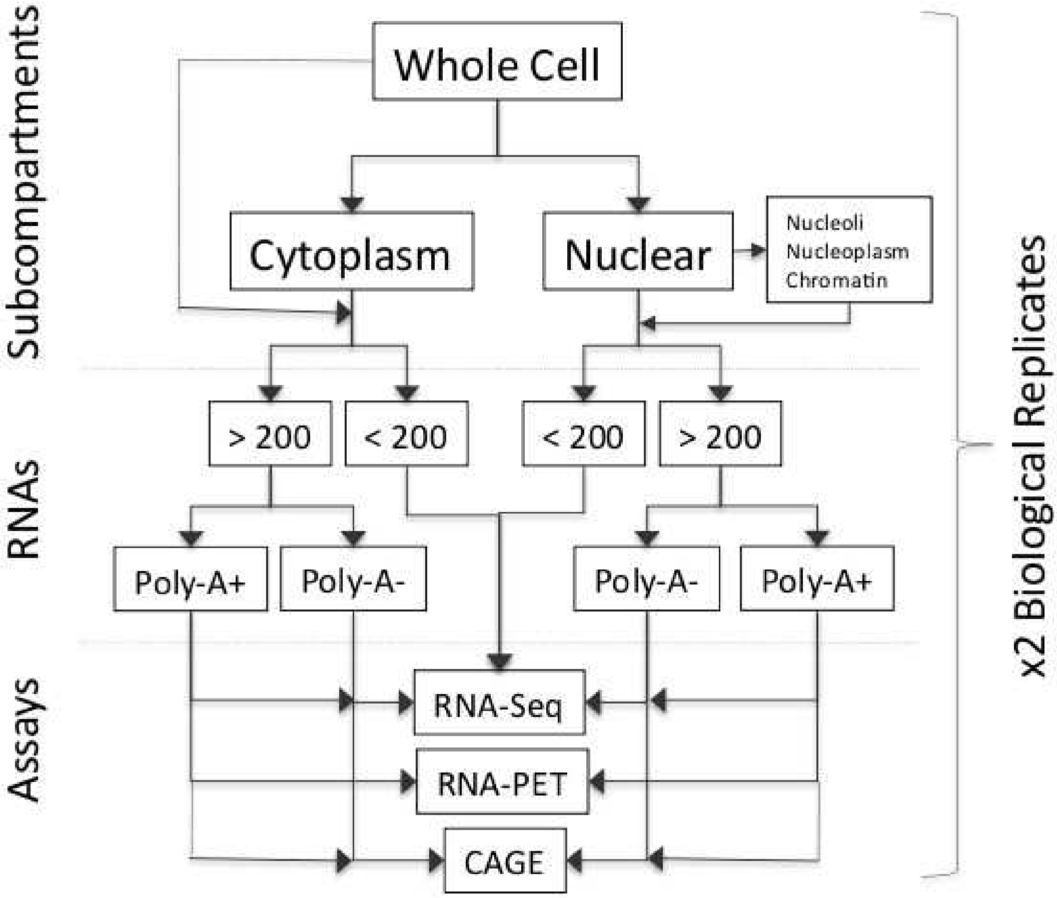
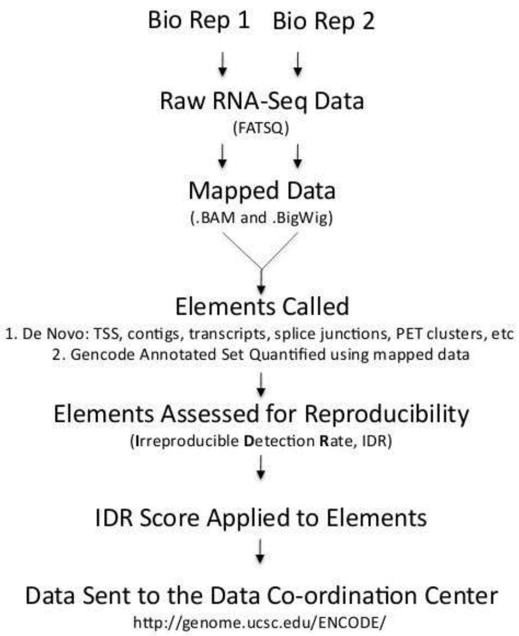
doi:10.1038/nature11233



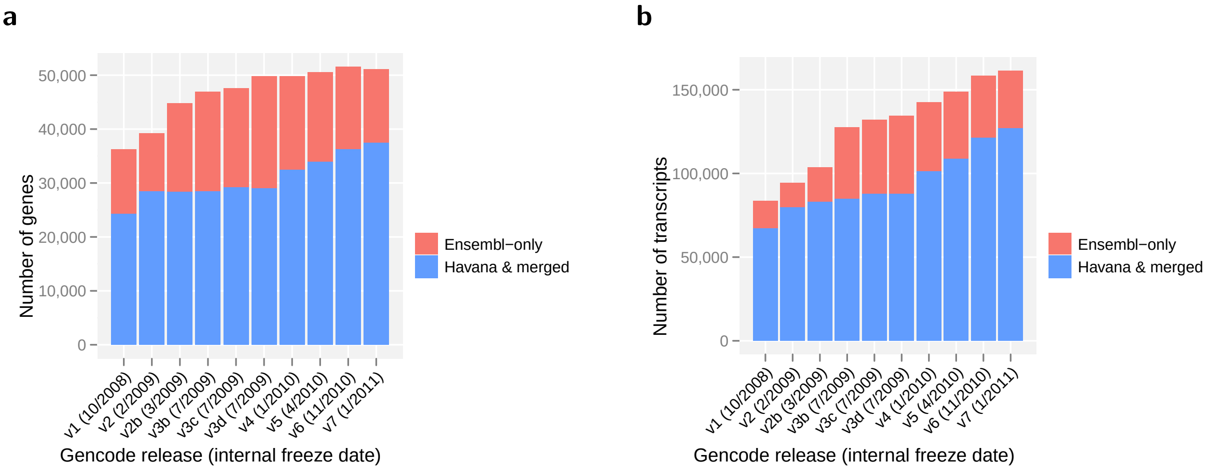
Supplementary Figure S1

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 partition into polyA+ and polyA- was done. RNA-seq was conducted on polyA+, polyA- and total (K562) RNA samples. CAGE was conducted primarily on polyA+ and total RNA but also on 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).

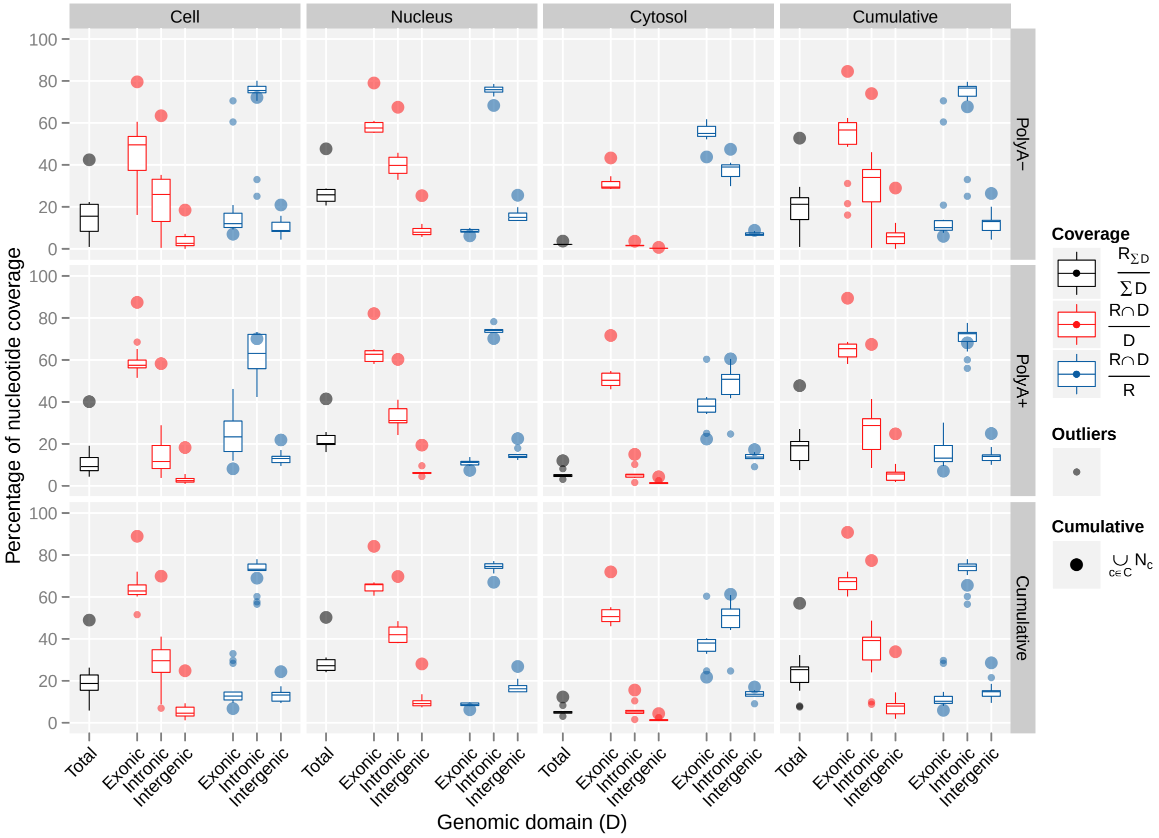

Bio Rep 1
Bio Rep 2
Raw RNA-Seq Data
(FATSQ)
Mapped Data
(.BAM and .BigWig)
Elements Called
1. De Novo: TSS, contigs, transcripts, splice junctions, PET clusters, etc
2. Gencode Annotated Set Quantified using mapped data
Elements Assessed for Reproducibility
(Irreproducible Detection Rate, IDR)
IDR Score Applied to Elements
Data Sent to the Data Co-ordination Center
http://genome.ucsc.edu/ENCODE/

Supplementary 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 section II). Finally, all data is sent to the Data Co-ordination Centre (DCC) at UCSC.

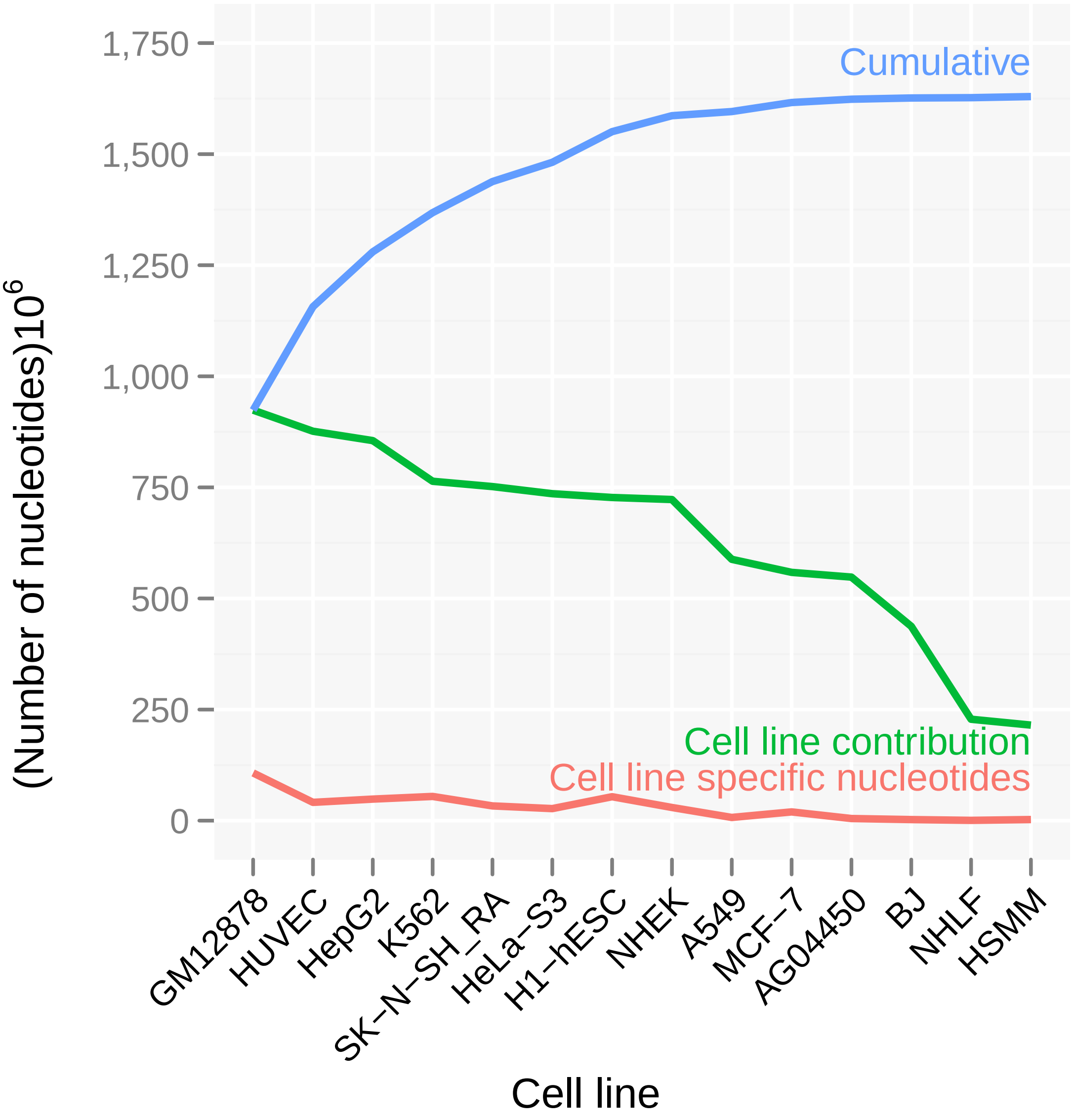


Supplementary Figure S3
Gencode annotation growth over time. This figure shows the number of Gencode (a) genes and (b) transcripts over time. Ensembl-only: found by the Ensembl pipeline only; Havana & merged: found by Havana manual annotation or confirmed by both Havana and Ensembl.



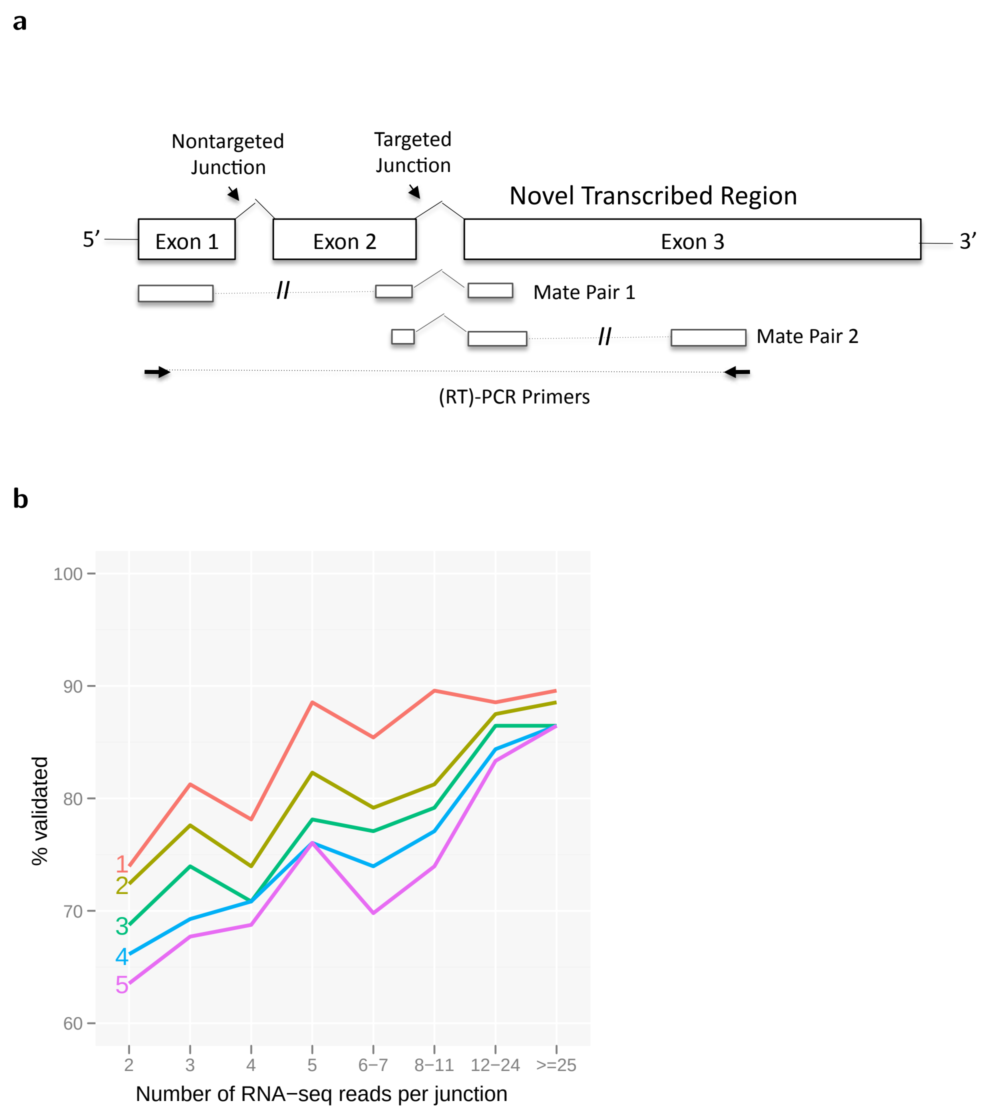
Supplementary Figure S4

Nucleotide coverage and relative expression of Gencode exonic, intronic and intergenic regions. Using RNA-seq data (RNA-seq contigs with $npIDR \leq 0.1$), 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; $\sum D$: set of nt in the genome.

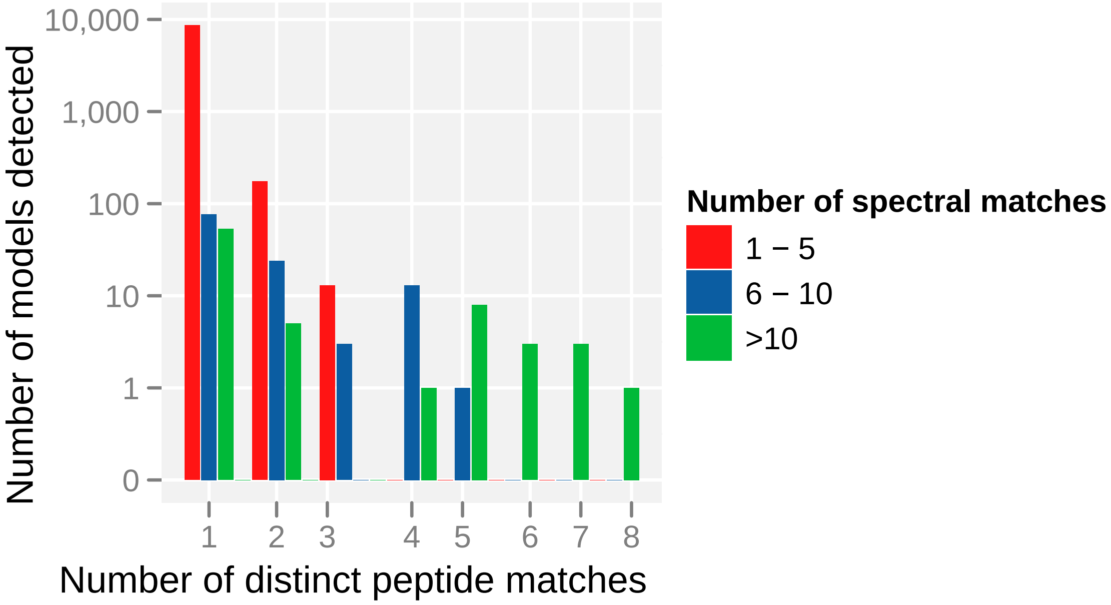


Supplementary Figure S5

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).

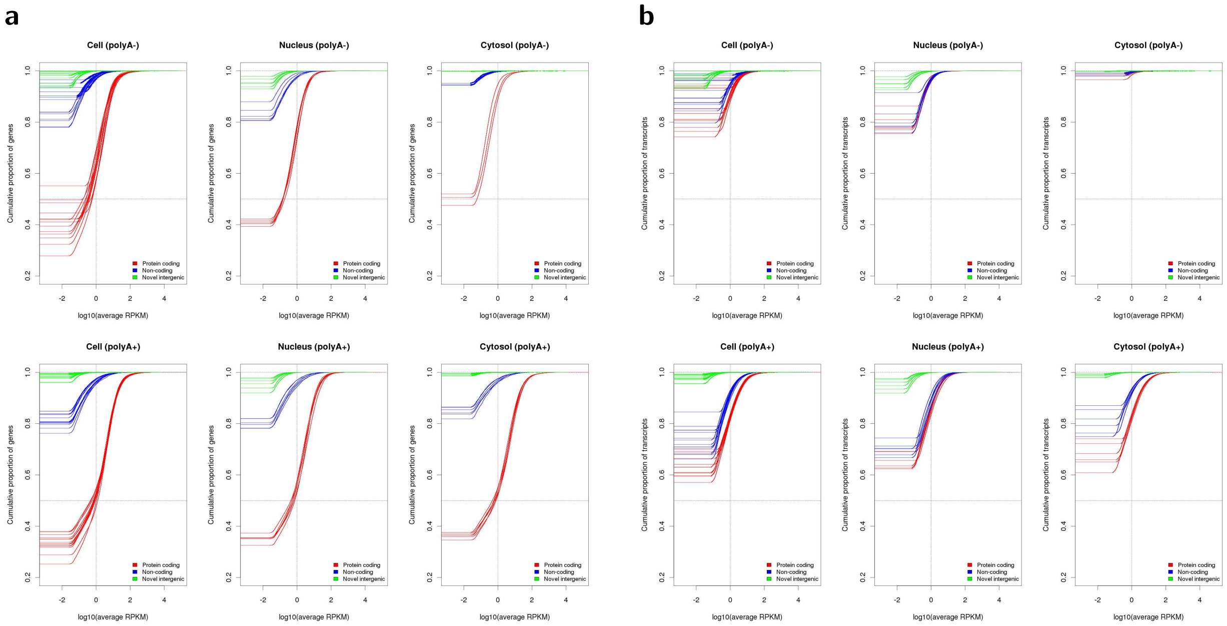


Supplementary Figure S6
RT-PCR Validation of Novel Splice Junctions. (a) A set of 3,000 GT/AG splice junctions identified from the Illumina RNA-seq data that are not annotated in Gencode and which map to intergenic and antisense regions of H1-hESC, HepG2 and Hela-S3 were selected for further validation 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.



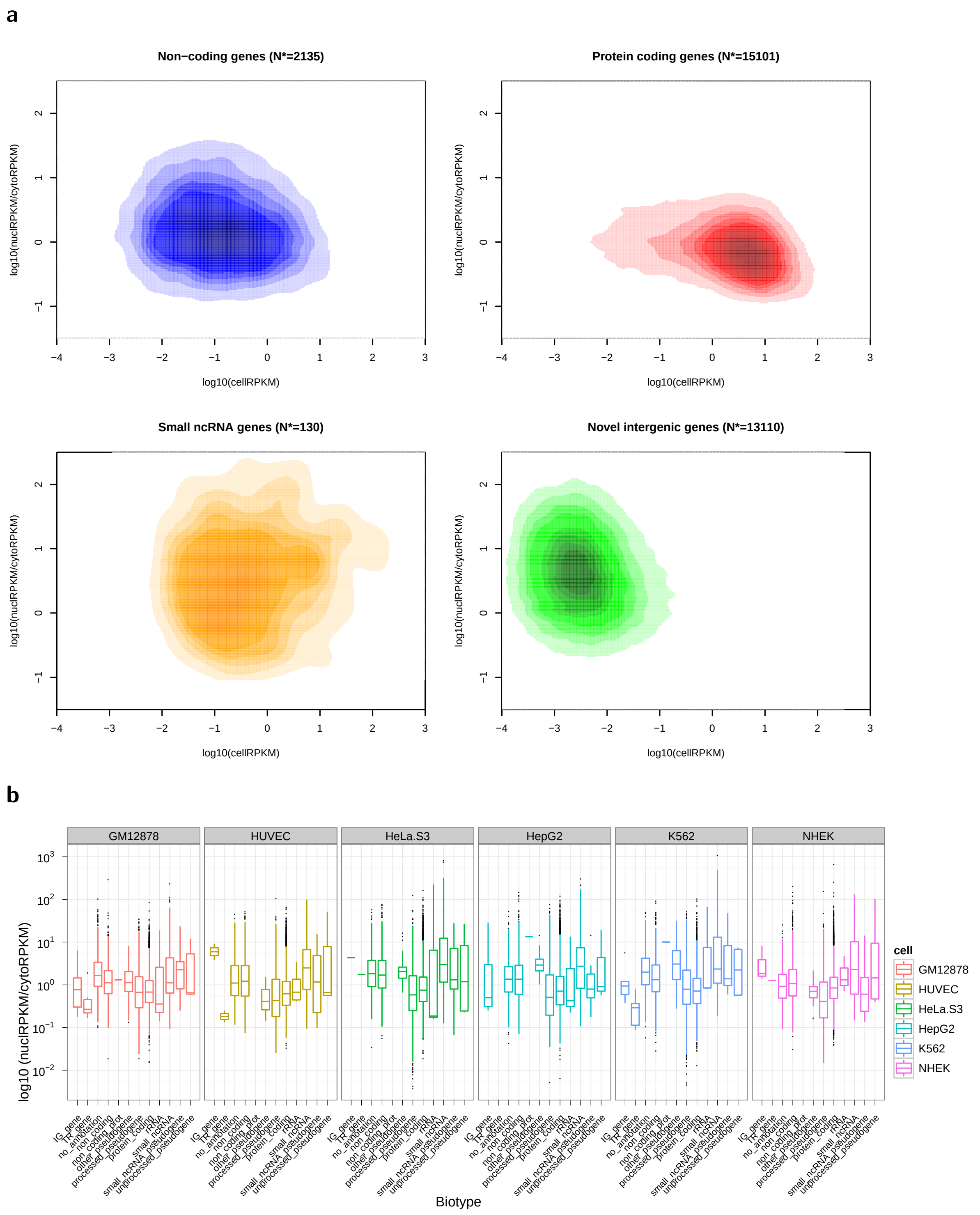
Supplementary Figure S7

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.



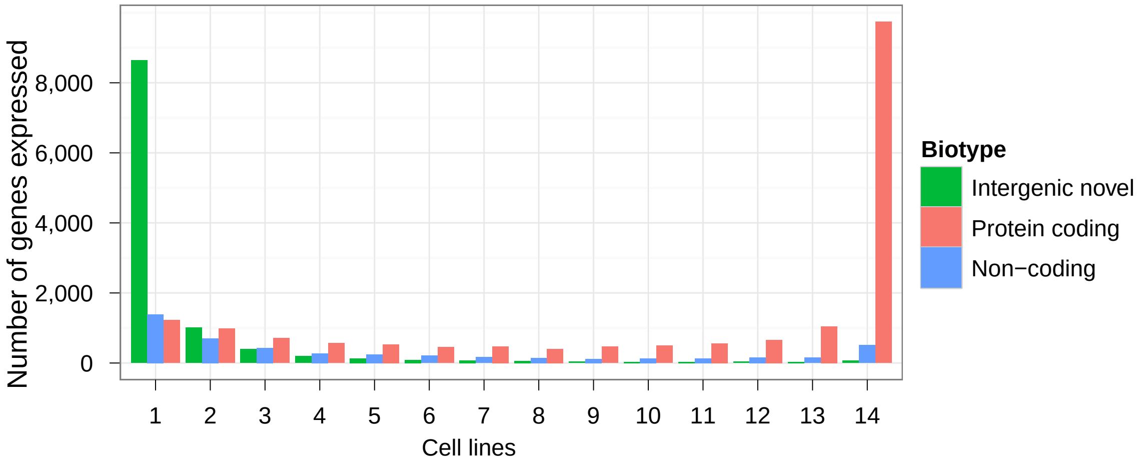
Supplementary Figure S8

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 $npIDR \leq 0.1$ are shown.



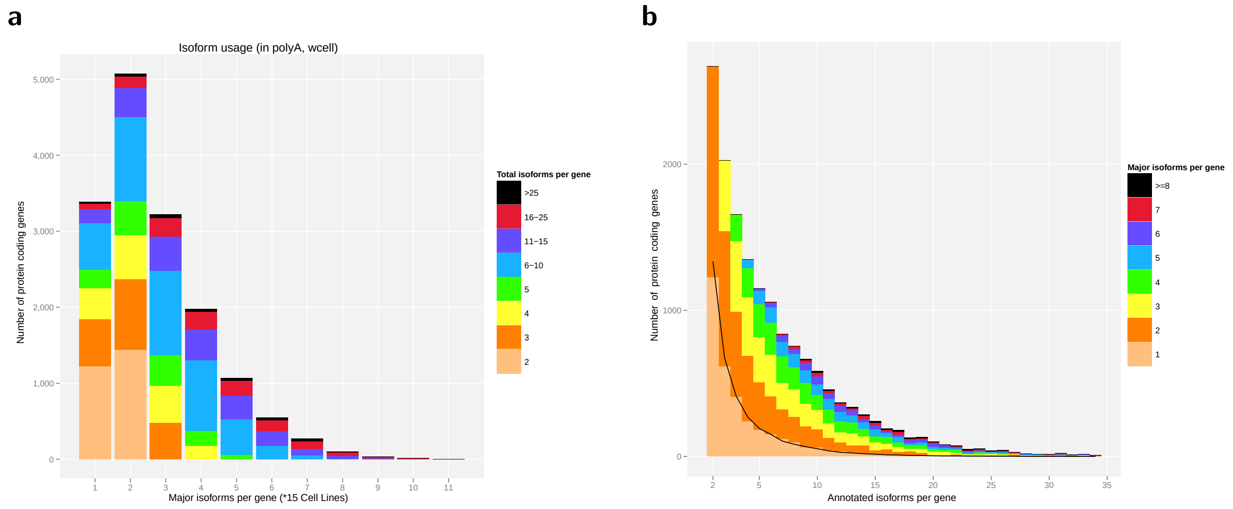
Supplementary Figure S9

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 ($npIDR \leq 0.1$).



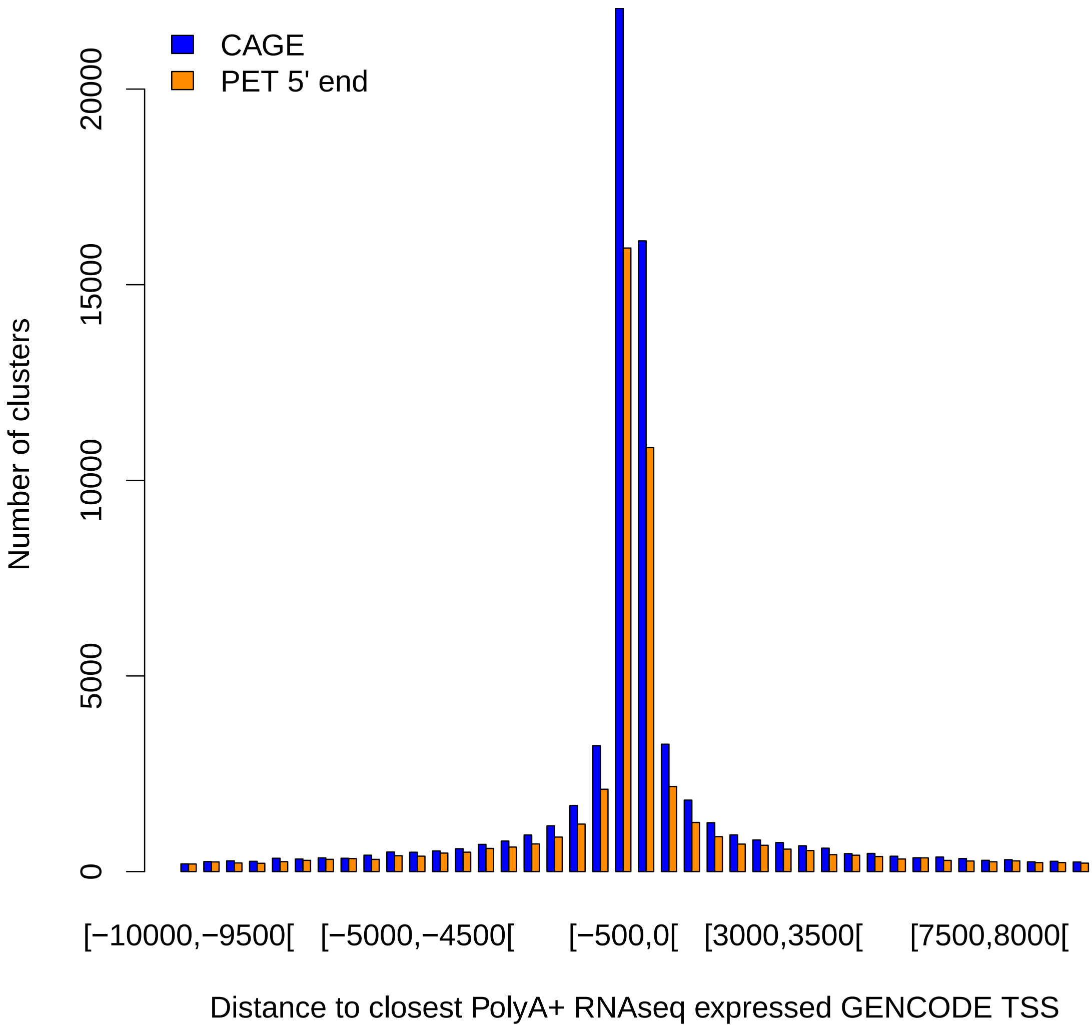
Supplementary Figure S10

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

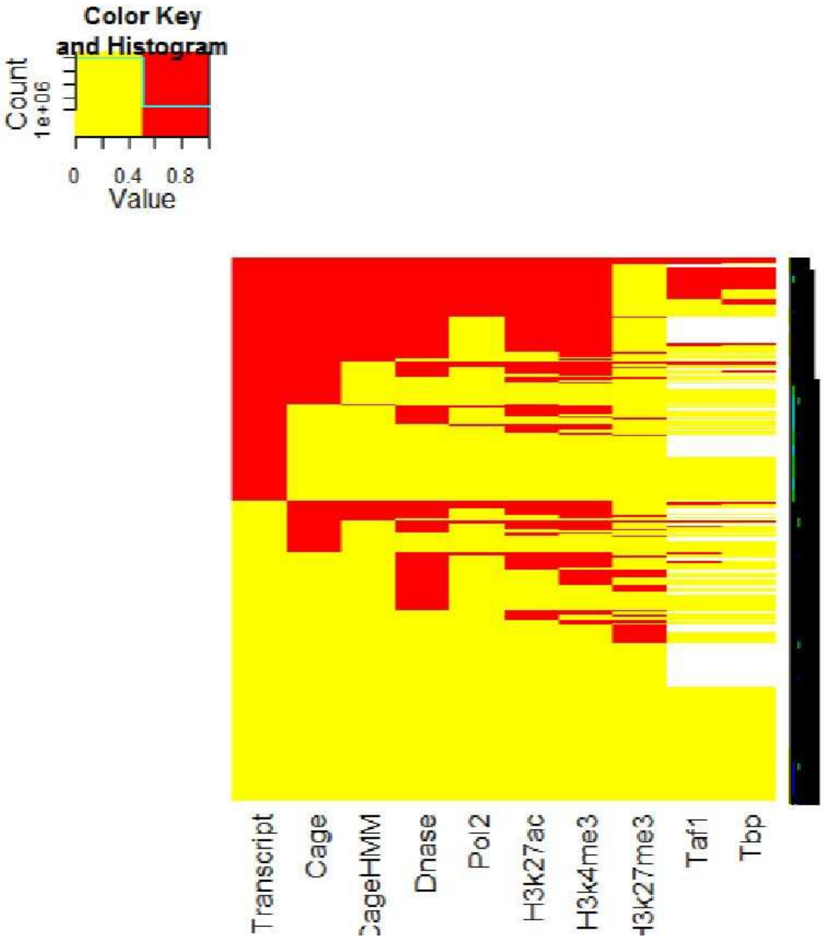


Supplementary Figure S11

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 protein-coding 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. b. Number of protein-coding genes that use two, three, etc, isoforms across all cell lines. Within each bar, the different colors represent the number of major isoforms per gene. The black line is the function $1/n$, where n is the number of annotated isoforms.

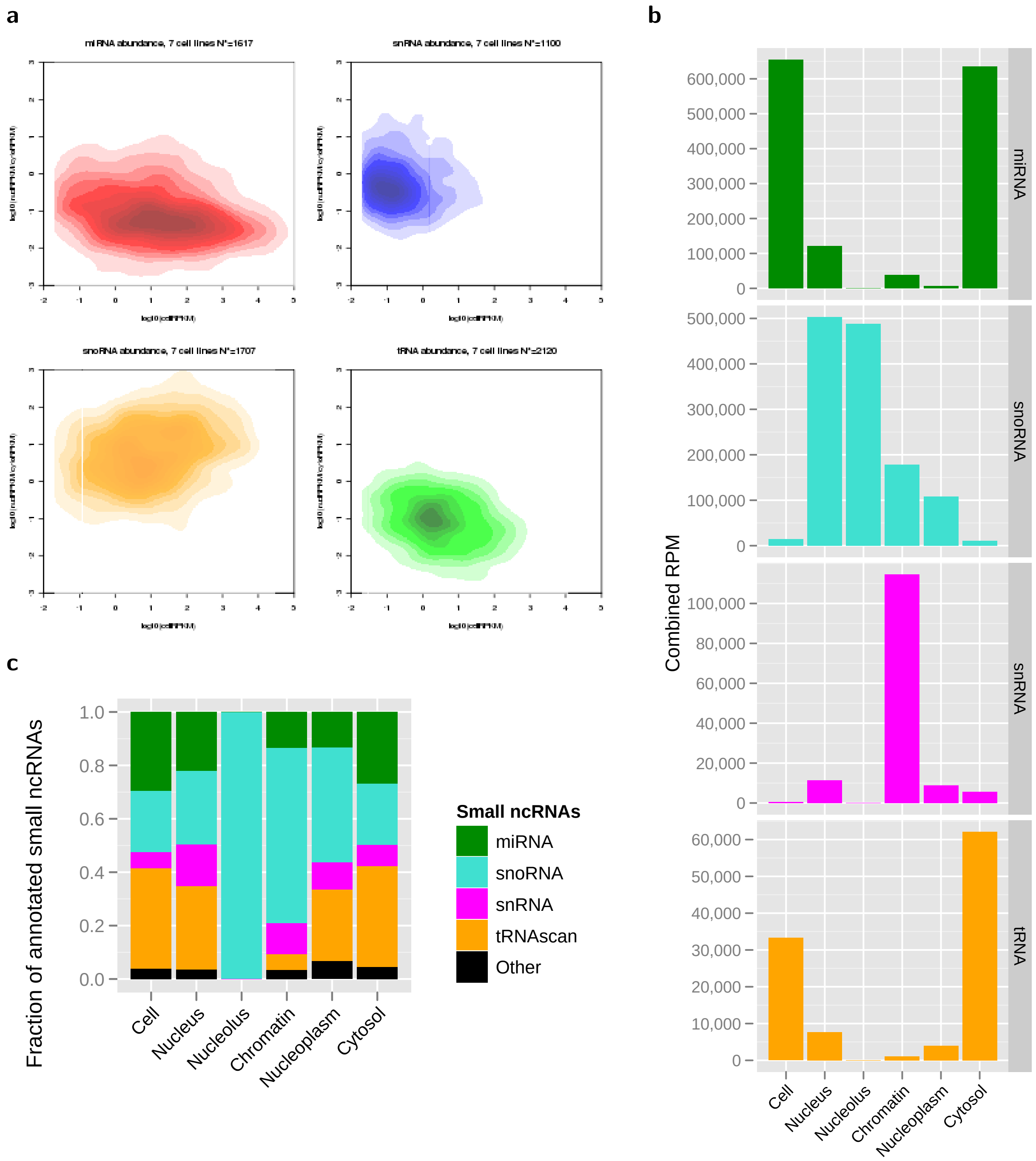


Supplementary Figure S12
Distance between CAGE and PET TSS to the closest RNA-seq expressed Gencode TSS. The 82,783 CAGE and the 63,864 PET 5' end clusters / TSS obtained from all CAGE and PET experiments were compared to the 97,778 polyA+ RNA-seq expressed Gencode TSS. Plotted here is for each such CAGE and PET TSS the distance to the closest expressed Gencode Transcription Start Site (TSS).



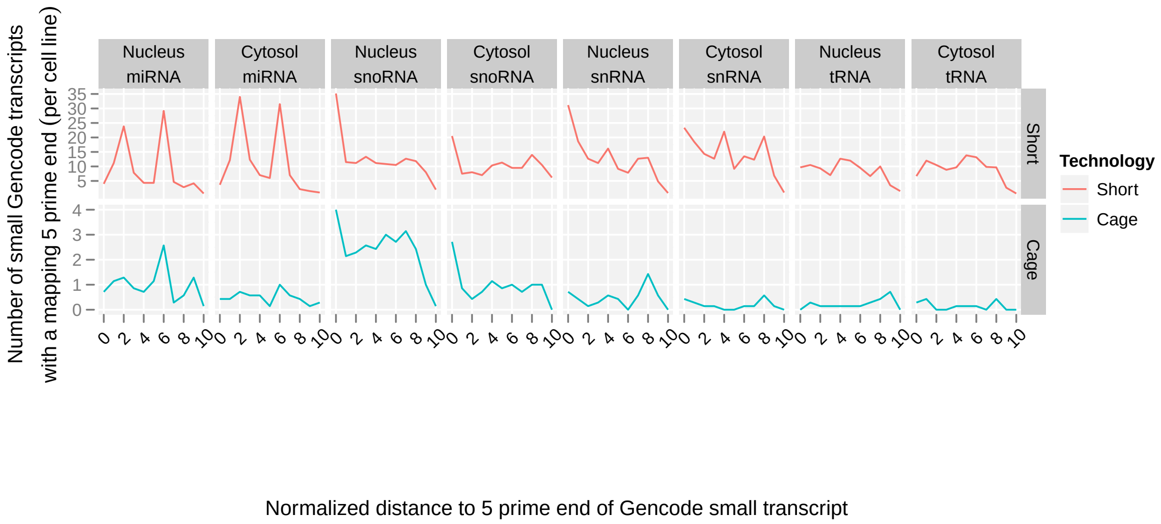
Supplementary Figure S13

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.



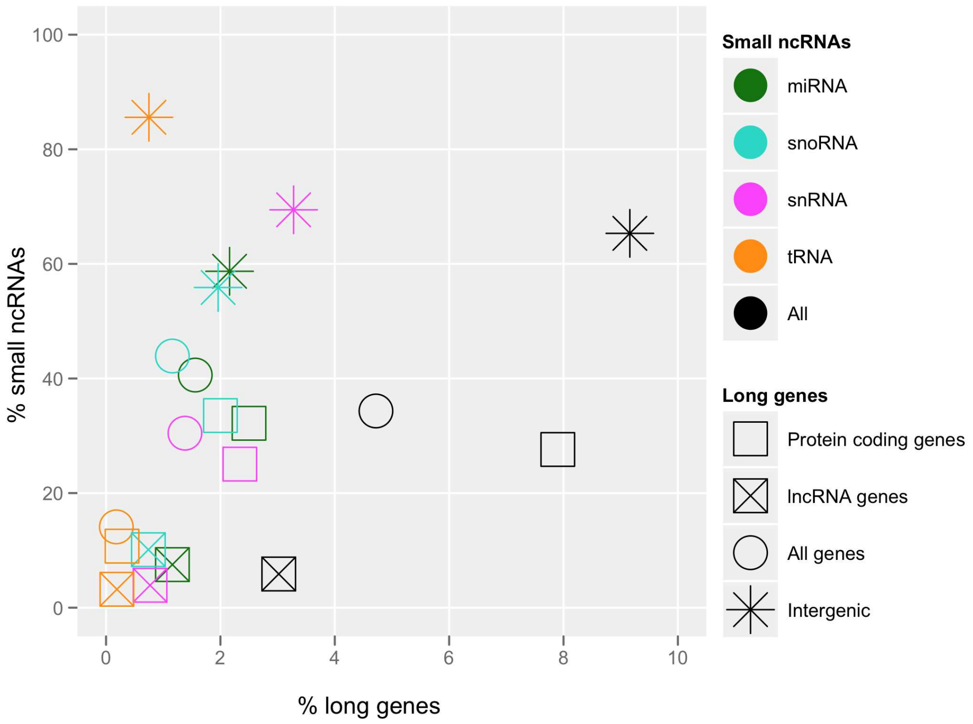
Supplementary Figure S14

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.



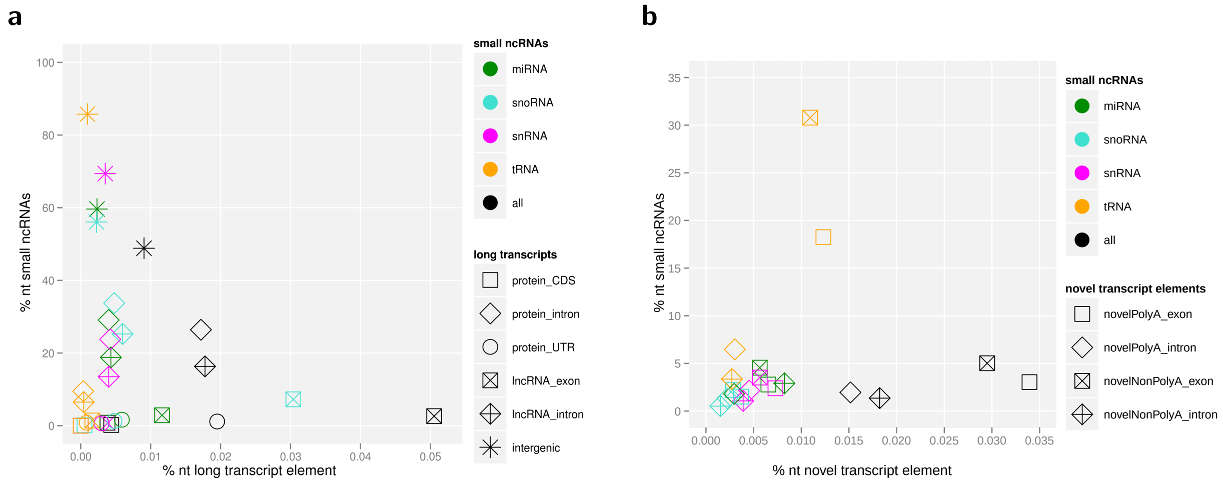
Supplementary Figure S15

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 prime ends in the nucleus and the cytosol. The coverage is calculated as the number of short RNA-seq reads / CAGE tags most 5 prime ends which fall at a given distance from the annotated small transcript 5 prime 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 section III for details).



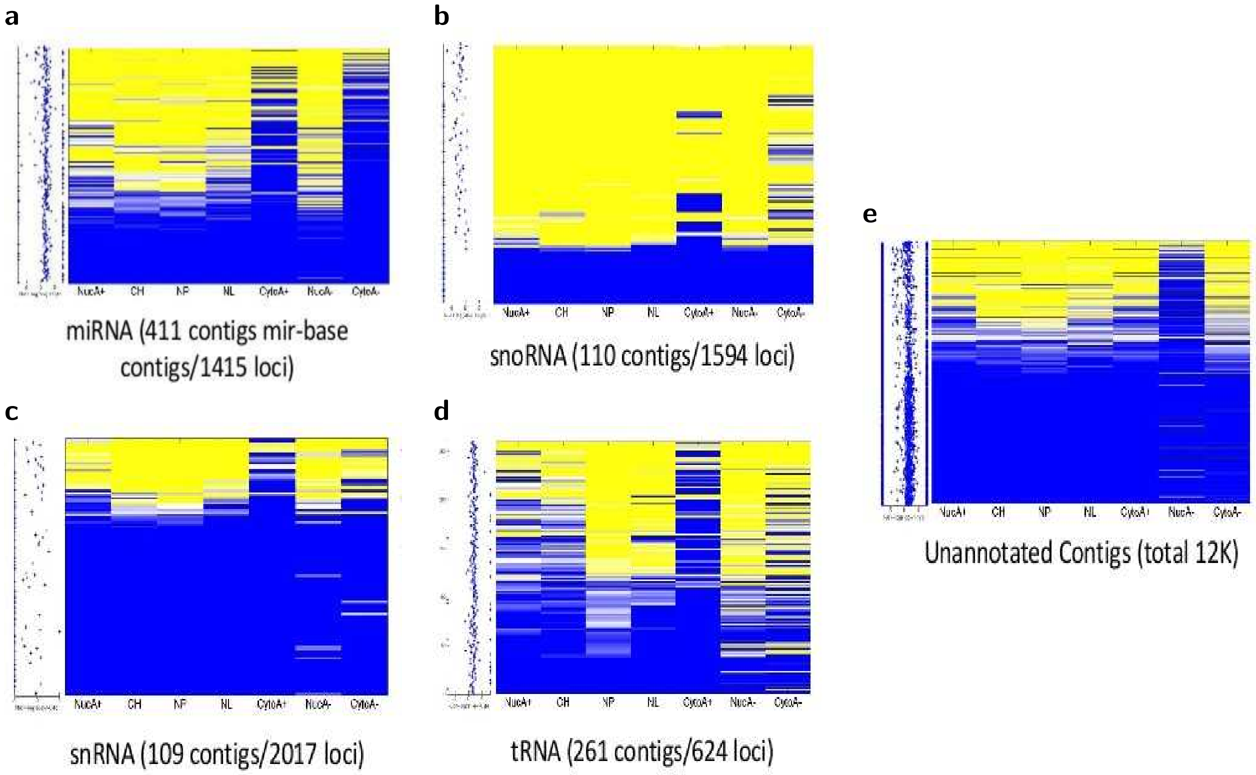
Supplementary Figure S16

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



Supplementary Figure S17

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.



miRNA (411 contigs mir-base contigs/1415 loci)

snoRNA (110 contigs/1594 loci)

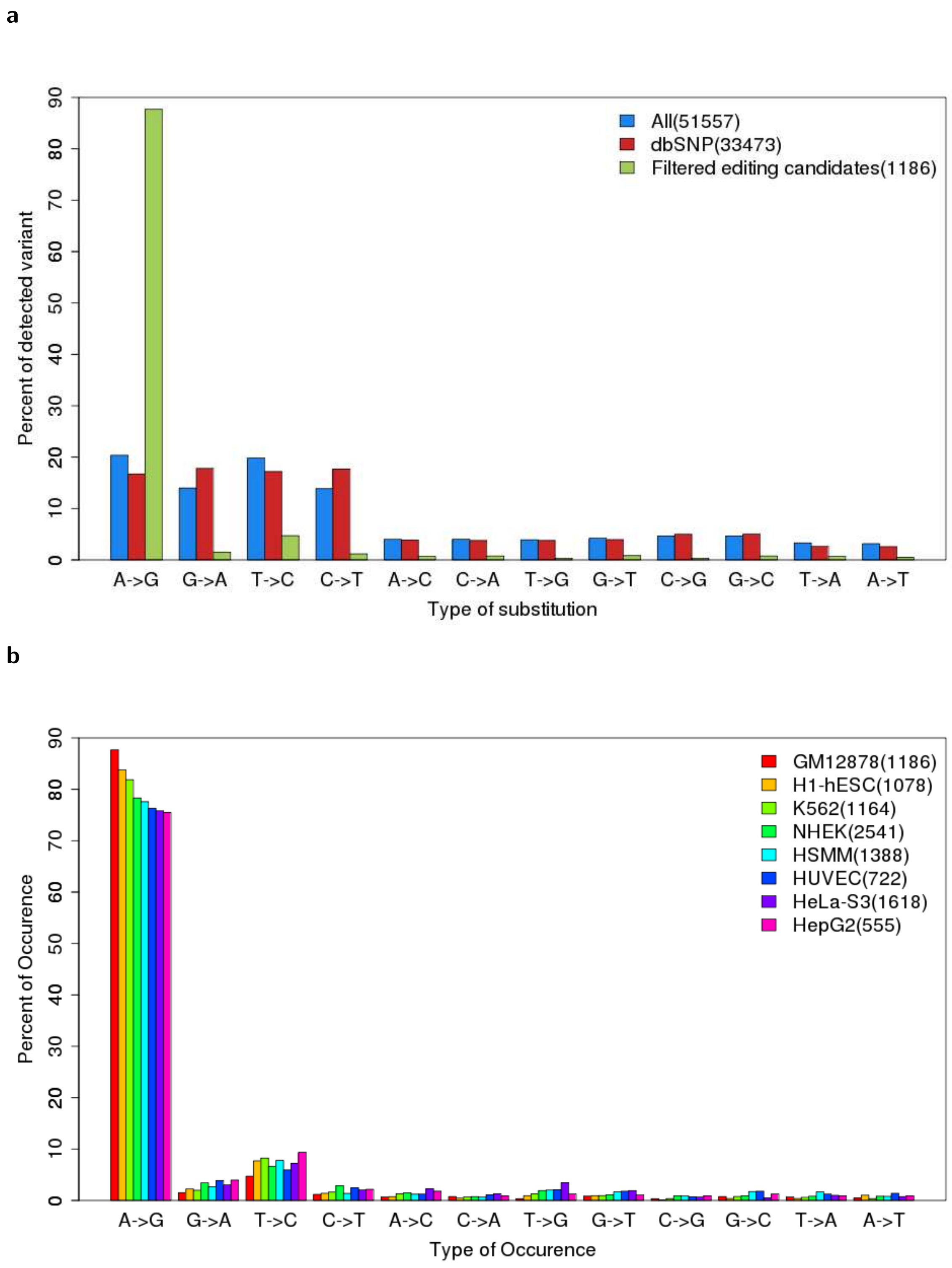
snRNA (109 contigs/2017 loci)

tRNA (261 contigs/624 loci)

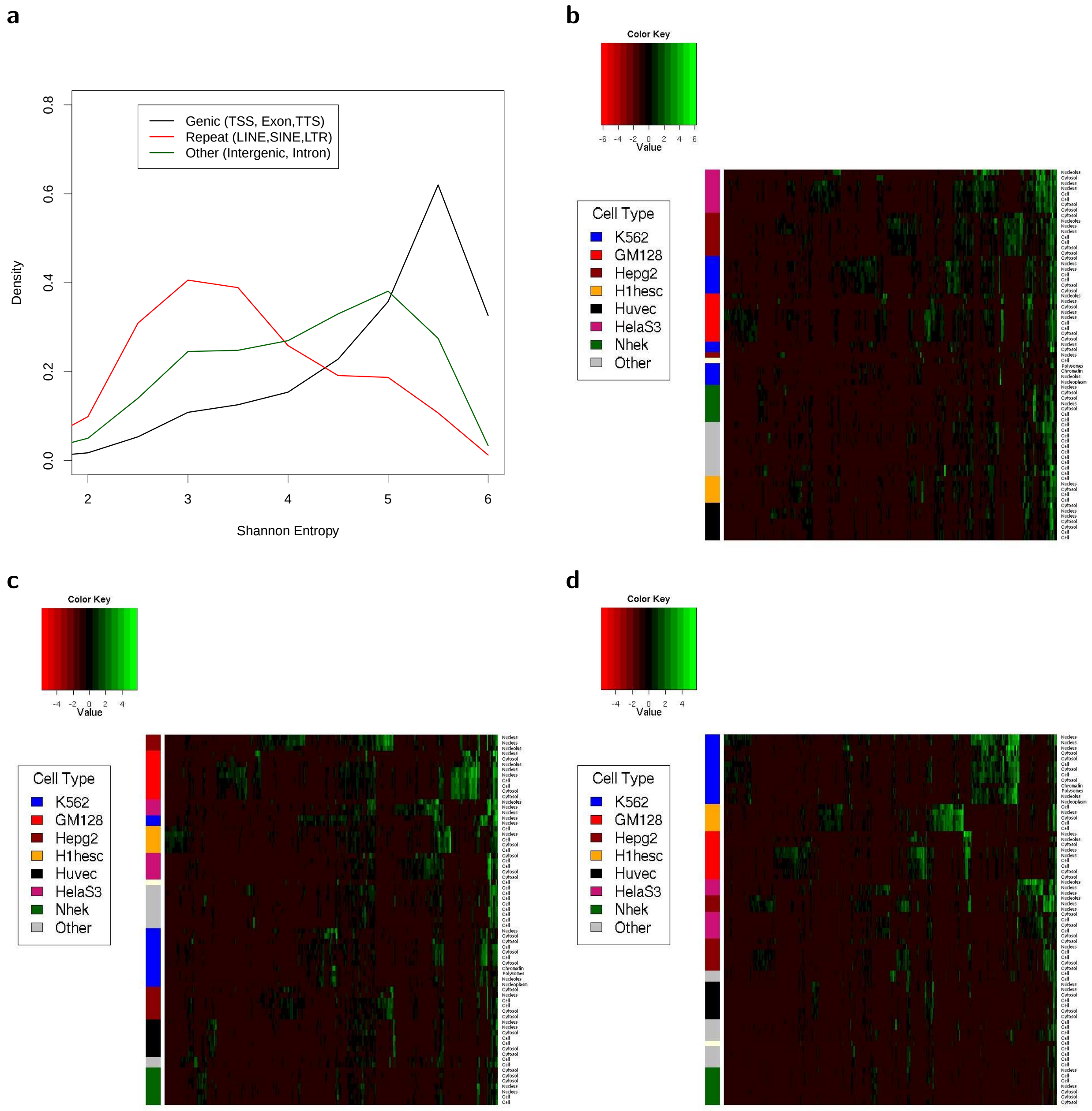
Unannotated Contigs (total 12K)

Supplementary Figure S18

Expression of long RNA contigs corresponding to detected short RNA. The expression of the long RNA contigs which corresponding to detected short RNA are color-coded in the heatmaps. Blue indicates no expression yellow indicates high expression. The log-ratio of detected short RNA expression in cytoplasm over nucleus is shown in the scatter plot on the left side. Cytoplasm enriched short RNA contigs are distributed to the right side of zero while nucleus enriched short RNA contigs to the left. a. for detected miRNA contigs; b. for snoRNA; c. for snRNA; d. for tRNA and e. for total unannotated short RNA.

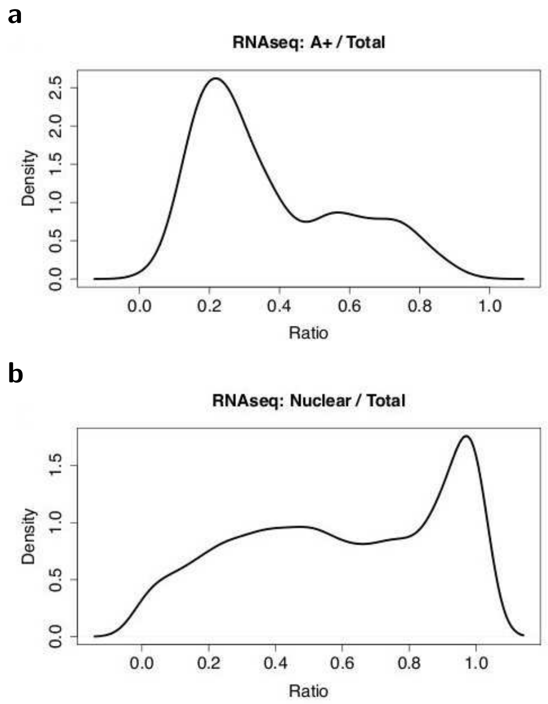


Supplementary Figure S19
Profile of RNA editing in ENCODE whole-cell datasets and compartments. a. The profile of RNA-detected single nucleotide variants (SNV) in GM12878 that are detected independently in both the Caltech whole-cell polyA-selected non-stranded dataset and the CSHL stranded dataset, 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.



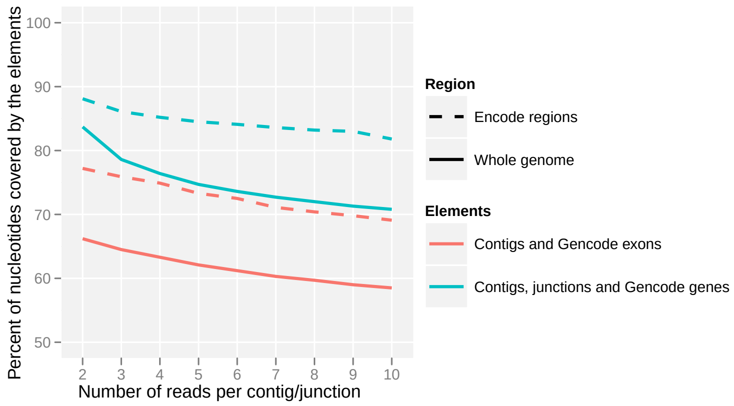
Supplementary Figure S20

Cell line specific expression of repeat elements. a. Shannon Entropy of CAGE cluster expression profiles across all experiments separated by broad annotation categories. A low entropy means a narrow expression across experiments, thus intergenic LINE, SINE and LTR repeats are noticeably more narrowly expressed than other genomic categories. Heatmaps of b. LINE c. SINE and d. LTR 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.



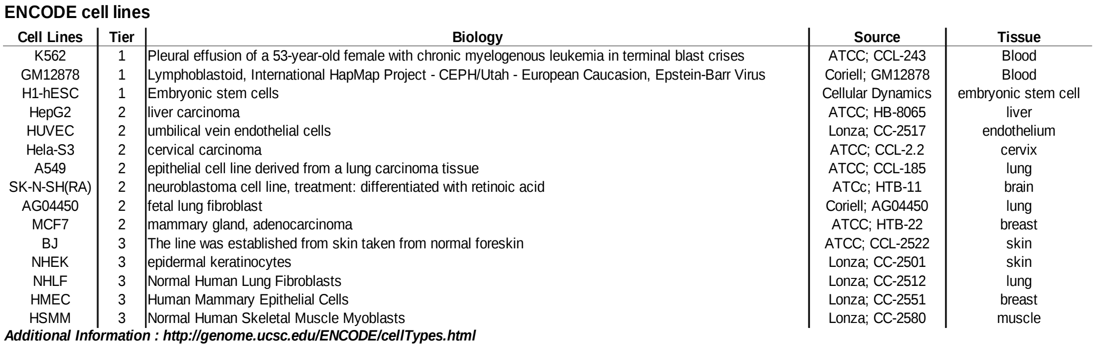
Supplementary Figure S21

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.



Supplementary Figure S22

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.

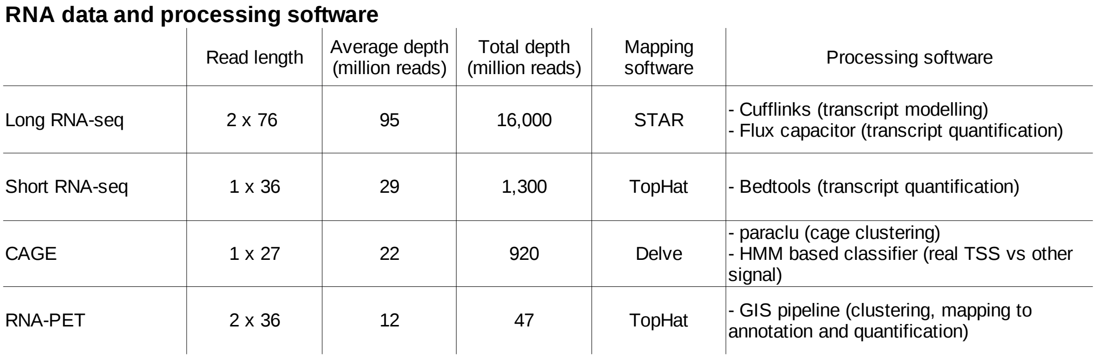


ENCODE cell lines

Cell Lines	Tier	Biology	Source	Tissue
K562	1	Pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises	ATCC; CCL-243	Blood
GM12878	1	Lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasion, Epstein-Barr Virus	Coriell; GM12878	Blood
H1-hESC	1	Embryonic stem cells	Cellular Dynamics	embryonic stem cell
HepG2	2	liver carcinoma	ATCC; HB-8065	liver
HUVEC	2	umbilical vein endothelial cells	Lonza; CC-2517	endothelium
Hela-S3	2	cervical carcinoma	ATCC; CCL-2.2	cervix
A549	2	epithelial cell line derived from a lung carcinoma tissue	ATCC; CCL-185	lung
SK-N-SH(RA)	2	neuroblastoma cell line, treatment: differentiated with retinoic acid	ATCc; HTB-11	brain
AG04450	2	fetal lung fibroblast	Coriell; AG04450	lung
MCF7	2	mammary gland, adenocarcinoma	ATCC; HTB-22	breast
BJ	3	The line was established from skin taken from normal foreskin	ATCC; CCL-2522	skin
NHEK	3	epidermal keratinocytes	Lonza; CC-2501	skin
NHLF	3	Normal Human Lung Fibroblasts	Lonza; CC-2512	lung
HMEC	3	Human Mammary Epithelial Cells	Lonza; CC-2551	breast
HSMM	3	Normal Human Skeletal Muscle Myoblasts	Lonza; CC-2580	muscle

Additional Information : http://genome.ucsc.edu/ENCODE/cellTypes.html

Supplementary 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.



RNA data and processing software

	Read length	Average depth (million reads)	Total depth (million reads)	Mapping software	Processing software
Long RNA-seq	2 x 76	95	16,000	STAR	- Cufflinks (transcript modelling) - Flux capacitor (transcript quantification)
Short RNA-seq	1 x 36	29	1,300	TopHat	- Bedtools (transcript quantification)
CAGE	1 x 27	22	920	Delve	- paraclu (cage clustering) - HMM based classifier (real TSS vs other signal)
RNA-PET	2 x 36	12	47	TopHat	- GIS pipeline (clustering, mapping to annotation and quantification)

Supplementary Table S2
RNA data and processing software.

a. Polyadenylated RNAs

1. Expression of Gencode (v7) annotated elements

Gene type	Detected exons[2] (annotation #)	Detected splice junctions[2] (annotation #)	Detected transcripts[2] (annotation #)	Detected genes[2] (annotation #)	Exon nucleotide coverage[3] (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only 1 cell line	Proportion over genes expressed (%)	Number of genes expressed in 14 cell lines	Proportion over genes expressed (%)
Long non coding	20,450 (41,467)	7,917 (26,872)	6,205 (14,880)	5,668 (9,277)	85.5	5,668	1,390	24.5	125	2.2
Protein coding	284,081 (318,514)	194,192 (244,158)	58,485 (76,006)	18,842 (20,679)	97.9	18,842	1,129	6.0	10,366	55.0
Other[1]	96,614 (133,937)	19,026 (47,663)	43,334 (71,113)	9,312 (21,750)	94.7	9,312	2,297	24.7	2,104	22.6
Total annotated	401,145 (493,918)	221,135 (318,693)	108,024 (161,999)	33,822 (51,706)	96.4	33,822	4,816	14.2	12,595	37.2

2. Expression of Gencode (v7) intergenic and antisense elements

Category	Detected exons[2]	Detected splice junctions[2]	Detected transcripts[2]	Detected genes[2]
Mono-exonic	18,932	NA	18,931	17,387
Multi-exonic	29,147	54,899	13,357	6,292
Total	48,079	54,899	32,288	23,679

b. Non-Polyadenylated RNAs

1. Expression of Gencode (v7) annotated elements

Gene type	Detected exons[2] (annotation #)	Detected splice junctions[2] (annotation #)	Detected transcripts[2] (annotation #)	Detected genes[2] (annotation #)	Exon nucleotide coverage[3] (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only 1 cell line	Proportion over genes expressed (%)	Number of genes expressed in 14 cell lines	Proportion over genes expressed (%)
Long non coding	15,653 (41,467)	3,918 (26,872)	3,397 (14,880)	3,952 (9,277)	87.9	3,952	1,365	34.5	125	3.2
Protein coding	261,410 (318,514)	168,039 (244,158)	42,652 (76,006)	17,297 (20,679)	97	17,297	1,577	9.1	6,950	40.2
Other[1]	85,590 (133,937)	9,731 (47,663)	31,672 (71,113)	7,398 (21,750)	95	7,398	2,085	28.2	724	9.8
Total annotated	362,653 (493,918)	181,688 (318,693)	77,721 (161,999)	28,647 (51,706)	96.1	28,647	5,027	17.5	7,799	27.2

2. Expression of Gencode (v7) intergenic and antisense elements

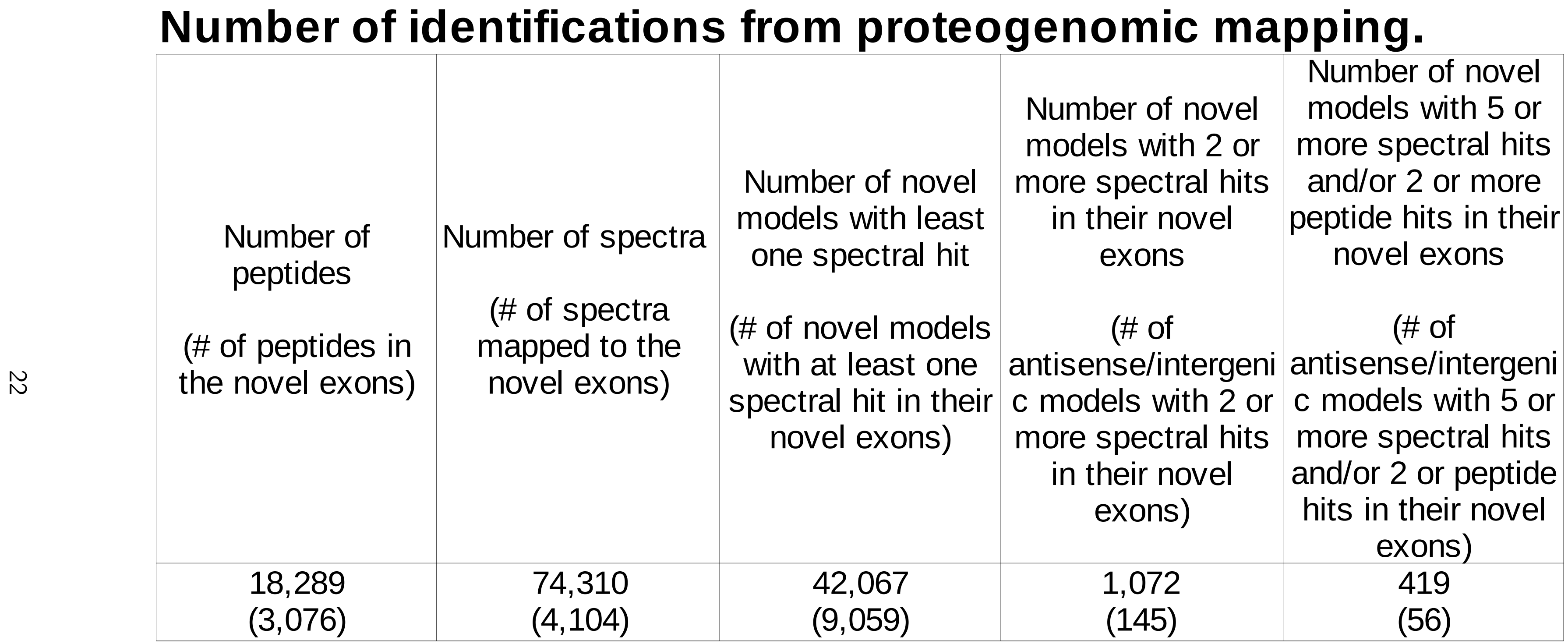
Category	Detected exons[2]	Detected splice junctions[2]	Detected transcripts[2]	Detected genes[2]
Mono-exonic	47,503	NA	47,503	45,409
Multi-exonic	10,674	27,178	4,791	3,290
Total	58,177	27,178	52,294	48,699

[1] includes pseudogenes, miRNAs, etc

[2] all elements that passed npIDR (0.1)

[3] cumulative detected nucleotide in detected exons / total nucleotides in detected exons

Supplementary Table S3
a. Polyadenylated and b. non-polyadenylated RNAs.



Number of identifications from proteogenomic mapping.

Number of peptides (# of peptides in the novel exons)	Number of spectra (# of spectra mapped to the novel exons)	Number of novel models with least one spectral hit (# of novel models with at least one spectral hit in their novel exons)	Number of novel models with 2 or more spectral hits in their novel exons (# of antisense/intergenic models with 2 or more spectral hits in their novel exons)	Number of novel models with 5 or more spectral hits and/or 2 or more peptide hits in their novel exons (# of antisense/intergenic models with 5 or more spectral hits and/or 2 or peptide hits in their novel exons)
18,289 (3,076)	74,310 (4,104)	42,067 (9,059)	1,072 (145)	419 (56)

Supplementary 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.

K562 nuclear sub-compartments (total RNA)

1. Expression of Gencode (v7) annotated elements

Gene type	Detected exons[2] (annotation #)	Detected splice junctions[2] (annotation #)	Detected transcripts[2] (annotation #)	Detected genes[2] (annotation #)	Exon nucleotide coverage[3] (%)
Long non coding	8,109 (41,467)	1,644 (26,872)	1,903 (14,880)	2,032 (9,277)	79
Protein coding	167,711 (318,514)	109,253 (244,158)	21,661 (76,006)	12,344 (20,679)	96.3
Other[1]	53,877 (133,937)	5,260 (47,663)	18,630 (71,113)	3,954 (21,750)	93.1
Total annotated	229,697 (493,918)	116,157 (318,693)	42,194 (161,999)	18,330 (51,706)	94.7

[1] includes pseudogenes, miRNAs, etc

[2] all elements that passed npIDR (0.1)

[3] cumulative detected nucleotide in detected exons / total nucleotides in detected exons

2. Expression of Gencode (v7) intergenic and antisense elements

Category	Detected exons[4]	Detected splice junctions[4]	Detected transcripts[4]	Detected genes[4]
Mono-exonic	40,319	NA	40,273	39,327
Multi-exonic	6,014	14,374	2,570	1,791
Total	46,333	14,374	42,843	41,118

[4] all elements that passed npIDR (0.1)

Supplementary Table S5
K562 nuclear subcompartments (total RNA).

K562 nuclear sub-compartment specific elements

1. Gencode (v7) annotated genes

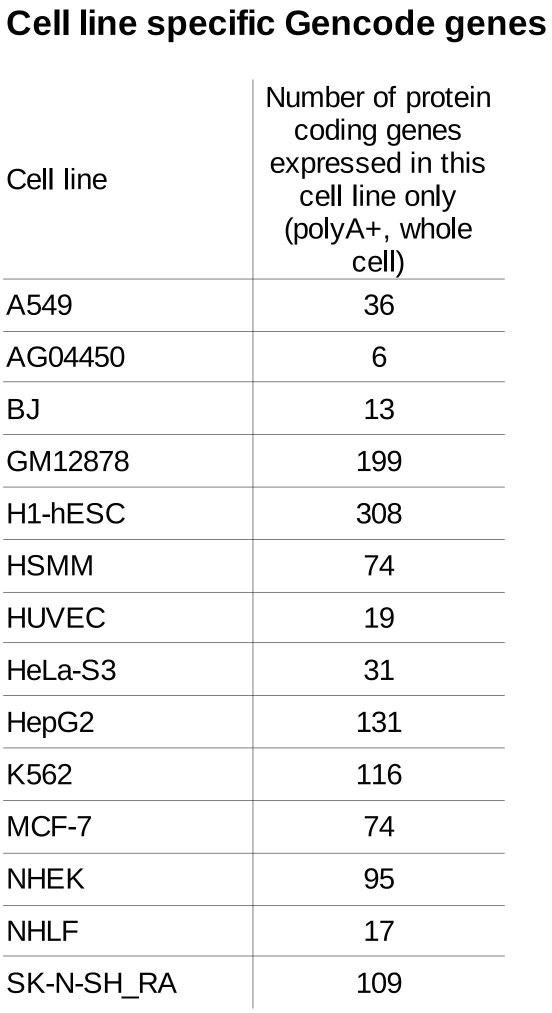
Cell compartment	Detected exons			Detected splice junctions			Detected transcripts			Detected genes		
	All	Unique to cell compartment		All	Unique to cell compartment		All	Unique to cell compartment		All	Unique to cell compartment	
		#	%		#	%		#	%		#	%
Nucleolus	170,484	3,177	1.9	101,665	135	0.1	27,754	1,387	5.0	12,940	826	6.4
Chromatin	189,818	1,093	0.6	110,578	422	0.4	28,047	333	1.2	16,604	145	0.9
Nucleoplasm	211,020	4,956	2.3	99,765	34	0.0	32,764	694	2.1	15,935	96	0.6

2. Gencode (v7) intergenic and antisense regions

Cell compartment	Detected exons			Detected splice junctions			Detected transcripts			Detected genes		
	All	Unique to cell compartment		All	Unique to cell compartment		All	Unique to cell compartment		All	Unique to cell compartment	
		#	%		#	%		#	%		#	%
Nucleolus	43,024	48	0.1	5,569	847	15.2	3,562	8	0.2	38,326	42	0.1
Chromatin	45,587	140	0.3	7,136	1,299	18.2	42,346	398	0.9	40,729	133	0.3
Nucleoplasm	45,549	148	0.3	11,020	2,248	20.4	42,389	441	1.0	40,827	141	0.3

Supplementary Table S6

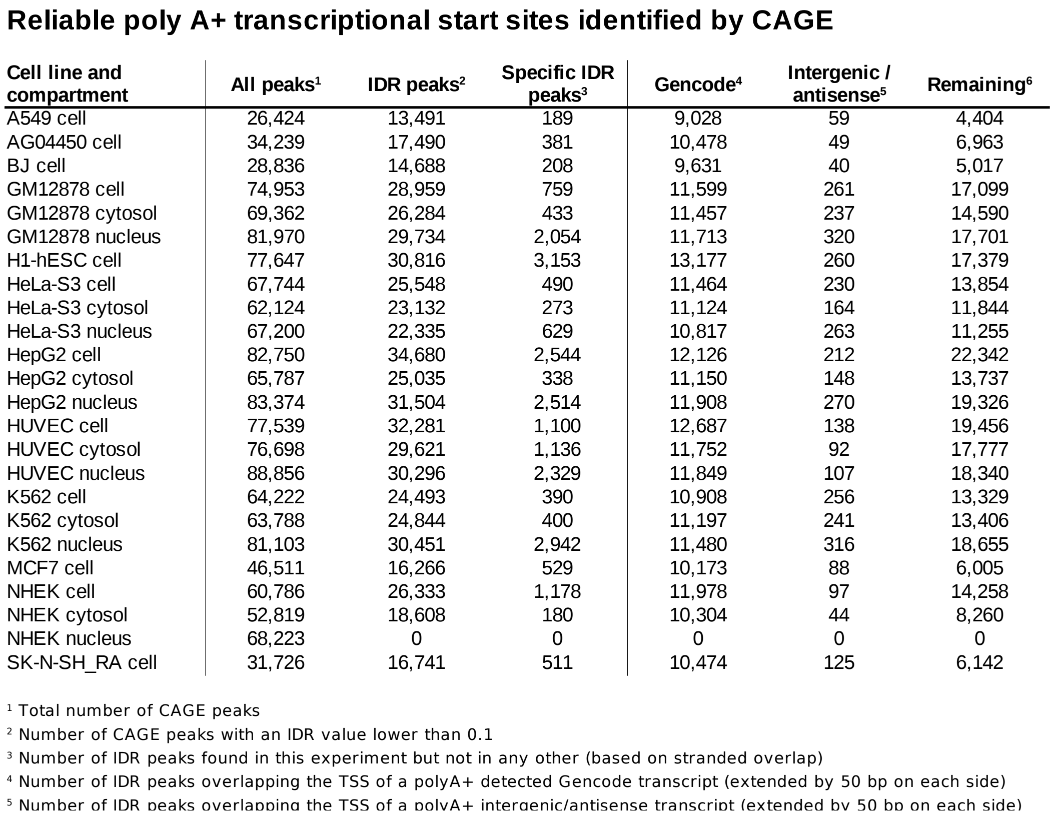
Elements specific to K562 nuclear subcompartments. This table shows the total and unique number of elements (exons, splice junctions, transcripts and genes) detected in the K562 nuclear subcompartments. (1) Annotated elements and (2) Novel intergenic / antisense elements.



Cell line specific Gencode genes

Cell line	Number of protein coding genes expressed in this cell line only (polyA+, whole cell)
A549	36
AG04450	6
BJ	13
GM12878	199
H1-hESC	308
HSMM	74
HUVEC	19
HeLa-S3	31
HepG2	131
K562	116
MCF-7	74
NHEK	95
NHLF	17
SK-N-SH_RA	109

Supplementary Table S7
Cell line specific Gencode genes. Annotated protein coding genes expressed in the different cells lines.

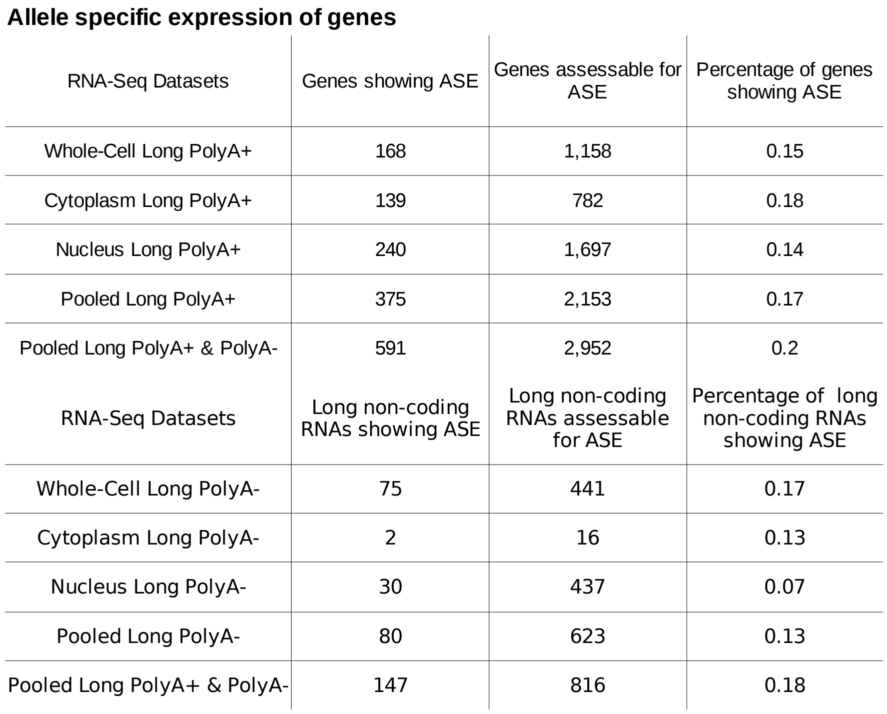


Reliable poly A+ transcriptional start sites identified by CAGE

Cell line and compartment	All peaks[1]	IDR peaks[2]	Specific IDR peaks[3]	Gencode[4]	Intergenic / antisense[5]	Remaining[6]
A549 cell	26,424	13,491	189	9,028	59	4,404
AG04450 cell	34,239	17,490	381	10,478	49	6,963
BJ cell	28,836	14,688	208	9,631	40	5,017
GM12878 cell	74,953	28,959	759	11,599	261	17,099
GM12878 cytosol	69,362	26,284	433	11,457	237	14,590
GM12878 nucleus	81,970	29,734	2,054	11,713	320	17,701
H1-hESC cell	77,647	30,816	3,153	13,177	260	17,379
HeLa-S3 cell	67,744	25,548	490	11,464	230	13,854
HeLa-S3 cytosol	62,124	23,132	273	11,124	164	11,844
HeLa-S3 nucleus	67,200	22,335	629	10,817	263	11,255
HepG2 cell	82,750	34,680	2,544	12,126	212	22,342
HepG2 cytosol	65,787	25,035	338	11,150	148	13,737
HepG2 nucleus	83,374	31,504	2,514	11,908	270	19,326
HUVEC cell	77,539	32,281	1,100	12,687	138	19,456
HUVEC cytosol	76,698	29,621	1,136	11,752	92	17,777
HUVEC nucleus	88,856	30,296	2,329	11,849	107	18,340
K562 cell	64,222	24,493	390	10,908	256	13,329
K562 cytosol	63,788	24,844	400	11,197	241	13,406
K562 nucleus	81,103	30,451	2,942	11,480	316	18,655
MCF7 cell	46,511	16,266	529	10,173	88	6,005
NHEK cell	60,786	26,333	1,178	11,978	97	14,258
NHEK cytosol	52,819	18,608	180	10,304	44	8,260
NHEK nucleus	68,223	0	0	0	0	0
SK-N-SH_RA cell	31,726	16,741	511	10,474	125	6,142

[1] Total number of CAGE peaks
[2] Number of CAGE peaks with an IDR value lower than 0.1
[3] Number of IDR peaks found in this experiment but not in any other (based on stranded overlap)
[4] Number of IDR peaks overlapping the TSS of a polyA+ detected Gencode transcript (extended by 50 bp on each side)
[5] Number of IDR peaks overlapping the TSS of a polyA+ intergenic/antisense transcript (extended by 50 bp on each side)

Supplementary 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.



Allele specific expression of genes

RNA-Seq Datasets	Genes showing ASE	Genes assessable for ASE	Percentage of genes showing ASE
Whole-Cell Long PolyA+	168	1,158	0.15
Cytoplasm Long PolyA+	139	782	0.18
Nucleus Long PolyA+	240	1,697	0.14
Pooled Long PolyA+	375	2,153	0.17
Pooled Long PolyA+ & PolyA-	591	2,952	0.2
RNA-Seq Datasets	Long non-coding RNAs showing ASE	Long non-coding RNAs assessable for ASE	Percentage of long non-coding RNAs showing ASE
Whole-Cell Long PolyA-	75	441	0.17
Cytoplasm Long PolyA-	2	16	0.13
Nucleus Long PolyA-	30	437	0.07
Pooled Long PolyA-	80	623	0.13
Pooled Long PolyA+ & PolyA-	147	816	0.18

Supplementary Table S9
Allele specific expression of genes. 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. 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.

Genome coverage

					whole genome, scaled to exclude gaps				
Read per junction/contig	2	3	4	5	6	7	8	9	10
contigs	66.0	64.2	63.0	61.8	60.9	60.0	59.3	58.7	58.1
introns	73.8	62.8	57.4	53.9	51.2	49.2	47.6	46.2	44.9
intersection introns+contigs	58.8	52.7	49.1	46.6	44.6	43.1	41.8	40.8	39.8
union introns+contigs	80.9	74.4	71.3	69.1	67.5	66.1	65.1	64.1	63.2
contigs + Gencode exons	66.2	64.5	63.3	62.1	61.2	60.3	59.7	59.0	58.5
contigs+introns+Gencode genes	83.7	78.6	76.4	74.7	73.6	72.7	72.0	71.3	70.8

					encode regions				
Read per junction/contig	2	3	4	5	6	7	8	9	10
contigs	77.0	75.7	74.7	73.2	72.3	71.0	70.2	69.6	68.8
introns	79.3	74.0	70.7	68.6	67.4	59.8	59.2	58.6	54.2
intersection introns+contigs	70.2	66.8	63.4	61.2	59.9	53.1	52.6	52.0	49.3
union introns+contigs	86.0	82.9	81.9	80.6	79.7	77.7	76.8	76.2	73.7
contigs + Gencode exons	77.2	75.9	74.9	73.3	72.5	71.1	70.4	69.8	69.1
contigs+introns+Gencode genes	88.1	86.1	85.2	84.5	84.1	83.6	83.2	83.0	81.8

Supplementary Table S10
Genome Coverage. Companion data for figure S23.