Landscape of transcription in human cells

S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokicinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, H. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Anonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, & T. R. Gingeras.

Nature **489** (2012) 101–108.

Speaker: Joseph Chuang-Chieh Lin

The Comparative & Evolutionary Genomics/Transcriptomics Lab.
Genomics Research Center, Academia Sinica
Taiwan

30 January 2013

Outline

1. Motivation
2. RNA data generation
3. Results
   - Long RNA expression landscape
   - Short RNA expression landscape
   - RNA editing & allele-specific expression
   - Repeat region transcription
   - Characterization of enhancer RNA

Motivation

- 2002–2007: The pilot phase of the ENCODE project:
  - Examine 1% of the human genome.
- 2007–2012: The second phase of the ENCODE project:
  - Interrogate the complete human genome.
- Goal of the paper:
  - Provide a genome-wide catalogue of the produced RNAs.
  - Identify the subcellular localization for the produced RNAs.

Motivation

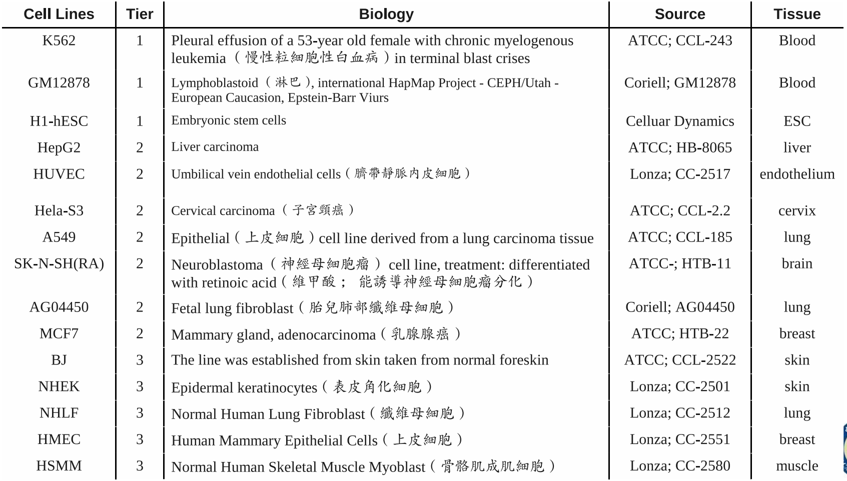
- 2002–2007: The pilot phase of the ENCODE project:
  - Examine 1% of the human genome.
- 2007–2012: The second phase of the ENCODE project:
  - Interrogate the complete human genome.
- Goal of the paper:
  - Provide a genome-wide catalogue of the produced RNAs.
  - Identify the subcellular localization for the produced RNAs.

Motivation

- 2002–2007: The pilot phase of the ENCODE project:
  - Examine 1% of the human genome.
- 2007–2012: The second phase of the ENCODE project:
  - Interrogate the complete human genome.
- Goal of the paper:
  - Provide a genome-wide catalogue of the produced RNAs.
  - Identify the subcellular localization for the produced RNAs.

Materials

- 15 ENCODE cell lines:

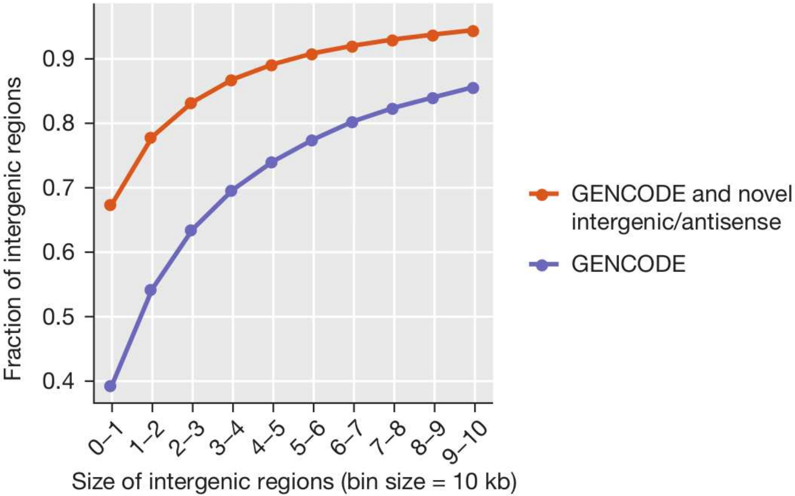


Cell Lines	Tier	Biology	Source	Tissue
K562	1	Pleural effusion of a 53-year old female with chronic myelogenous leukemia（慢性粒細胞性白血病）in terminal blast crises	ATCC; CCL-243	Blood
GM12878	1	Lymphoblastoid（淋巴）, international HapMap Project - CEPH/Utah - European Caucasion, Epstein-Barr Viurs	Coriell; GM12878	Blood
H1-hESC	1	Embryonic stem cells	Celluar Dynamics	ESC
HepG2	2	Liver carcinoma	ATCC; HB-8065	liver
HUVEC	2	Umbilical vein endothelial cells（臍帶靜脈內皮細胞）	Lonza; CC-2517	endothelium
Hela-S3	2	Cervical carcinoma（子宮頸癌）	ATCC; CCL-2.2	cervix
A549	2	Epithelial（上皮細胞）cell line derived from a lung carcinoma tissue	ATCC; CCL-185	lung
SK-N-SH(RA)	2	Neuroblastoma（神經母細胞瘤）cell line, treatment: differentiated with retinoic acid（維甲酸； 能誘導神經母細胞瘤分化）	ATCC-; HTB-11	brain
AG04450	2	Fetal lung fibroblast（胎兒肺部纖維母細胞）	Coriell; AG04450	lung
MCF7	2	Mammary gland, adenocarcinoma（乳腺腺癌）	ATCC; HTB-22	breast
BJ	3	The line was established from skin taken from normal foreskin	ATCC; CCL-2522	skin
NHEK	3	Epidermal keratinocytes（表皮角化細胞）	Lonza; CC-2501	skin
NHLF	3	Normal Human Lung Fibroblast（纖維母細胞）	Lonza; CC-2512	lung
HMEC	3	Human Mammary Epithelial Cells（上皮細胞）	Lonza; CC-2551	breast
HSMM	3	Normal Human Skeletal Muscle Myoblast（骨骼肌成肌細胞）	Lonza; CC-2580	muscle

Contribution of this paper

- Extend the current genome-wide annotated catalogue of long poly-adenylated & small RNAs of GENCODE.
- 62.1% and 74.7% of the human genome are covered by either processed or primary transcripts.
  - ⋆ **primary** = contigs + introns + GENCODE genes;
  - ⋆ **processed** = contigs + GENCODE exons.

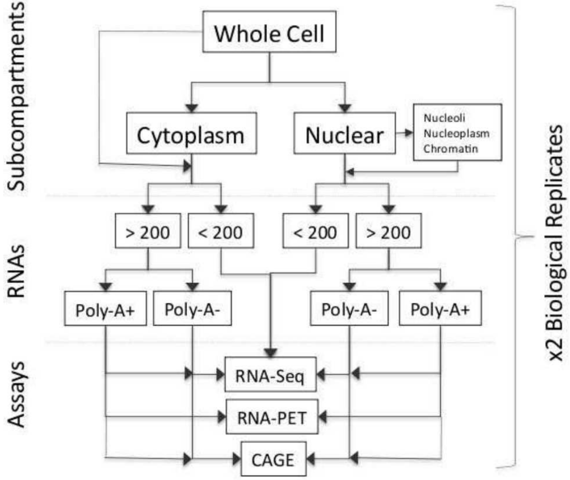
  - No cell line shows more than 56.7% of the union of the expressed transcriptomes across all cell lines.
  - The consequent reduction in the length of *intergenic regions* leads to a significant overlapping of neighbouring genic regions and prompts a redefinition of a gene.

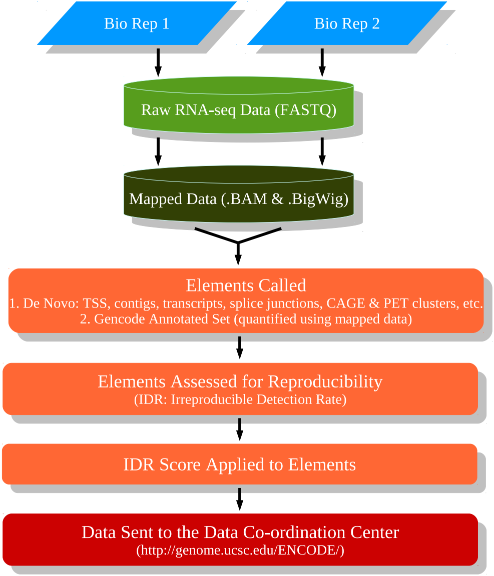

0.9
0.8
0.7
0.6
0.5
0.4
Fraction of intergenic regions
0–1
1–2
2–3
3–4
4–5
5–6
6–7
7–8
8–9
9–10
Size of intergenic regions (bin size = 10 kb)
GENCODE and novel intergenic/antisense
GENCODE

Contribution of this paper (contd.)

- A tendency for genes to express many isoforms simultaneously (plateau: 10–12 expressed isoforms per gene per cell line).
- Cell-type-specific enhancers are promoters differentiable from other regulatory regions.
- Coding & non-coding transcripts are predominantly localized in the cytosol and nucleus, resp.
- $\approx$ 6% of all annotated coding and non-coding transcripts overlap with small RNAs and probably precursors to these small RNAs.

Sample Flowchart





Bio Rep 1

Bio Rep 2

Raw RNA-seq Data (FASTQ)

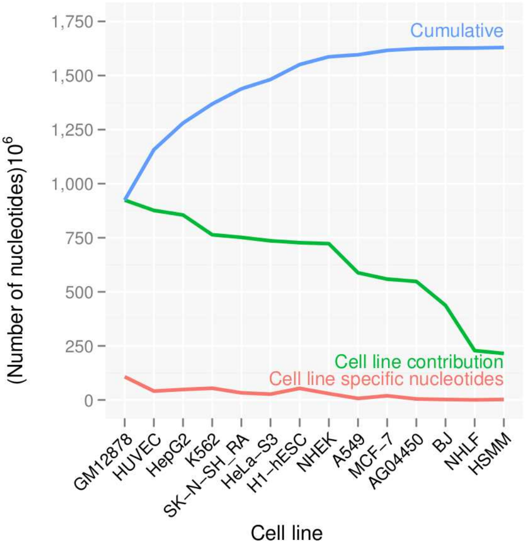
Mapped Data (.BAM & .BigWig)

Elements Called
1. De Novo: TSS, contigs, transcripts, splice junctions, CAGE & PET clusters, etc.
2. Gencode Annotated Set (quantified using mapped data)

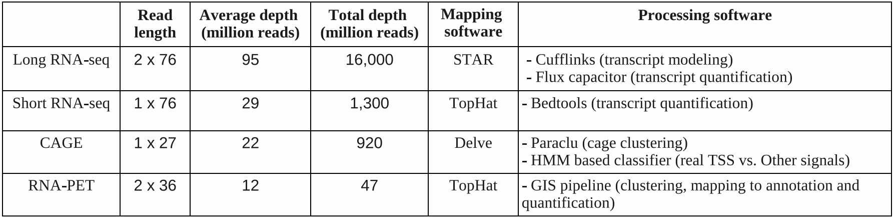
Elements Assessed for Reproducibility
(IDR: Irreproducible Detection Rate)

IDR Score Applied to Elements

Data Sent to the Data Co-ordination Center
(http://genome.ucsc.edu/ENCODE/)


Cumulative
Cell line contribution
Cell line specific nucleotides
(Number of nucleotides)10^6
1,750
1,500
1,250
1,000
750
500
250
0
GM12878
HUVEC
HepG2
K562
SK–N–SH_RA
HeLa–S3
H1–hESC
NHEK
A549
MCF–7
AG04450
BJ
NHLF
HSMM
Cell line

RNA data & 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 modeling) - Flux capacitor (transcript quantification)
Short RNA-seq	1 x 76	29	1,300	TopHat	- Bedtools (transcript quantification)
CAGE	1 x 27	22	920	Delve	- Paraclu (cage clustering) - HMM based classifier (real TSS vs. Other signals)
RNA-PET	2 x 36	12	47	TopHat	- GIS pipeline (clustering, mapping to annotation and quantification)

- The mapped data was used to *assemble* and *quantify* annotated GENCODE v7 elements.
- Elements and quantifications were further assessed for reproducibility between replicates using a non-parametric version of *IDR*.

Long RNA expression landscape

Detection of annotated & novel transcripts

Expression of GENCODE (v7) annotated elements (a)

Gene type	Detected exons† (annotation no.)	Detected splice junctions† (annotation no.)	Detected transcripts† (annotation no.)	Detected genes† (annotation no.)	Exon nucleotide coverage‡ (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only one cell line	Proportion over genes expressed§ (%)	Number of genes expressed in 14 cell lines	Proportion over genes expressed‖ (%)
Long non-coding	22,381 (41,467)	8,017 (26,872)	6,521 (14,880)	5,906 (9,277)	87.5	5,906	1,386	23.5	631	10.7
Protein coding	288,322 (318,514)	194,752 (244,158)	59,822 (76,006)	18,939 (20,679)	98.1	18,939	1,082	5.7	10,571	55.8
Other*	102,000 (133,937)	19,277 (47,663)	45,410 (71,113)	10,649 (21,750)	95.2	10,649	2,453	23.0	1,896	17.8
Total annotated	412,703 (493,918)	222,046 (318,693)	111,753 (161,999)	35,494 (51,706)	96.7	35,394	4,921	13.9	13,098	37.0

NA, not applicable.

* Includes pseudogenes, miRNAs, etc.

† All elements that passed npIDR (0.1).

‡ Cumulative detected nucleotide in detected exons/total nucleotides in detected exons.

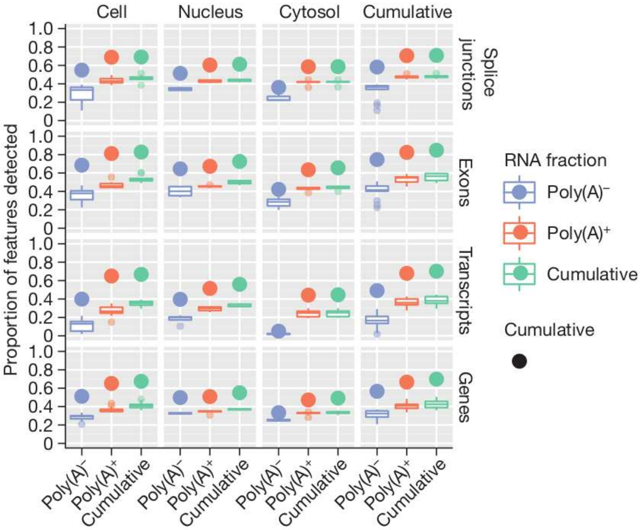
§ Proportion for genes expressed in only one cell line.

‖ Proportion for genes expressed in 14 cell lines.

Detection of annotated & novel transcripts

- 70% of annotated splice junctions, transcripts and genes were cumulatively detected.
  - ≈ 85% of annotated exons with an average of coverage 96% (by RNA-seq).
- Small variation in the proportion of detected GENCODE elements.
- Only a small proportion of GENCODE elements are detected *exclusively in the Poly-A$^-$ RNA fraction*.

A large majority of GENCODE elements are detected by RNA-seq data



Detection of annotated & novel transcripts (contd.)

- The identified novel elements covered 78% of the intronic nucleotides and 34% of the intergenic sequences.
- Use **Cufflinks** to predict (over all long RNA-seq samples) the following elements in intergenic and antisense regions:
    - 94,800 exons (↑ 19%);
    - 69,052 splice junctions (↑ 22%); ⇒ mono-exonic transcripts?
    - 73,325 transcripts (↑ 45%);
    - 41,204 genes (↑ 80%). ⇒ mono-exonic transcripts?

  ⋆ DNA contamination or incomplete determination of transcript structures?

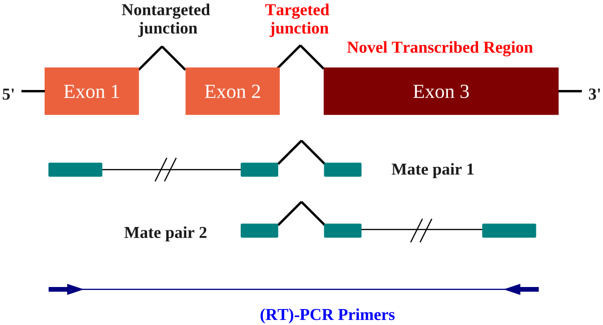
Detection of annotated & novel transcripts (contd.)

- The identified novel elements covered 78% of the intronic nucleotides and 34% of the intergenic sequences.
- Use **Cufflinks** to predict (over all long RNA-seq samples) the following elements in intergenic and antisense regions:
  - 94,800 exons (↑ 19%);
  - 69,052 splice junctions (↑ 22%); ⇒ mono-exonic transcripts?
  - 73,325 transcripts (↑ 45%);
  - 41,204 genes (↑ 80%). ⇒ mono-exonic transcripts?
  - ⋆ DNA contamination or incomplete determination of transcript structures?

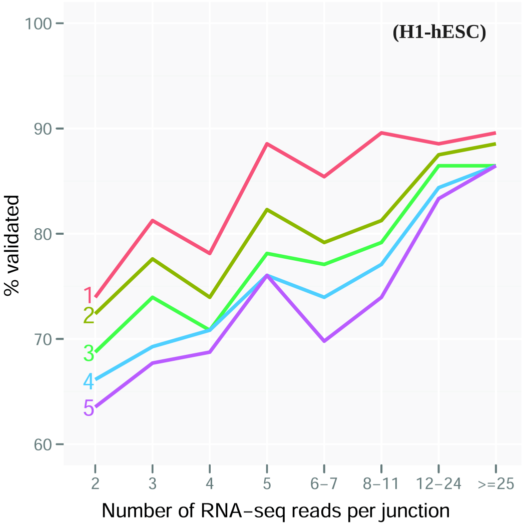
Independent validation of multi-exonic transcript models

- Using overlapping targeted Roche FLX 454 paired-end reads and mass spectrometry.

⋆ The selected 3,000 GT/AG splice junctions identified from Illumina RNA-seq:
  a. Not annotated in GENCODE;
  b. Map to intergenic & antisense regions of H1-hESC, HepG2, Hela-S3 (poly-A^{+}, whole cell).

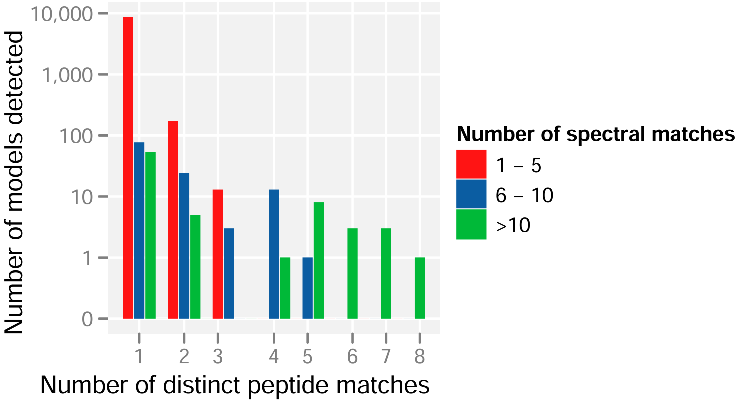


Independent validation of multi-exonic transcript models (contd.)



Most novel transcripts seem to lack protein-coding capacity

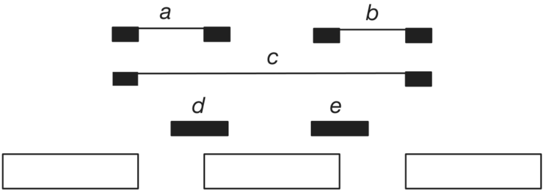
- Distribution of spectral and peptide identifications in novel exons.



The transcriptome of nuclear subcompartments (K562 cell line)

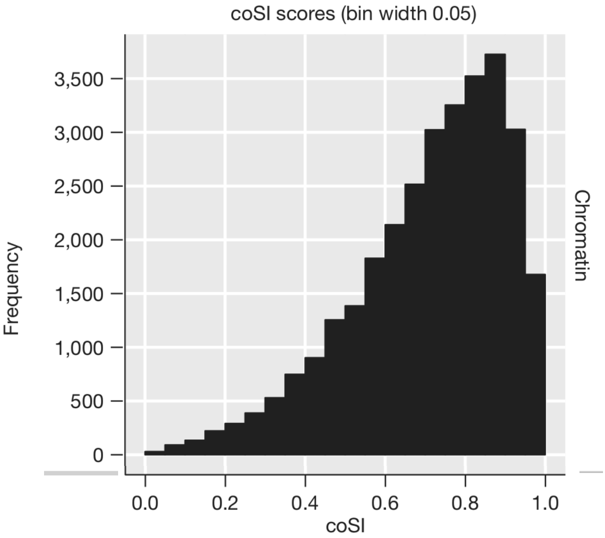
- K562 cell line; total RNA isolated from chromatin, nucleolus and nucleoplasm.
- 51.64% (18,330) of the GENCODE (v7) annotated genes detected for all 15 cell lines (35,494) were identified.
- Only a small fraction of annotated/novel elements was unique to that compartment.

Co-transcriptional splicing (CoSI)

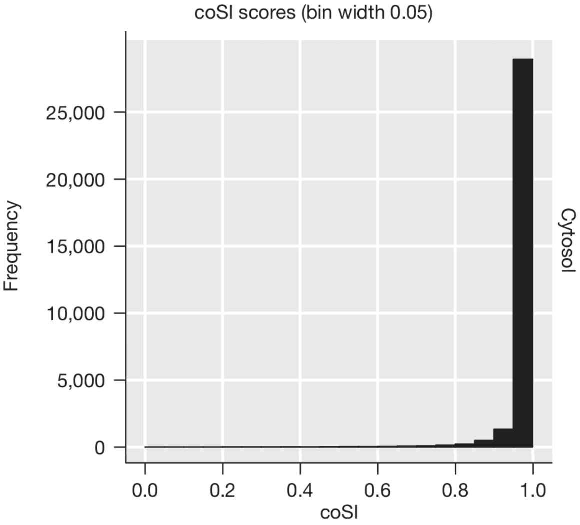


- The complete splicing index (CoSI):
  - The ratio of $\frac{0.5(a+b)+c}{0.5(a+b)+c+0.5(d+e)}$.

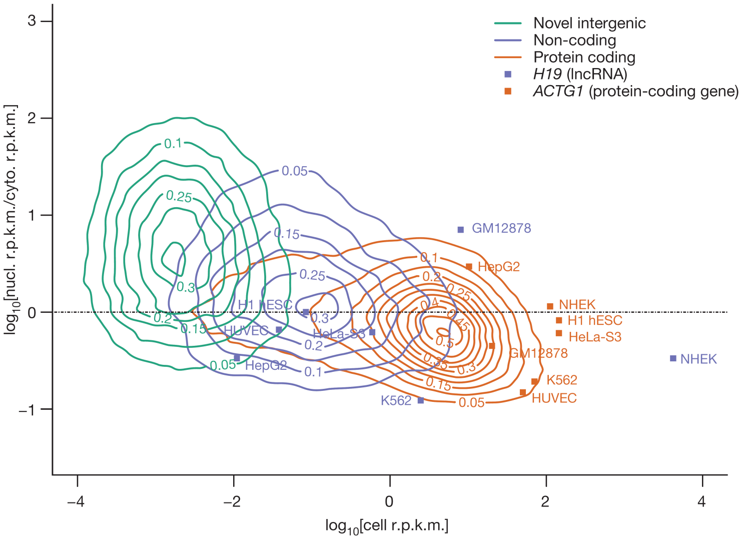
Co-transcriptional splicing (chromatin)



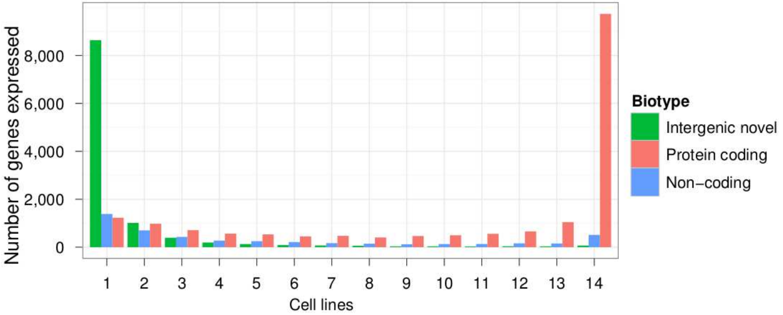
Co-transcriptional splicing (cytosol)



Abundance of gene types in cellular compartments



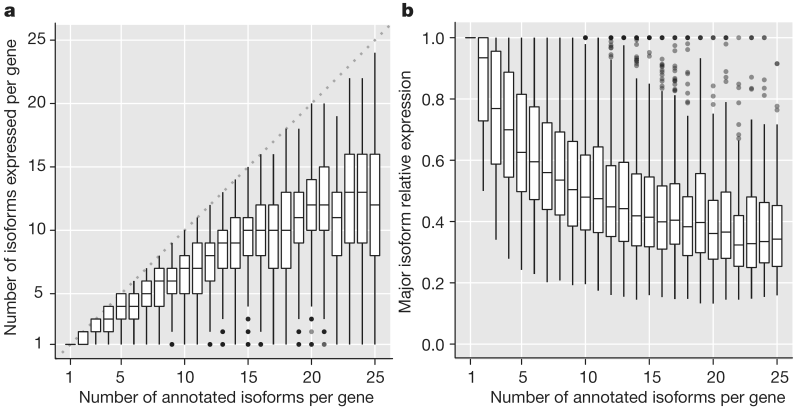
lncRNAs contributes more to cell-line specificity than protein-coding genes



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.

Isoform expression within a gene



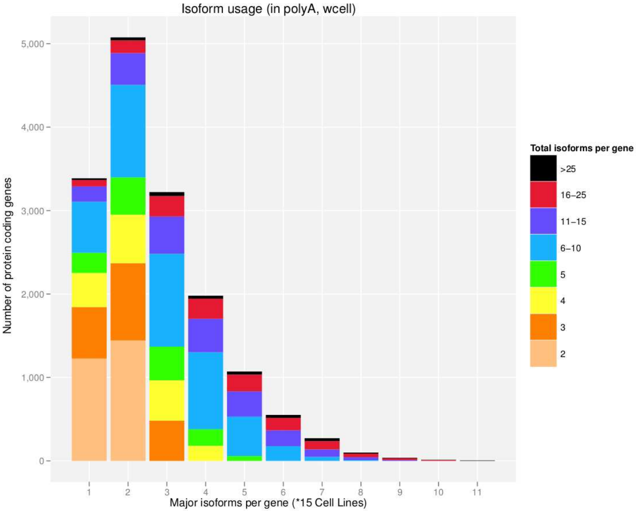
Splice site usage

- For each protein coding gene in each cell line, the authors computed:
  - # detected splice junctions (# detected isoforms)
  - relative expression of the most frequently used splice junction (major isoform)
  - the Shannon's diversity index on the relative usage of gene's annotated splice junctions.

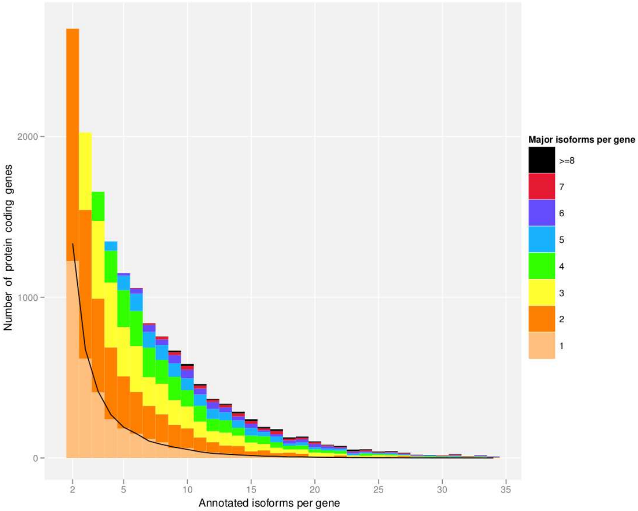
- Let g be a gene with n annotated isoforms with relative frequencies $p_1, p_2, \ldots, p_n$ in a given cell line. The entropy of g is

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

Splice site usage (contd.)



Splice site usage (contd.)



Splice site usage (contd.)

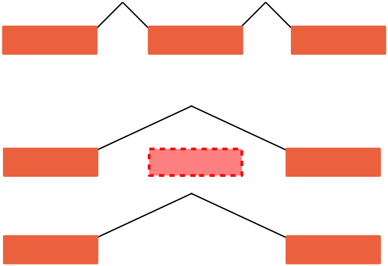
- The average entropy is slightly *higher* when compute on splice site usage than isoform usage

Splice site usage (contd.)

- The average entropy is slightly *higher* when compute on splice site usage than isoform usage ⇒ why?

Splice site usage (contd.)

- The average entropy is slightly *higher* when compute on splice site usage than isoform usage ⇒ why?

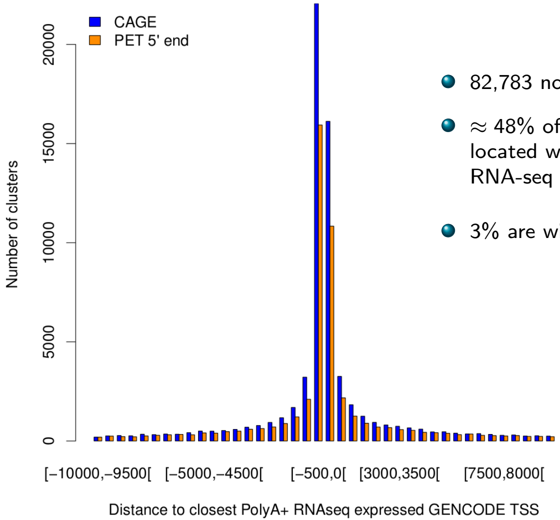


Transcription initiation and termination

Workflow of CAGE processing and elements:

- Raw CAGE reads ⇒ mapped to hg19 genome (using Delve); reads with bad mapping quality were discarded.
  - Delve: a probabilistic mapper using HMM;
    - Iteratively map reads to the genome & estimate position dependent error prob.
- Mapped reads ⇒ clustered using paraclu → hierarchal organization of overlapping clusters; clusters shorter than 200bp were selected ($\because$ length(nucleosome)).
- Using TSS predictor.
  - A non-supervised classifier based on modeling sequences surrounding CAGE regions via HMMs.
  - To capture sequence motifs of length 2–8 present at a certain distance from the middle of each cluster.

Transcription initiation and termination (contd.)



- 82,783 non-redundant TSSs were identified.
- ≈ 48% of the CAGE-identified TSSs are located within 500bp of an annotated RNA-seq detected GENCODE TSS.
- 3% are within 500bp of a novel TSS.

Transcription initiation and termination (contd.)

Correlations to chromatin and features of initiation of transcription:

- 44.7% (199,146) of the RNA-seq-supported TSSs also displayed evidence of CAGE.
- $\approx$ half of the TSSs are associated with ≥ 1 transcription initiation features (DNase I, H3K27ac & H3K4me3 chromatin modifications).
- Only a small minority of the TSSs identified by either CAGE or RNA-seq/GENCODE displayed **all** of the characteristics of transcription start.

As to transcription termination:

- 128,824 sites mapping within annotated GENCODE transcripts were identified.
  - Trim unmapped RNA-seq reads with long terminal poly-As first.
  - ⋆ $\approx$ 20% mapped proximal to annotated poly-A sites (PAS).
  - ⋆ $\approx$ 80% correspond to novel PAS.
- A cell-type preference for proximal PAS in the cytosol compared to the nucleus.

Transcription initiation and termination (contd.)

Correlations to chromatin and features of initiation of transcription:

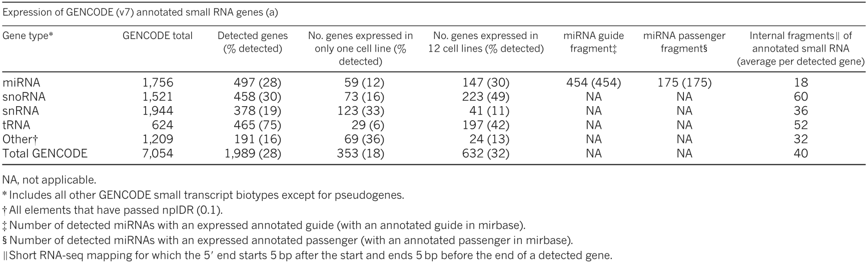
- 44.7% (199,146) of the RNA-seq-supported TSSs also displayed evidence of CAGE.
- $\approx$ half of the TSSs are associated with ≥ 1 transcription initiation features (DNase I, H3K27ac & H3K4me3 chromatin modifications).
- Only a small minority of the TSSs identified by either CAGE or RNA-seq/GENCODE displayed **all** of the characteristics of transcription start.

As to transcription termination:

- 128,824 sites mapping within annotated GENCODE transcripts were identified.
  - Trim unmapped RNA-seq reads with long terminal poly-As first.
  - ⋆ $\approx$ 20% mapped proximal to annotated poly-A sites (PAS).
  - ⋆ $\approx$ 80% correspond to novel PAS.
- A cell-type preference for proximal PAS in the cytosol compared to the nucleus.

Short RNA expression landscape

Annotated small RNAs



Expression of GENCODE (v7) annotated small RNA genes (a)

Gene type*	GENCODE total	Detected genes (% detected)	No. genes expressed in only one cell line (% detected)	No. genes expressed in 12 cell lines (% detected)	miRNA guide fragment‡	miRNA passenger fragment§	Internal fragments‖ of annotated small RNA (average per detected gene)
miRNA	1,756	497 (28)	59 (12)	147 (30)	454 (454)	175 (175)	18
snoRNA	1,521	458 (30)	73 (16)	223 (49)	NA	NA	60
snRNA	1,944	378 (19)	123 (33)	41 (11)	NA	NA	36
tRNA	624	465 (75)	29 (6)	197 (42)	NA	NA	52
Other†	1,209	191 (16)	69 (36)	24 (13)	NA	NA	32
Total GENCODE	7,054	1,989 (28)	353 (18)	632 (32)	NA	NA	40

NA, not applicable.

* Includes all other GENCODE small transcript biotypes except for pseudogenes.

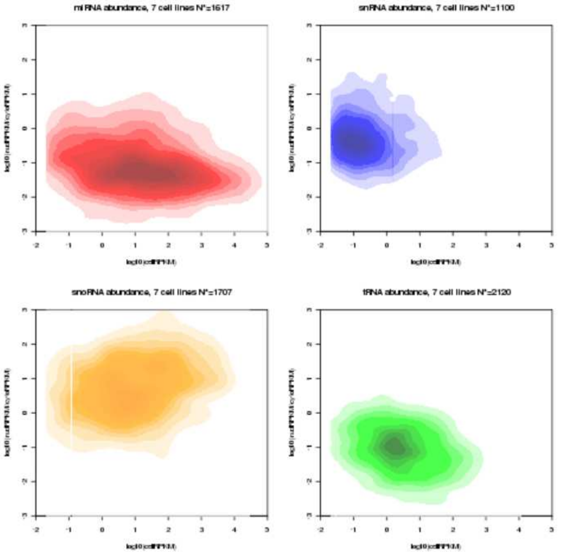
† All elements that have passed npIDR (0.1).

‡ Number of detected miRNAs with an expressed annotated guide (with an annotated guide in mirbase).

§ Number of detected miRNAs with an expressed annotated passenger (with an annotated passenger in mirbase).

‖ Short RNA-seq mapping for which the 5′ end starts 5 bp after the start and ends 5 bp before the end of a detected gene.

Annotated small RNAs (all cell lines)



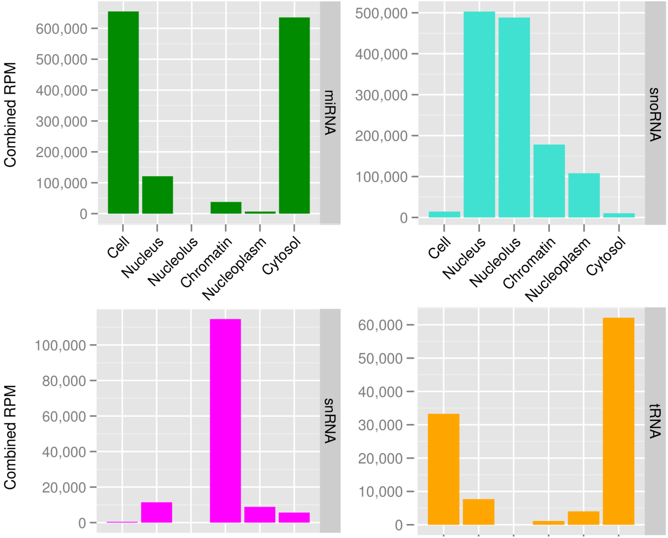
miRNA abundance, 7 cell lines N*=1617

snRNA abundance, 7 cell lines N*=1100

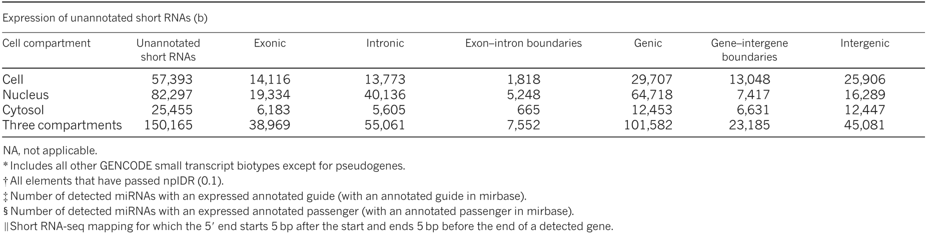
snoRNA abundance, 7 cell lines N*=1707

tRNA abundance, 7 cell lines N*=2120

Annotated small RNAs (K562)



Unannotated small RNAs



Expression of unannotated short RNAs (b)

Cell compartment	Unannotated short RNAs	Exonic	Intronic	Exon–intron boundaries	Genic	Gene–intergene boundaries	Intergenic
Cell	57,393	14,116	13,773	1,818	29,707	13,048	25,906
Nucleus	82,297	19,334	40,136	5,248	64,718	7,417	16,289
Cytosol	25,455	6,183	5,605	665	12,453	6,631	12,447
Three compartments	150,165	38,969	55,061	7,552	101,582	23,185	45,081

NA, not applicable.

* Includes all other GENCODE small transcript biotypes except for pseudogenes.

† All elements that have passed npIDR (0.1).

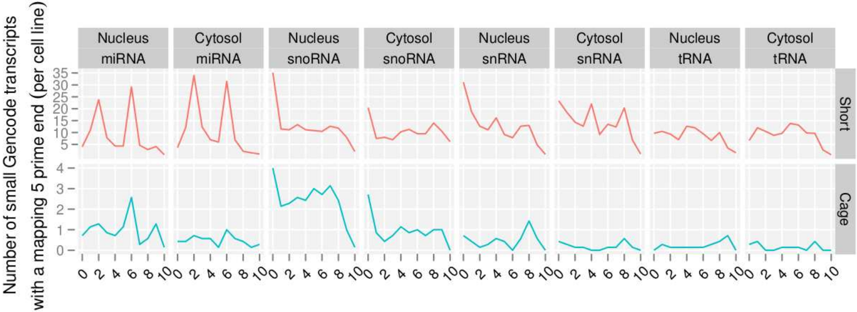
‡ Number of detected miRNAs with an expressed annotated guide (with an annotated guide in mirbase).

§ Number of detected miRNAs with an expressed annotated passenger (with an annotated passenger in mirbase).

‖ Short RNA-seq mapping for which the 5′ end starts 5 bp after the start and ends 5 bp before the end of a detected gene.

Unannotated small RNAs (contd.)

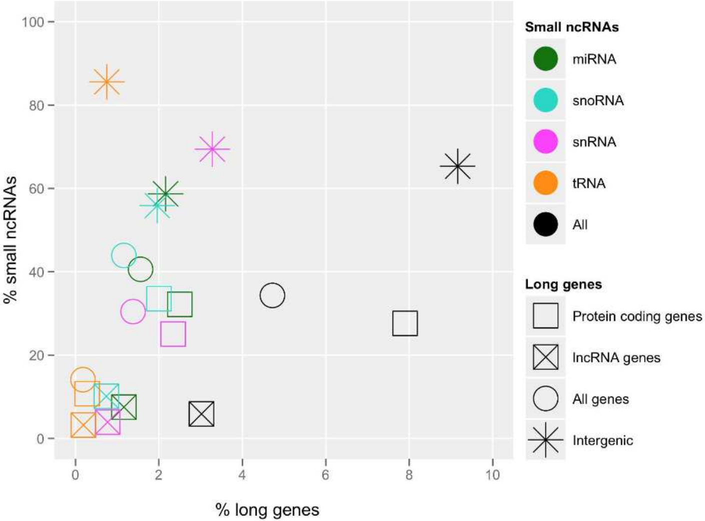
- Two types detected:
  - Subfragments of annotated small RNAs.



Unannotated small RNAs (contd.)

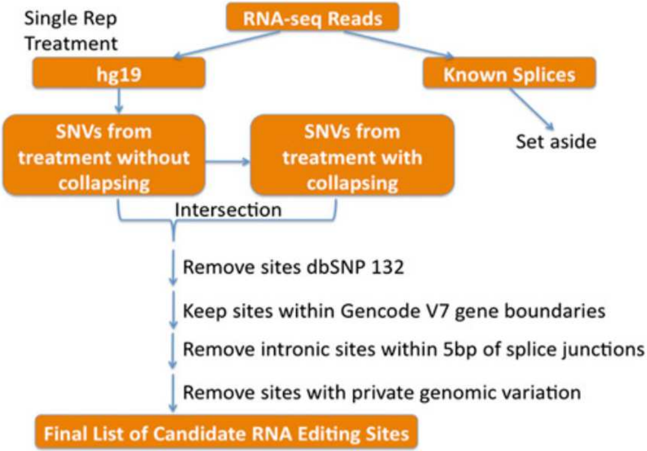
- Two types detected:
  - Subfragments of annotated small RNAs.
  - Novel short RNAs mapping outside of annotated ones.
    - Almost 90% of these are only observed in one cell line (low copy numbers).
    - Nearly 40% of these are associated with promoter & terminator regions of annotated genes.

Genealogy of short RNAs



RNA editing & allele-specific expression

The pipeline [Park *et al.* *Genome Research* 2012]



Single Rep Treatment

RNA-seq Reads

hg19

Known Splices

Set aside

SNVs from treatment without collapsing

SNVs from treatment with collapsing

Intersection

Remove sites dbSNP 132

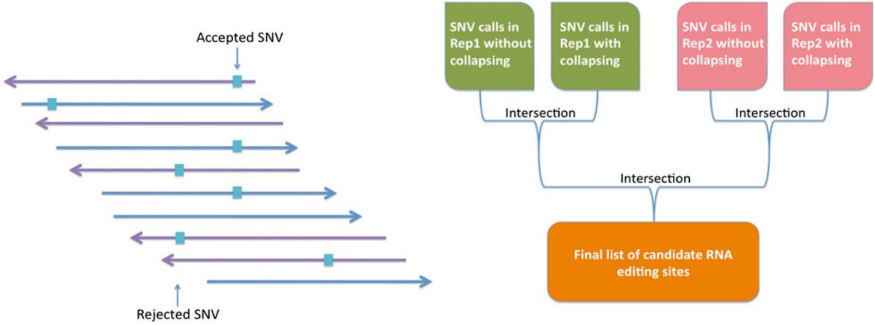
Keep sites within Gencode V7 gene boundaries

Remove intronic sites within 5bp of splice junctions

Remove sites with private genomic variation

Final List of Candidate RNA Editing Sites

The pipeline [Park *et al.* *Genome Research* 2012]



SNV calls in Rep1 without collapsing

SNV calls in Rep1 with collapsing

SNV calls in Rep2 without collapsing

SNV calls in Rep2 with collapsing

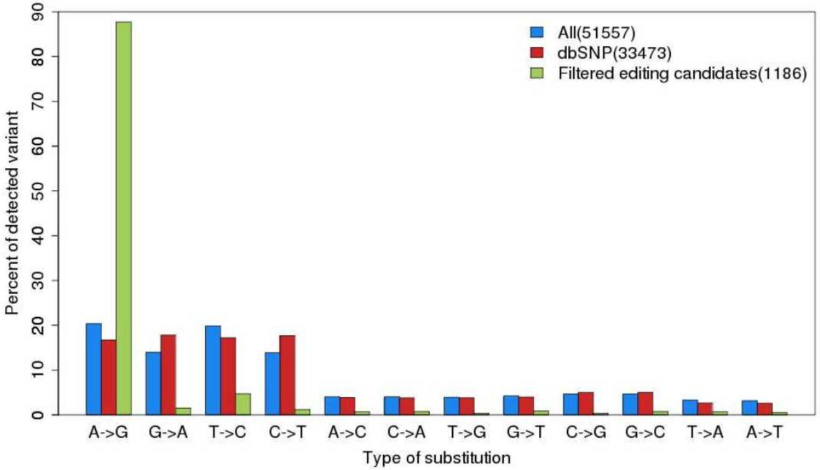
Intersection

Intersection

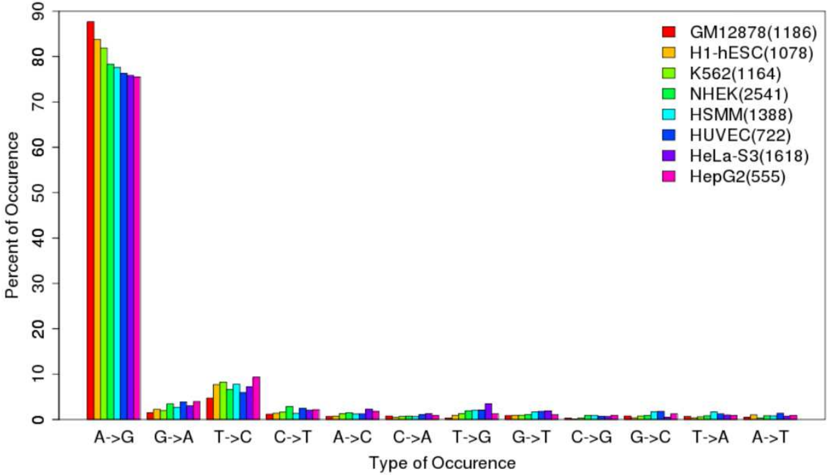
Intersection

Final list of candidate RNA editing sites

RNA-detected SNVs in GM12878



RNA-detected SNVs in 8 cell lines



Allele-specific expression (GM12878 RNA-seq datasets)

The AlleleSeq pipeline [Rozowsky *et al.* *Mol. Syst. Biol*. 2011]

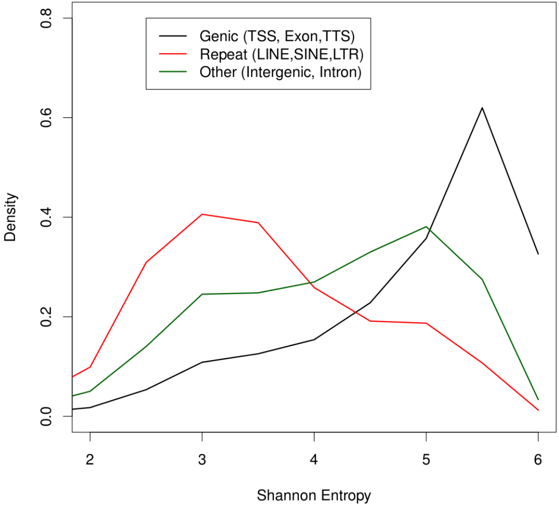
- RNA-seq reads were independently mapped using Bowtie against both **maternal** and **paternal** haplotype sequences.
  - Constructed for the NA12878 genome using phased variant calls from the pilot phase of the 1000 Genome Project Consortium.
- Heterozygous SNPs in sufficiently highly transcribed regions can be used to distinguish those regions that exhibit allele-specific expression from those that are not (by counting reads mapping to each allele).
- ≈ 18% of both GENCODE annotated protein-coding & non-coding genes exhibit allele-specific expression.
  - Similar proportion of genes with allele-specific expression in whole-cell, cytoplasm, & nucleus.

Repeat region transcription

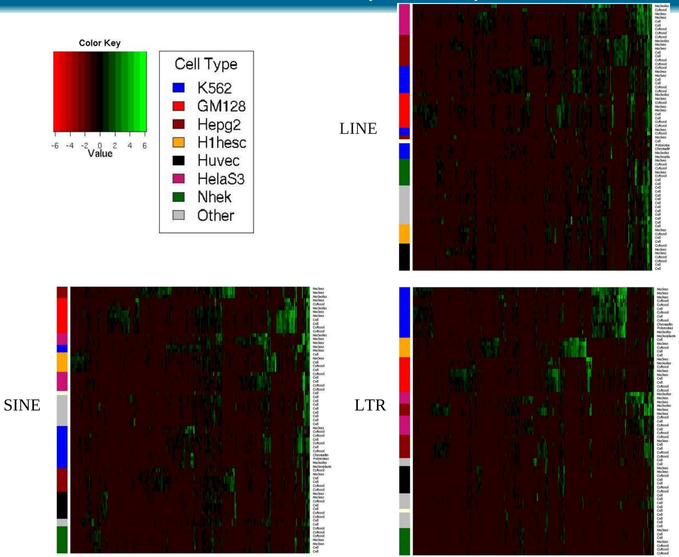
Repeat region transcription

- 18% (14,828) of CAGE-defined TSS regions overlap repetitive elements.
- # Intergenic CAGE clusters overlapping repeat elements:
  - 322 for long interspersed elements (LINE);
  - 315 for short interspersed elements (SINE);
  - 507 for long terminal repeat (LTR);
  - 1,262 for the other repeat elements.

Repeat region transcription (contd.)



Repeat region transcription (contd.)

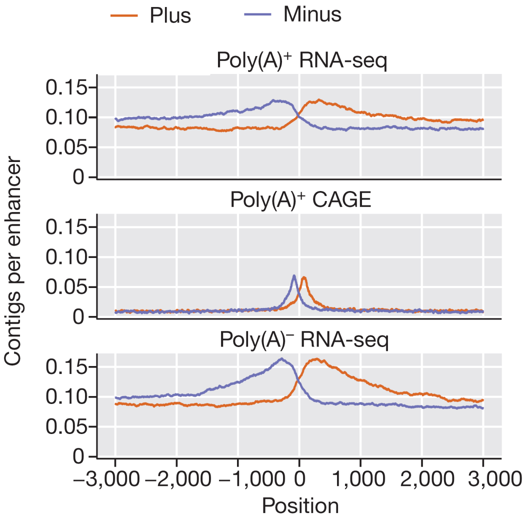


Characterization of enhancer RNA

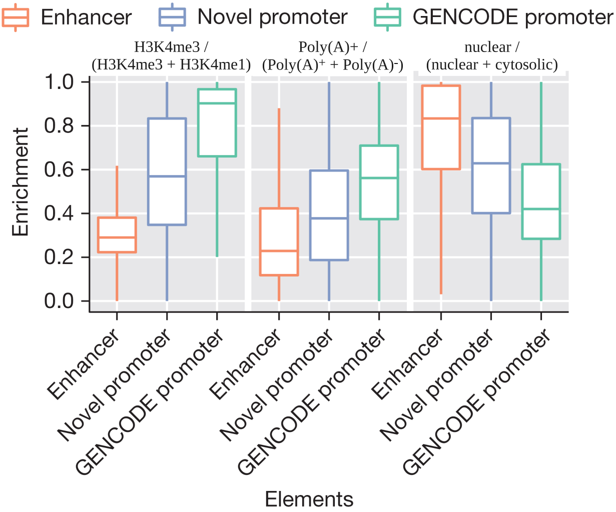
Transcription at enhancers

- RNA polymerase II binds some distal enhancer regions & produce enhancer-associated transcripts (eRNAs).
- Material used:
  - enhancer loci predicted from ENCODE ChIP-seq data.

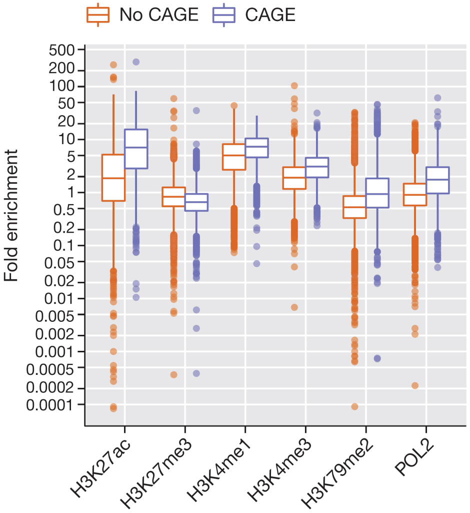
Transcription at enhancers



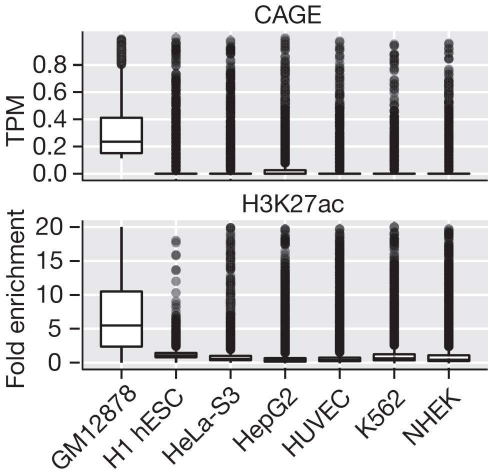
Enhancer transcripts differ from promoter transcripts



Chromatin state at transcribed enhancers



Enhancer activity & transcription is cell-type specific



Thanks for your listening!