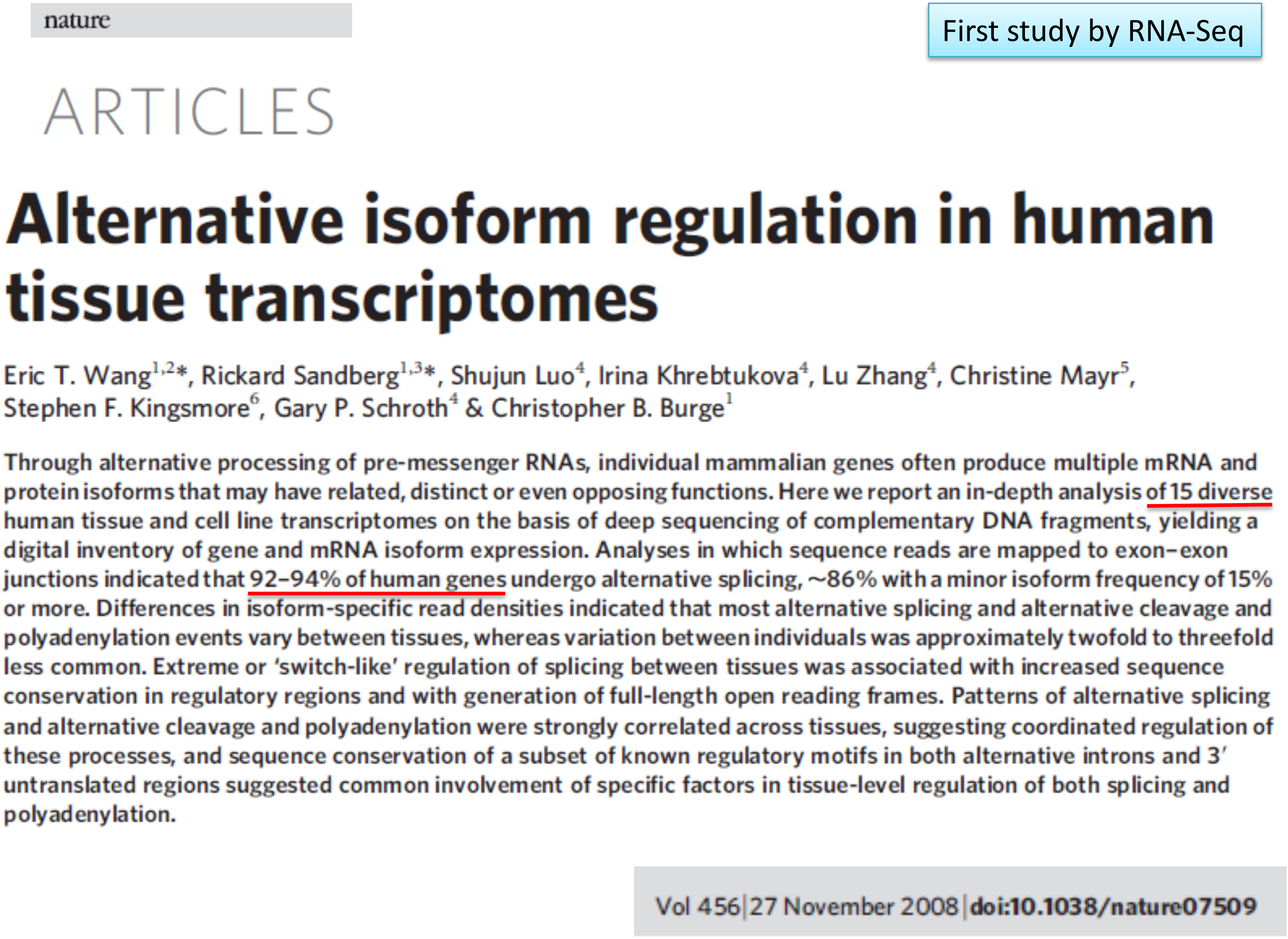
L3.3 Transcriptomes (More on RNA-seq)

AGENDA

1. Representative RNAseq papers (ENCODE, modENCODE)

2. Discuss Research Paper for this unit (ENCODE)

First study by RNA-Seq



nature

ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

Vol 456|27 November 2008|**doi:10.1038/nature07509**

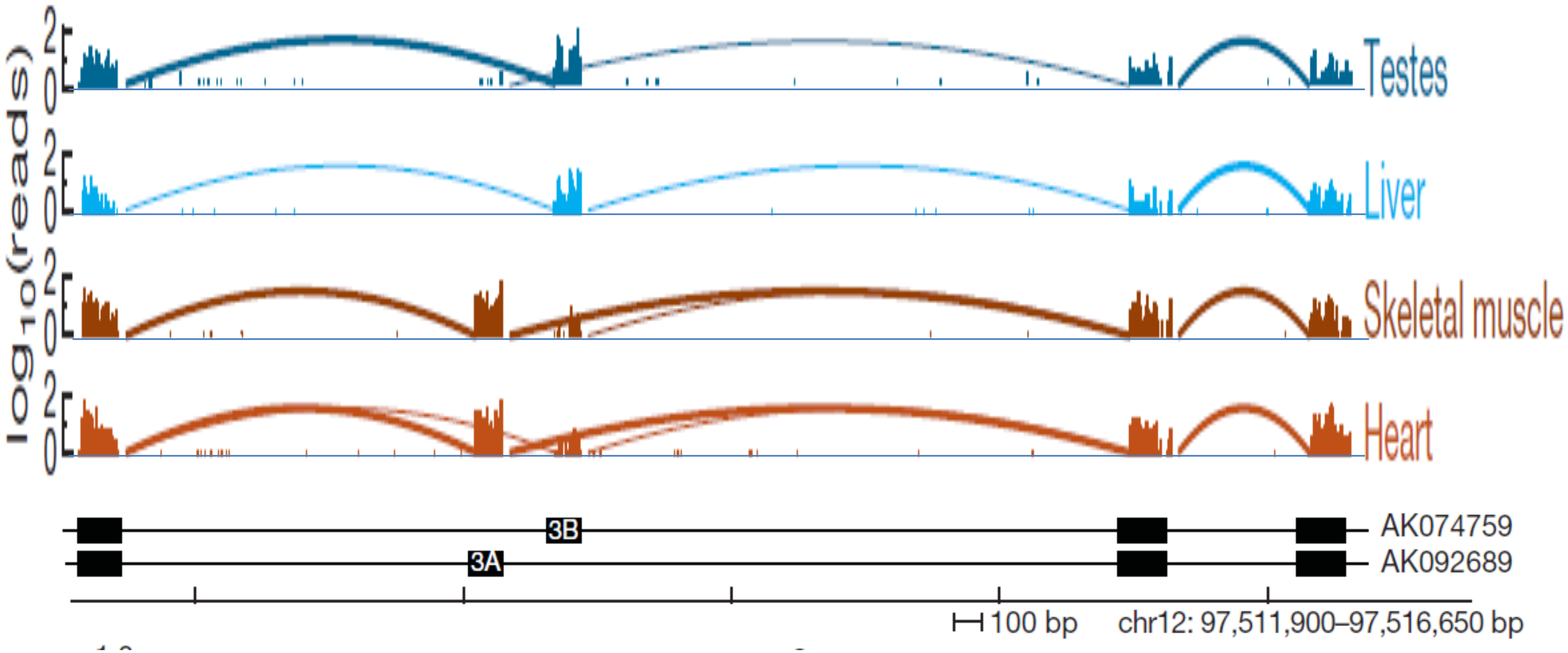
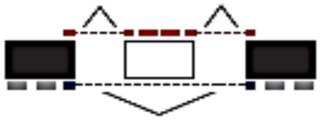
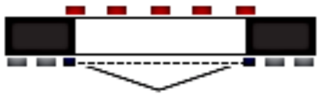
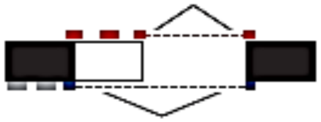
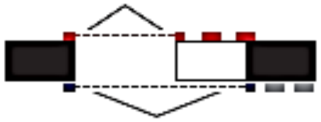
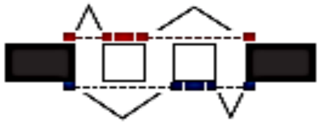
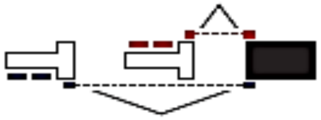
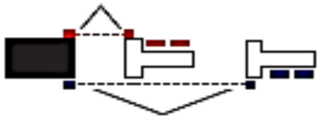
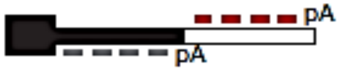


Figure 1 | Frequency and relative abundance of alternative splicing isoforms in human genes.

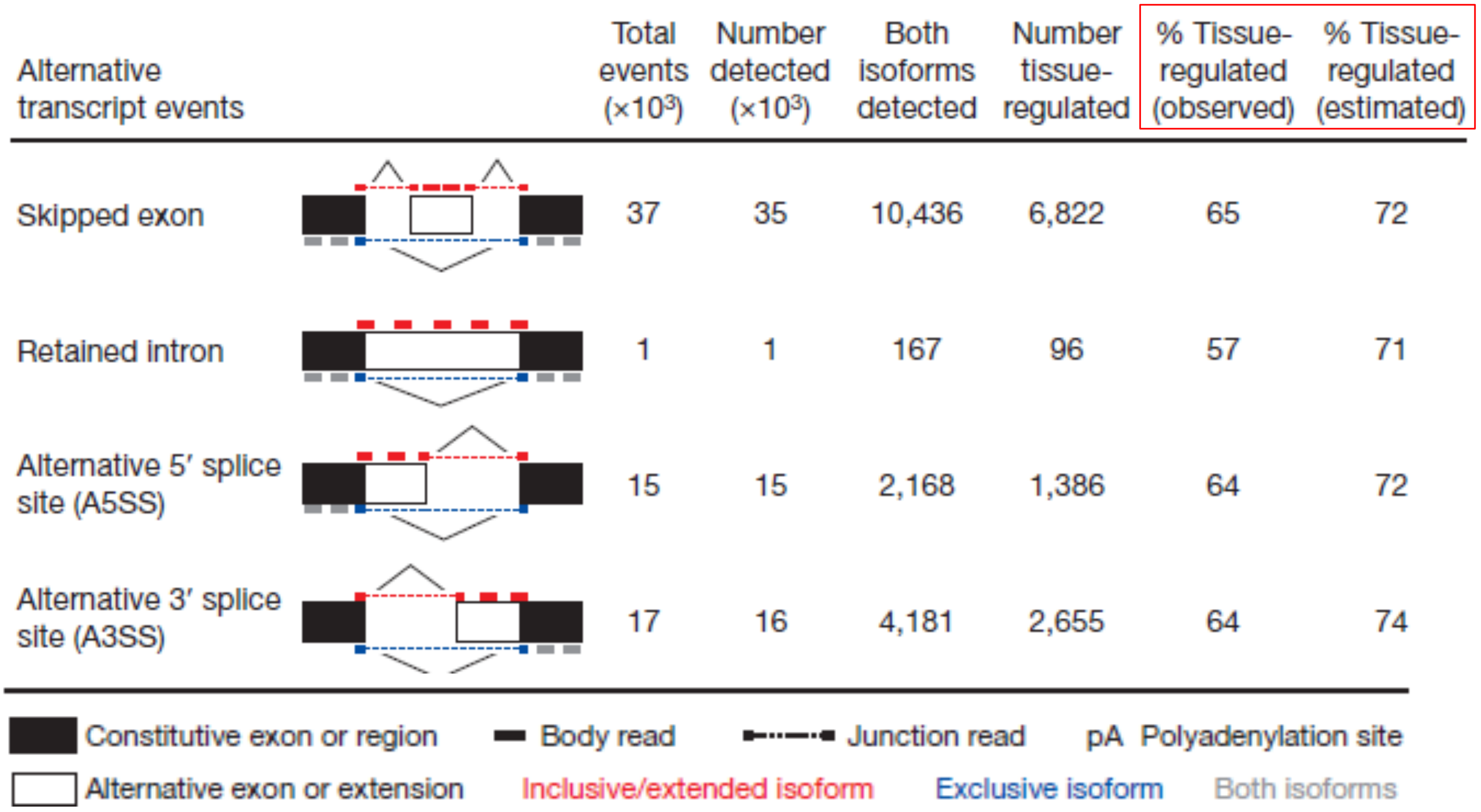
a, mRNA-Seq reads mapping to a portion of the SLC25A3 gene locus. The number of mapped reads starting at each nucleotide position is displayed (log10) for the tissues listed at the right. Arcs represent junctions detected by splice junction reads.

Bottom: exon/intron structures of representative transcripts containing mutually exclusive exons 3A and 3B (GenBank accession numbers shown at the right).

Alternative transcript events		Total events (×10^3)	Number detected (×10^3)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

■ Constitutive exon or region ▬ Body read ■--------■ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

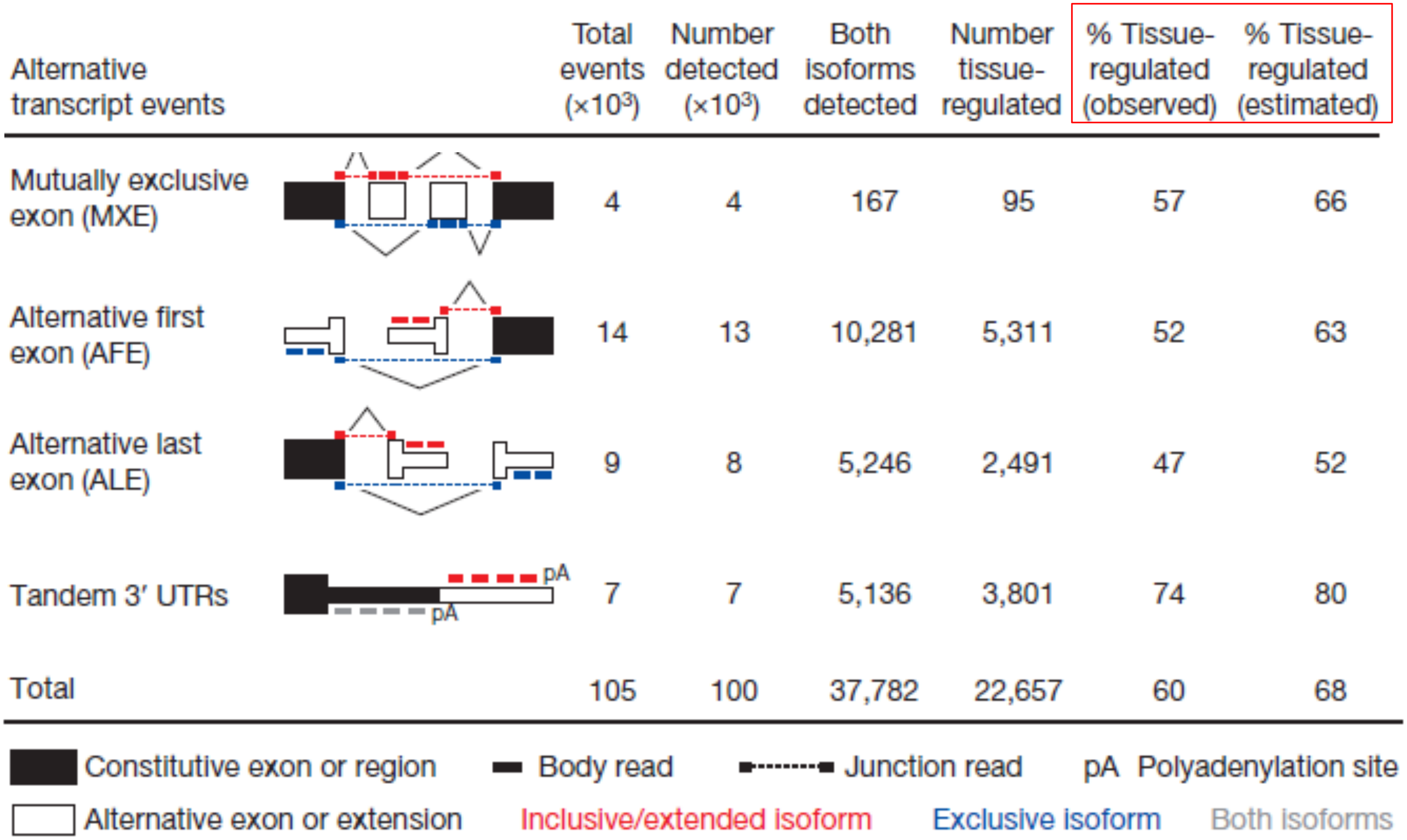


Alternative transcript events	Total events (×10³)	Number detected (×10³)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74

■ Constitutive exon or region ▬ Body read ▪┈┈▪ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1–4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥ 1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue regulated (Fisher’s exact test) at an FDR of 5% (assuming negligible technical variation).



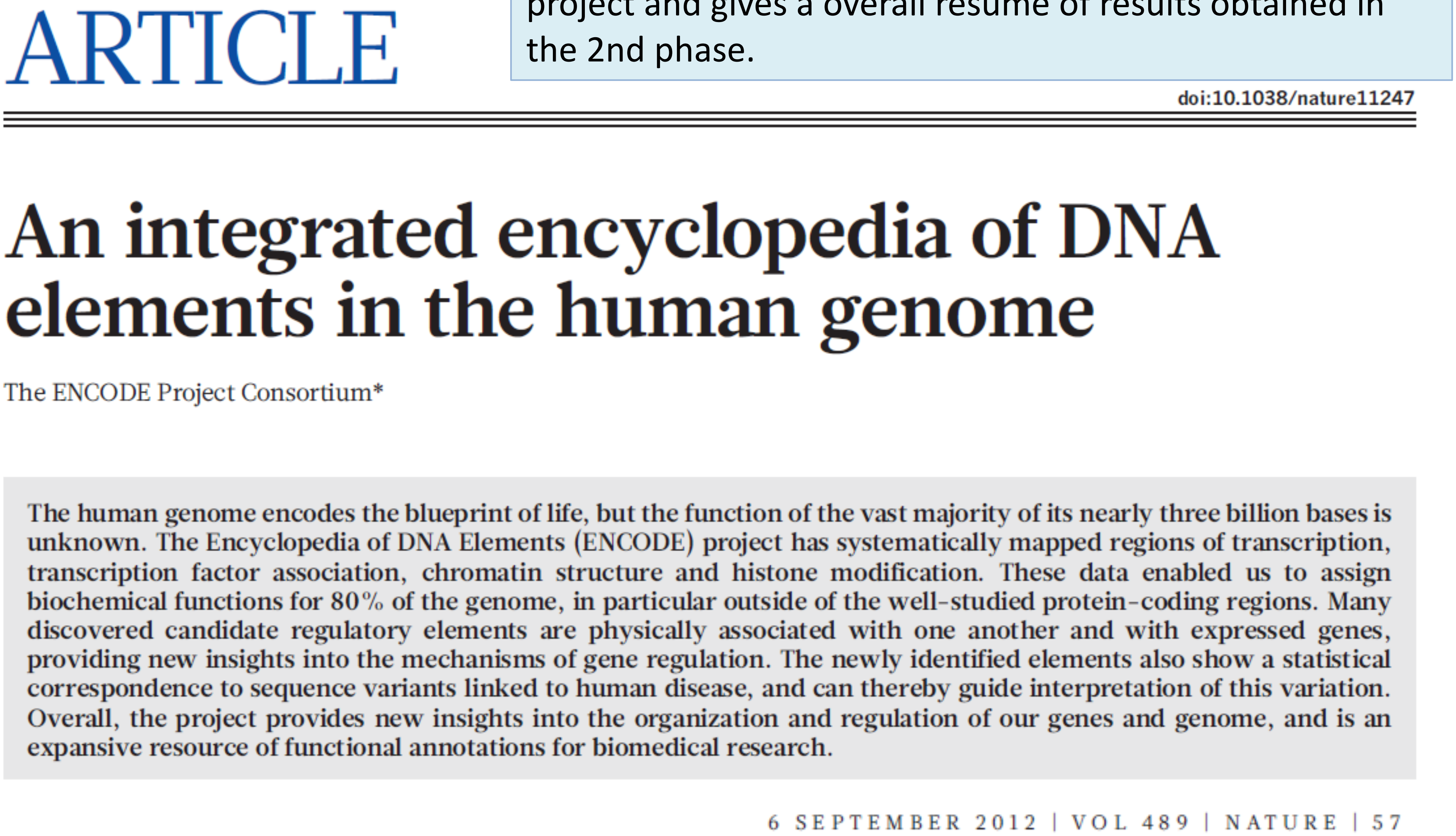
Alternative transcript events	Total events (×10³)	Number detected (×10³)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

■ Constitutive exon or region ▬ Body read ■--------■ Junction read pA Polyadenylation site

□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, ‘common reads’ (grey bars) were used in lieu of (for tandem 39UTR events) or in addition to ‘exclusion’ reads for detection of changes in isoform levels between tissues.
Note that Aa use the following definition for “tissue-specific”: at least 10% variation in isoforms.

This is the leading article that describes all the ENCODE project and gives a overall resumé of results obtained in the 2nd phase.



ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

6 SEPTEMBER 2012 | VOL 489 | NATURE | 57

ENCODE official website: https://www.encodeproject.org/
ENCODE at the NHGRI: http://www.genome.gov/encode/
Nature ENCODE: http://www.nature.com/encode/#/threads

ENCODE

Encyclopedia of DNA Elements

nature.com/encode

Overall mission of the Encyclopedia of DNA Elements (ENCODE) project: identifying and characterizing the functional elements present in the human genome sequence. A key part of it is to catalogue the entire repertoire of RNAs produced by human cells.

Initial pilot phase (2003): approximately 1% of the human genome was examined. it was observed that both gene-rich and gene-poor regions were pervasively transcribed.

Second phase of the ENCODE project (2007-2012): the scope was broadened to interrogate the complete human genome, to provide a genome-wide catalogue of human transcripts and to identify the subcellular localization for the RNAs produced.

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

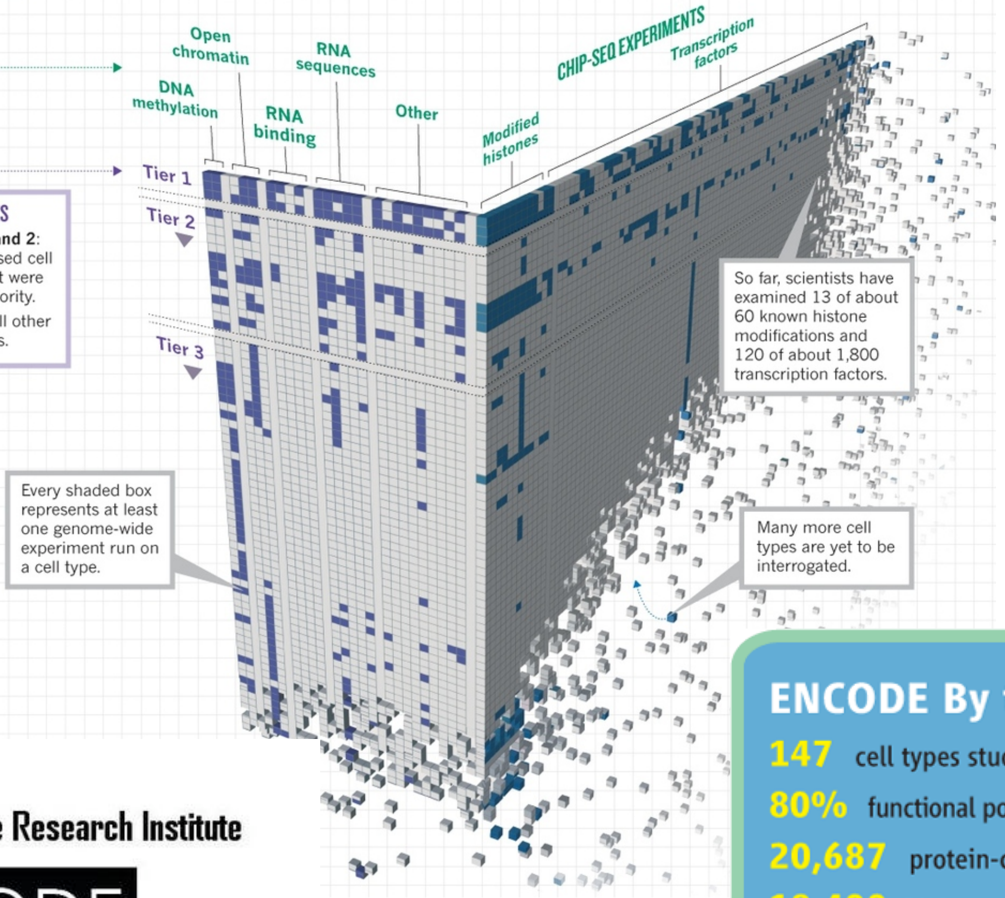
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.



Every shaded box represents at least one genome-wide experiment run on a cell type.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Many more cell types are yet to be interrogated.

National Human Genome Research Institute

ENCODE

ENCODE By the Numbers

147 cell types studied

80% functional portion of human genome

20,687 protein-coding genes

18,400 RNA genes

1640 data sets

30 papers published this week

442 researchers

$288 million funding for pilot, technology, model organism, and current project

Navigate to :

ENCODE website

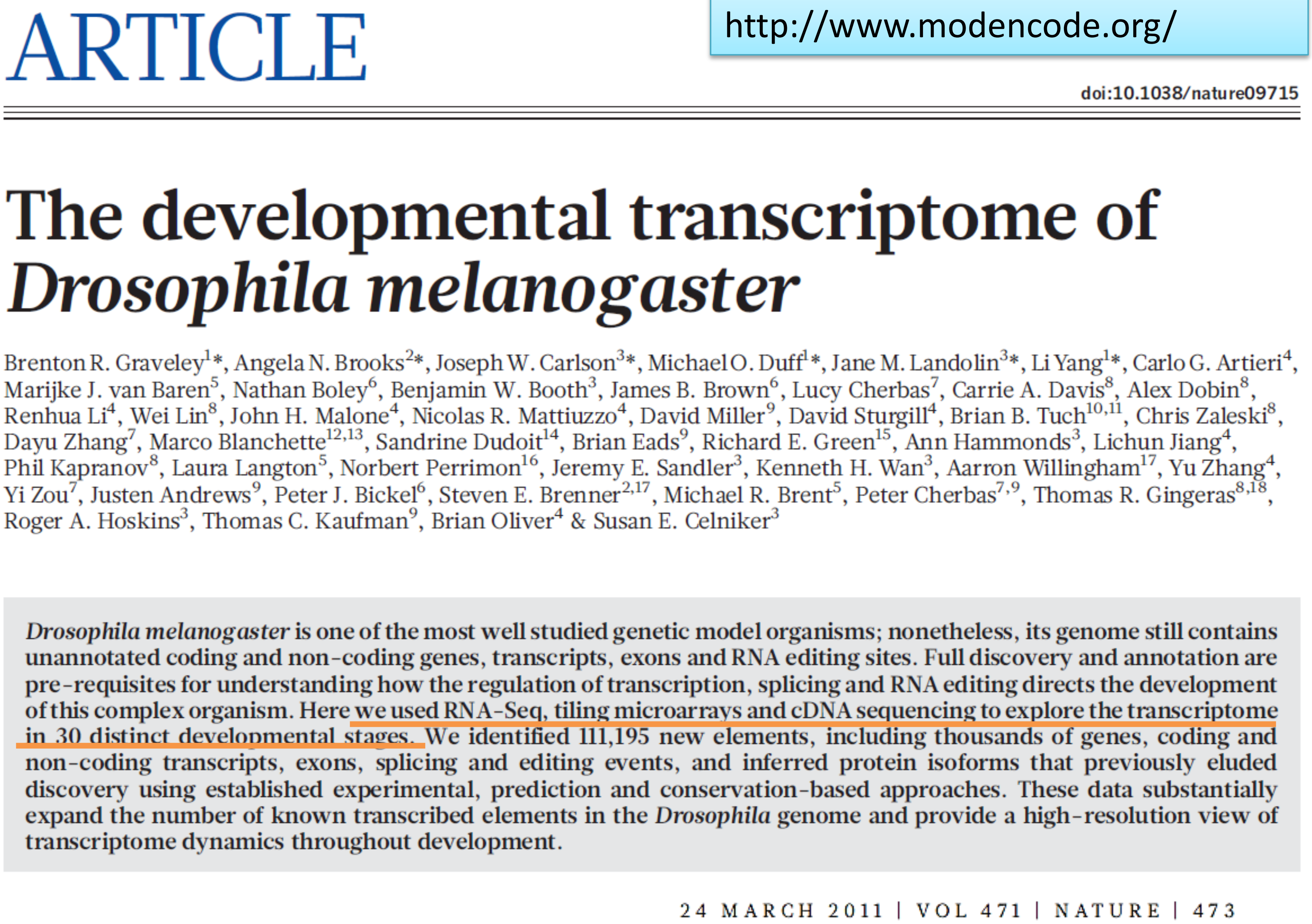
Description of Encyclopedia https://www.encodeproject.org/data/annotations/
Matrix https://www.encodeproject.org/matrix/?type=Experiment
ENCODE data on the USCS browser http://genome.ucsc.edu/ENCODE/

Nature ENCODE website

Nature Explorer http://www.nature.com/encode/
How to navigate the ENCODE Papers through Threads
I.e. transcriptome (03)
http://www.nature.com/encode/threads/characterization-of-intergenic-regions-and-gene-definition

Research paper: Djebali et al., Nature 2012
Will discuss later in details

modENCODE project (Worm and Fly)
http://www.modencode.org/



ARTICLE

doi:10.1038/nature09715

The developmental transcriptome of *Drosophila melanogaster*

Brenton R. Graveley[1]*, Angela N. Brooks[2]*, Joseph W. Carlson[3]*, Michael O. Duff[1]*, Jane M. Landolin[3]*, Li Yang[1]*, Carlo G. Artieri[4], Marijke J. van Baren[5], Nathan Boley[6], Benjamin W. Booth[3], James B. Brown[6], Lucy Cherbas[7], Carrie A. Davis[8], Alex Dobin[8], Renhua Li[4], Wei Lin[8], John H. Malone[4], Nicolas R. Mattiuzzo[4], David Miller[9], David Sturgill[4], Brian B. Tuch[10,11], Chris Zaleski[8], Dayu Zhang[7], Marco Blanchette[12,13], Sandrine Dudoit[14], Brian Eads[9], Richard E. Green[15], Ann Hammonds[3], Lichun Jiang[4], Phil Kapranov[8], Laura Langton[5], Norbert Perrimon[16], Jeremy E. Sandler[3], Kenneth H. Wan[3], Aarron Willingham[17], Yu Zhang[4], Yi Zou[7], Justen Andrews[9], Peter J. Bickel[6], Steven E. Brenner[2,17], Michael R. Brent[5], Peter Cherbas[7,9], Thomas R. Gingeras[8,18], Roger A. Hoskins[3], Thomas C. Kaufman[9], Brian Oliver[4] & Susan E. Celniker[3]

Drosophila melanogaster is one of the most well studied genetic model organisms; nonetheless, its genome still contains unannotated coding and non-coding genes, transcripts, exons and RNA editing sites. Full discovery and annotation are pre-requisites for understanding how the regulation of transcription, splicing and RNA editing directs the development of this complex organism. Here we used RNA-Seq, tiling microarrays and cDNA sequencing to explore the transcriptome in 30 distinct developmental stages. We identified 111,195 new elements, including thousands of genes, coding and non-coding transcripts, exons, splicing and editing events, and inferred protein isoforms that previously eluded discovery using established experimental, prediction and conservation-based approaches. These data substantially expand the number of known transcribed elements in the *Drosophila* genome and provide a high-resolution view of transcriptome dynamics throughout development.

SAMPLES:

RNA was isolated from 30 whole-animal samples representing 27 distinct stages of development. These included 12 embryonic samples collected at 2-h intervals for 24 h, six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion.

COMPLEMENTARY APPROACHES:

- 38-base-pair (bp) resolution genome **tiling microarrays**
- non-strand-specific poly(A) RNA-Seq from all 30 samples generating a combination of **single and paired-end, 75-bp reads** *(Illumina platform)*
- 12 embryonic time points were also interrogated with **strand-specific 50-bp sequence reads** from partially rRNA-depleted total RNA on the *Applied Biosystems SOLiD platform*

Discovery of new transcribed regions

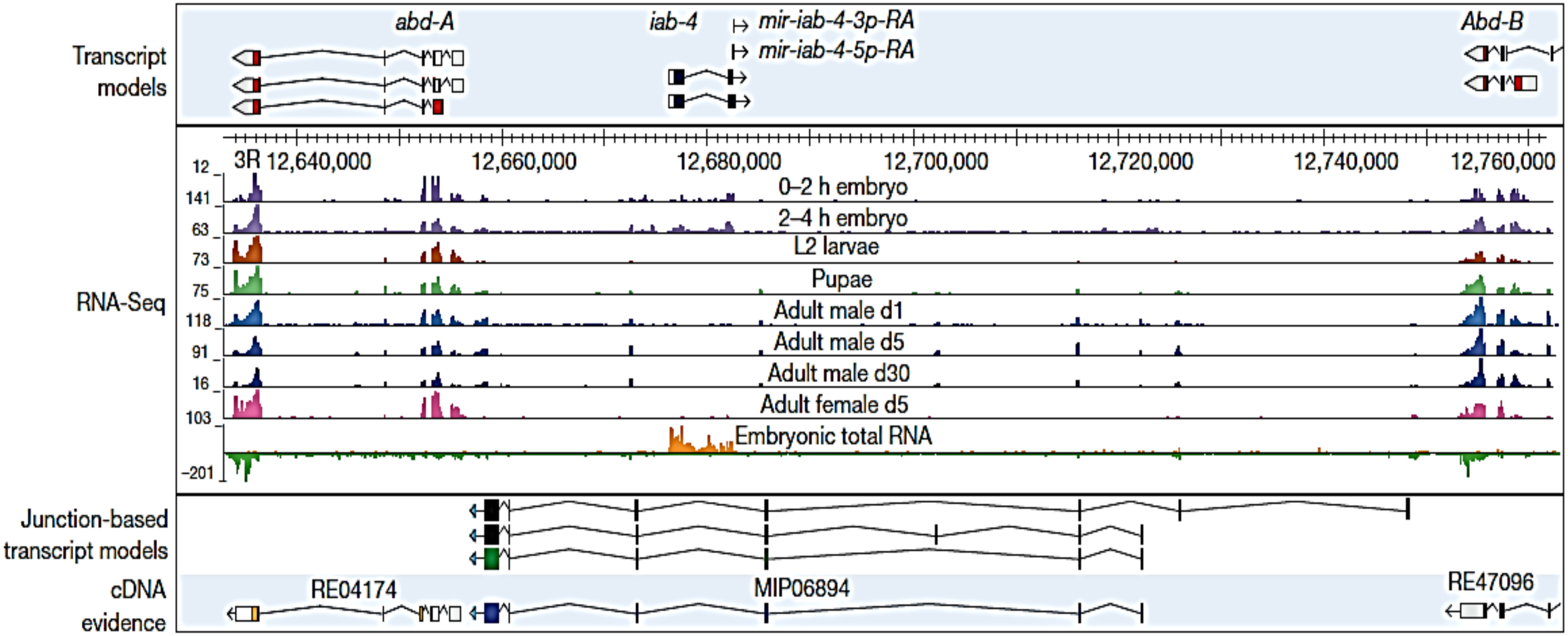
We identified 1,938 new transcribed regions (NTRs) not linked to any annotated gene models.

Discovery of small ncRNAs

We identified 37 unannotated intron-encoded and two unannotated intergenic small ncRNAs (,300 nucleotides)

Discovery of microRNA primary transcripts

We identified 23 putative independently transcribed pri-miRNAs from the total embryonic RNA-Seq and tiling array data that encode 37 annotated miRNAs



Discovery of new RNAs in the Bithorax complex. Genomic organization and experimental evidence for new transcripts located between the HOX genes, abd-A and Abd-B, based on short poly(A) RNA and total RNA-Seq expression profiles. The numbers to the left of each track indicate the maximal number of reads for that sample. Three manually curated junction-based transcript models are shown; the green transcript model was fully validated by a cDNA, MIP06894.

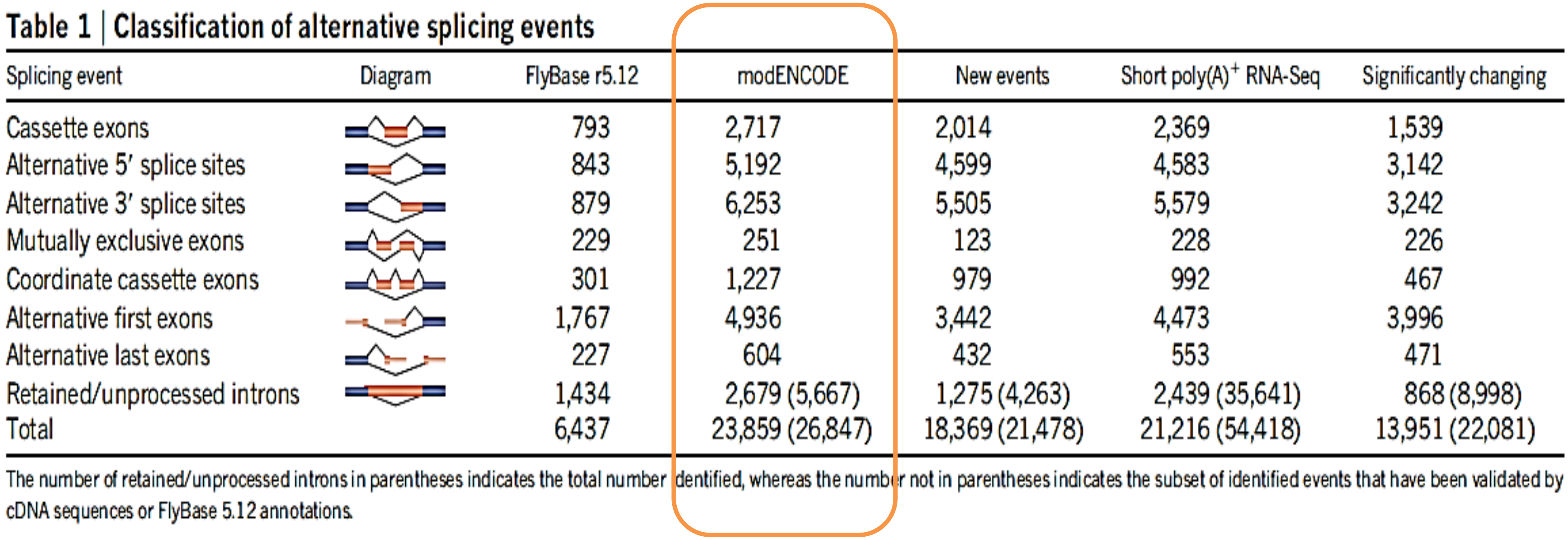


Table 1 | Classification of alternative splicing events

Splicing event	Diagram	FlyBase r5.12	modENCODE	New events	Short poly(A)$^+$ RNA-Seq	Significantly changing
Cassette exons		793	2,717	2,014	2,369	1,539
Alternative 5′ splice sites		843	5,192	4,599	4,583	3,142
Alternative 3′ splice sites		879	6,253	5,505	5,579	3,242
Mutually exclusive exons		229	251	123	228	226
Coordinate cassette exons		301	1,227	979	992	467
Alternative first exons		1,767	4,936	3,442	4,473	3,996
Alternative last exons		227	604	432	553	471
Retained/unprocessed introns		1,434	2,679 (5,667)	1,275 (4,263)	2,439 (35,641)	868 (8,998)
Total		6,437	23,859 (26,847)	18,369 (21,478)	21,216 (54,418)	13,951 (22,081)

The number of retained/unprocessed introns in parentheses indicates the total number identified, whereas the number not in parentheses indicates the subset of identified events that have been validated by cDNA sequences or FlyBase 5.12 annotations.

Discovery and dynamics of alternative splicing.
To characterize constitutive and alternative splicing, we identified 71,316 splice junctions, of which 22,965 were new discoveries.

All data produced by transcriptomics are required to be deposited in Public Databases.

Microarray and RNA-Seq data are stired in specialized databases:

Gene expression databases

GEO (gene expression omnibus) - http://www.ncbi.nlm.nih.gov/gds/

ArrayExpress - https://www.ebi.ac.uk/arrayexpress/

Public databases:

Primary (unannotated sequences)
E.g. GenBank - http://www.ncbi.nlm.nih.gov/genbank/

Secondary (annotated sequences)
NCBI, ENSEMBL, UCSC etc.

Sequence variations
1000 Genomes - HapMap

Specialized databases
TCGA – The Cancer Genome Atlas http://cancergenome.nih.gov/

Knowledge databases
Gene Ontology - http://geneontology.org/

Medical databases
OMIM - http://www.ncbi.nlm.nih.gov/omim

The **NAR Database**

Collects all databases on-line
Every year, one paper describing new databases and utilities

The database page:
https://www.oxfordjournals.org/our_journals/nar/database/c/

The 2016 database issue paper:
http://nar.oxfordjournals.org/content/44/D1/D1.abstract

Important **question**: do all databases report the same information ?

Answer: not really

Although ENSEMBL-GenBank-DDBJ synchronize data (primary), secondary and specialized database differ in annotation and reconstruction algorithms.

See for example:

ENSEMBL (http://www.ensembl.org/index.html)

NCBI (Gene) (http://www.ncbi.nlm.nih.gov/gene/)

UCSC Genome browser (https://genome.ucsc.edu/)

GENCODE browser (http://www.gencodegenes.org/human_biodalliance.html)

Compare one gene of your choice in the different browsers.

Data integration

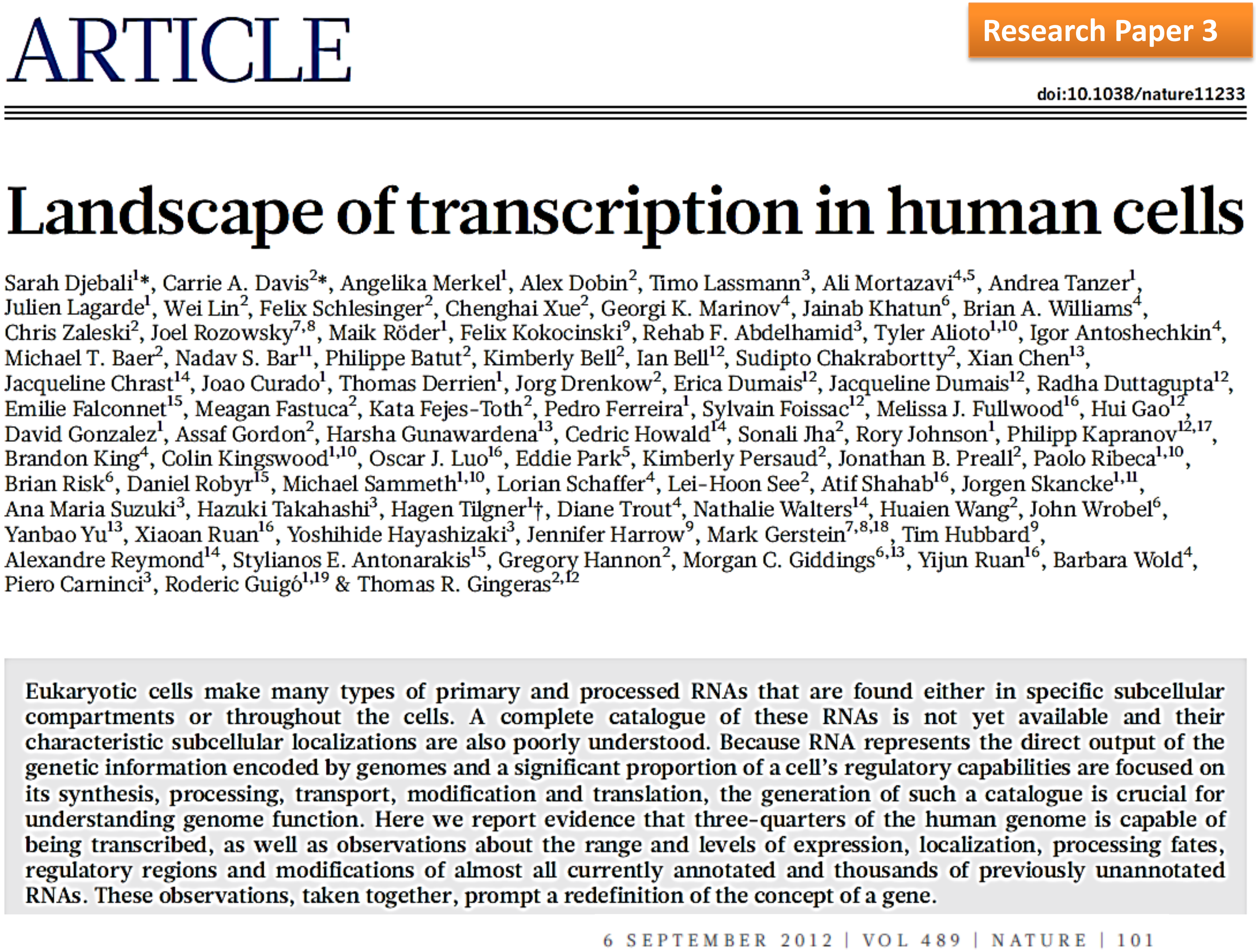
Considering the large amount of genomic/transcriptomics data that is now available in public databases, one important application today is to re-analyze data by integrating them with other, coherent data.

This «integrative « approach is extremely powerful since it may unravel connections, functional categories, and even hidden relationship between genes and proteins.

Needs:

- huge computing power
- enormous storage capability
- new algorithms to manage data
- new statistics
- new representations
- simplification models

Research Paper 3



ARTICLE

doi:10.1038/nature11233

Landscape of transcription in human cells

Sarah Djebali[1]*, Carrie A. Davis[2]*, Angelika Merkel[1], Alex Dobin[2], Timo Lassmann[3], Ali Mortazavi[4,5], Andrea Tanzer[1], Julien Lagarde[1], Wei Lin[2], Felix Schlesinger[2], Chenghai Xue[2], Georgi K. Marinov[4], Jainab Khatun[6], Brian A. Williams[4], Chris Zaleski[2], Joel Rozowsky[7,8], Maik Röder[1], Felix Kokocinski[9], Rehab F. Abdelhamid[3], Tyler Alioto[1,10], Igor Antoshechkin[4], Michael T. Baer[2], Nadav S. Bar[11], Philippe Batut[2], Kimberly Bell[2], Ian Bell[12], Sudipto Chakrabortty[2], Xian Chen[13], Jacqueline Chrast[14], Joao Curado[1], Thomas Derrien[1], Jorg Drenkow[2], Erica Dumais[12], Jacqueline Dumais[12], Radha Duttagupta[12], Emilie Falconnet[15], Meagan Fastuca[2], Kata Fejes-Toth[2], Pedro Ferreira[1], Sylvain Foissac[12], Melissa J. Fullwood[16], Hui Gao[12], David Gonzalez[1], Assaf Gordon[2], Harsha Gunawardena[13], Cedric Howald[14], Sonali Jha[2], Rory Johnson[1], Philipp Kapranov[12,17], Brandon King[4], Colin Kingswood[1,10], Oscar J. Luo[16], Eddie Park[5], Kimberly Persaud[2], Jonathan B. Preall[2], Paolo Ribeca[1,10], Brian Risk[6], Daniel Robyr[15], Michael Sammeth[1,10], Lorian Schaffer[4], Lei-Hoon See[2], Atif Shahab[16], Jorgen Skancke[1,11], Ana Maria Suzuki[3], Hazuki Takahashi[3], Hagen Tilgner[1]†, Diane Trout[4], Nathalie Walters[14], Huaien Wang[2], John Wrobel[6], Yanbao Yu[13], Xiaoan Ruan[16], Yoshihide Hayashizaki[3], Jennifer Harrow[9], Mark Gerstein[7,8,18], Tim Hubbard[9], Alexandre Reymond[14], Stylianos E. Antonarakis[15], Gregory Hannon[2], Morgan C. Giddings[6,13], Yijun Ruan[16], Barbara Wold[4], Piero Carninci[3], Roderic Guigó[1,19] & Thomas R. Gingeras[2,12]

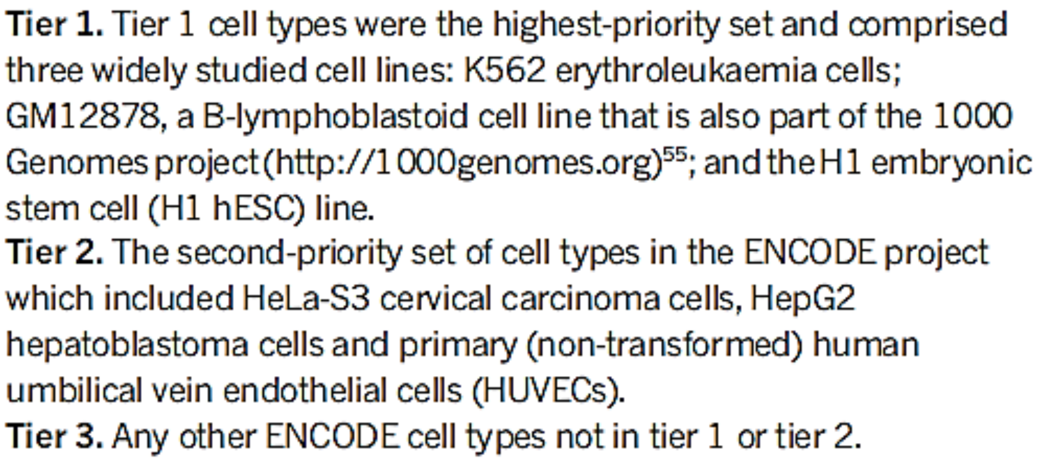
Eukaryotic cells make many types of primary and processed RNAs that are found either in specific subcellular compartments or throughout the cells. A complete catalogue of these RNAs is not yet available and their characteristic subcellular localizations are also poorly understood. Because RNA represents the direct output of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modification and translation, the generation of such a catalogue is crucial for understanding genome function. Here we report evidence that three-quarters of the human genome is capable of being transcribed, as well as observations about the range and levels of expression, localization, processing fates, regulatory regions and modifications of almost all currently annotated and thousands of previously unannotated RNAs. These observations, taken together, prompt a redefinition of the concept of a gene.

ENCODE - Transcriptome

Djebali et al., 2012

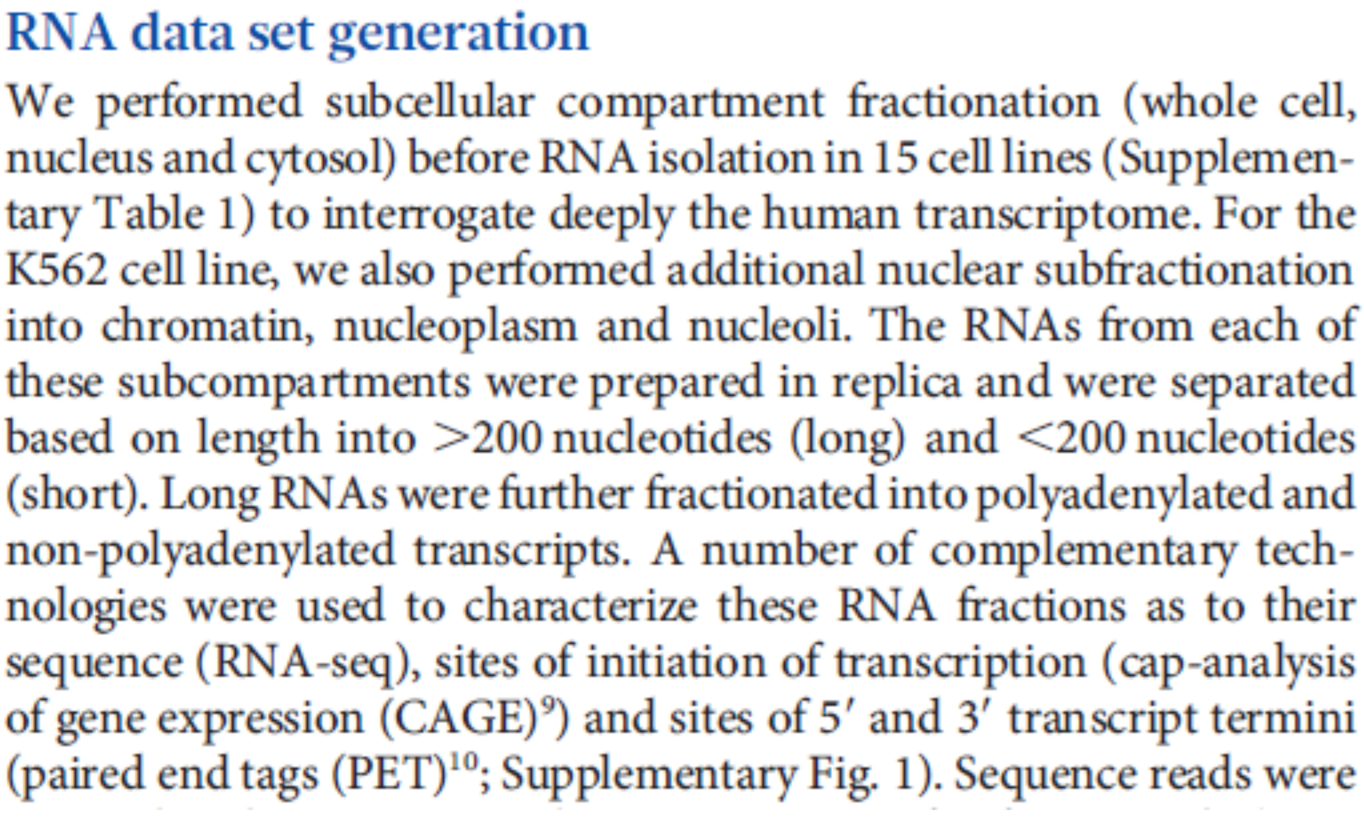
RNA-Seq: identification of annotated and novel RNAs from either of the two major cellular subcompartments (nucleus and cytosol) for 15 cell lines (3 tiers of coverage).

To see the EXPERIMENTAL GRID :
http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html



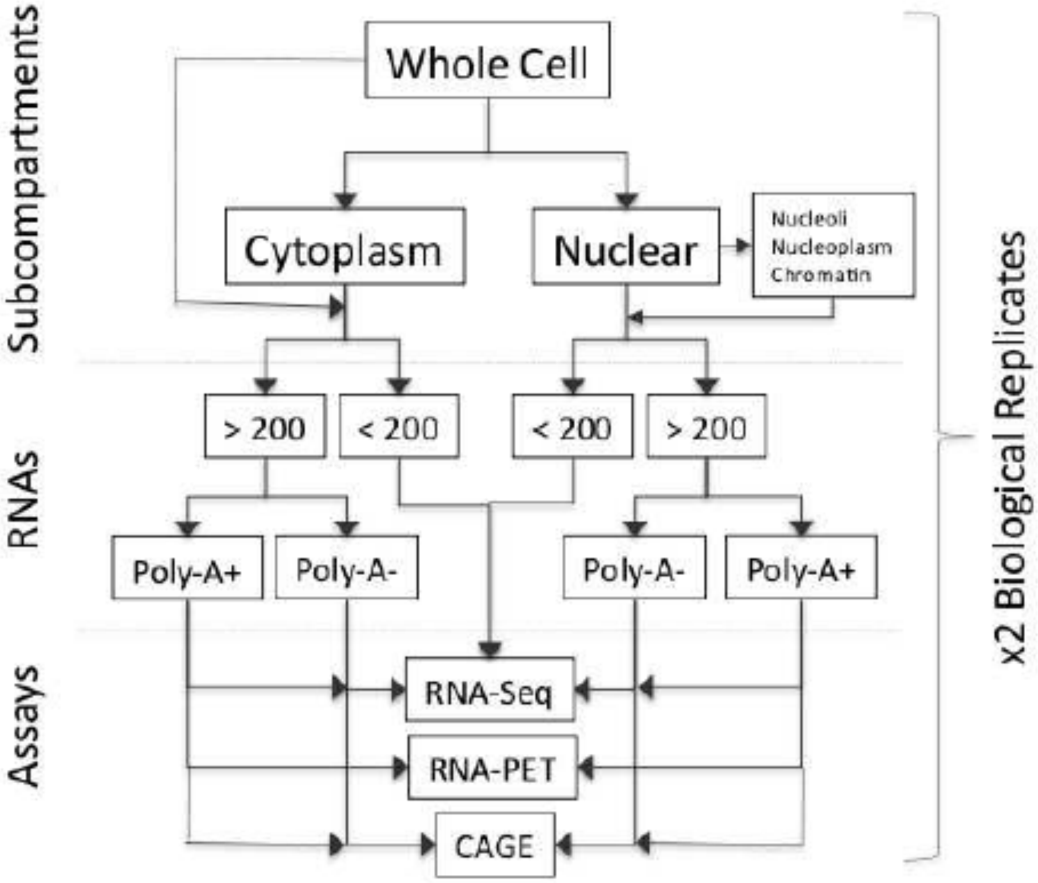
Tier 1. Tier 1 cell types were the highest-priority set and comprised three widely studied cell lines: K562 erythroleukaemia cells; GM12878, a B-lymphoblastoid cell line that is also part of the 1000 Genomes project (http://1000genomes.org)[55]; and the H1 embryonic stem cell (H1 hESC) line.
Tier 2. The second-priority set of cell types in the ENCODE project which included HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs).
Tier 3. Any other ENCODE cell types not in tier 1 or tier 2.

A- Experimental set up



RNA data set generation

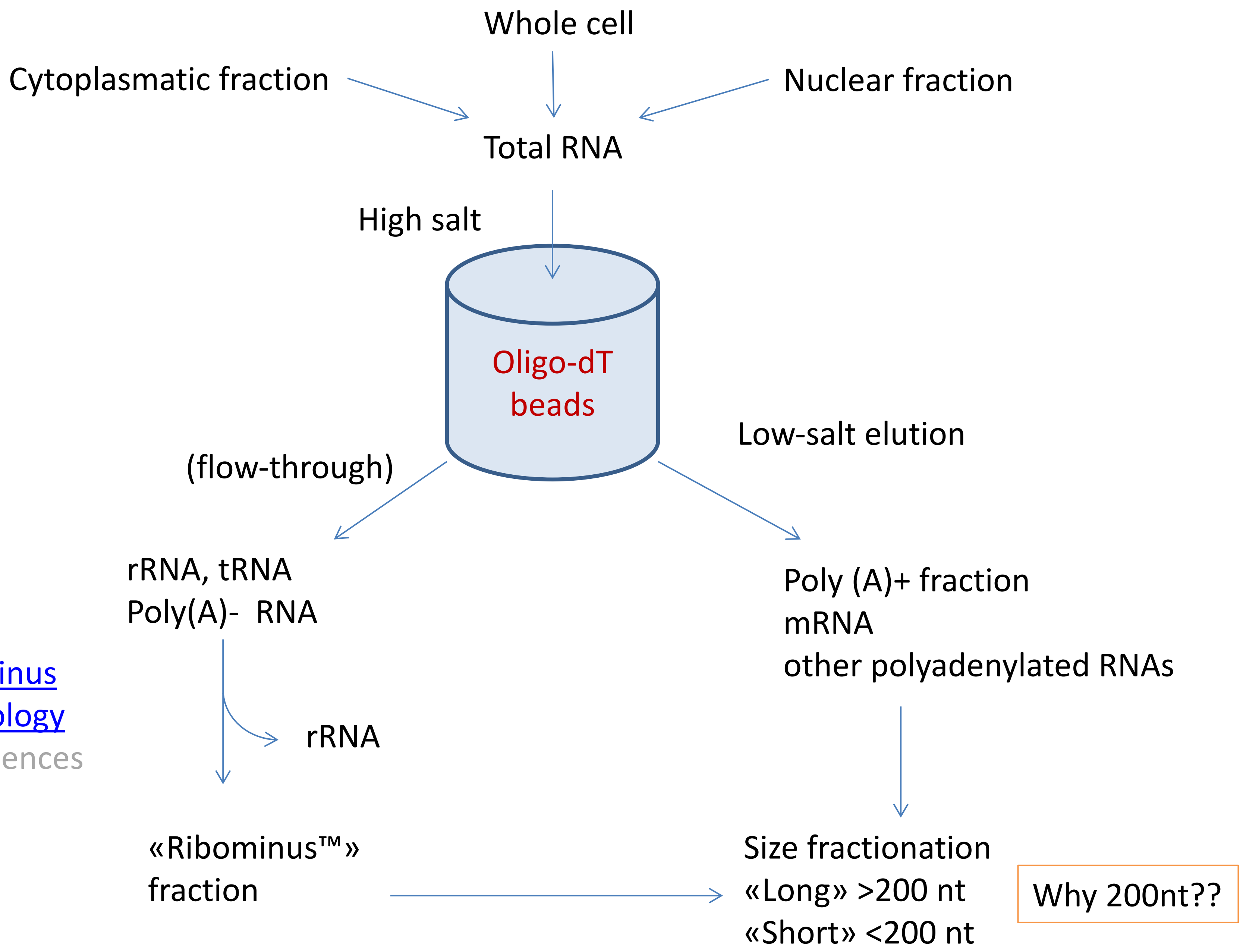
We performed subcellular compartment fractionation (whole cell, nucleus and cytosol) before RNA isolation in 15 cell lines (Supplementary Table 1) to interrogate deeply the human transcriptome. For the K562 cell line, we also performed additional nuclear subfractionation into chromatin, nucleoplasm and nucleoli. The RNAs from each of these subcompartments were prepared in replica and were separated based on length into >200 nucleotides (long) and <200 nucleotides (short). Long RNAs were further fractionated into polyadenylated and non-polyadenylated transcripts. A number of complementary technologies were used to characterize these RNA fractions as to their sequence (RNA-seq), sites of initiation of transcription (cap-analysis of gene expression (CAGE)[9]) and sites of 5′ and 3′ transcript termini (paired end tags (PET)[10]; Supplementary Fig. 1). Sequence reads were



From Supplementary 2

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).



Whole cell

Cytoplasmatic fraction

Nuclear fraction

Total RNA

High salt

Oligo-dT beads

(flow-through)

Low-salt elution

rRNA, tRNA
Poly(A)- RNA

Poly (A)+ fraction
mRNA
other polyadenylated RNAs

rRNA

«Ribominus™» fraction

Size fractionation
«Long» >200 nt
«Short» <200 nt

Why 200nt??

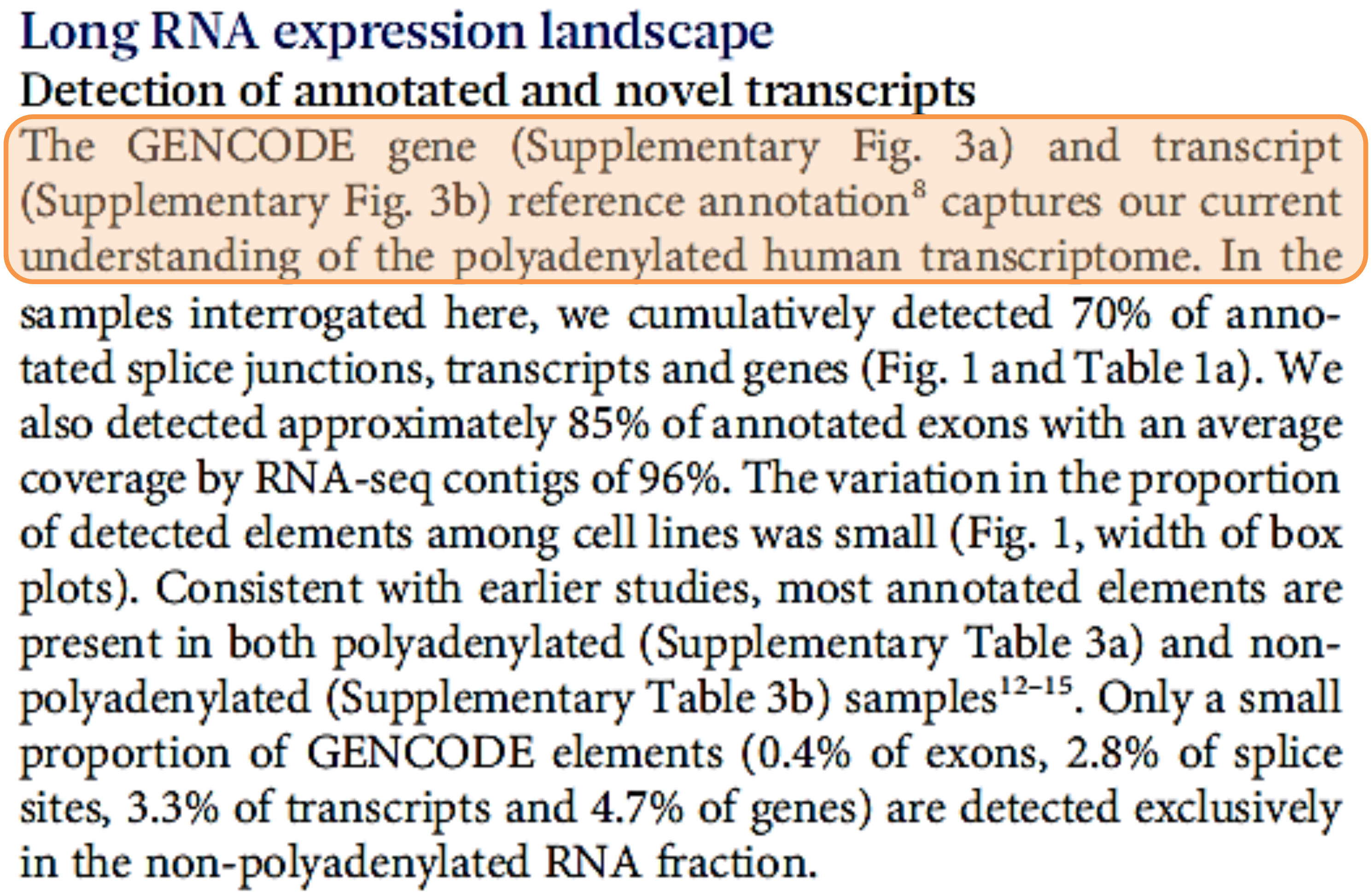
RiboMinus Technology
Life Sciences

We used CAGE-seq (5’ cap-targeted RNA isolation and sequencing) to identify 62,403 transcription start sites (TSSs) at high confidence in **tier 1 and 2 cell** types.

Of these, 27,362 (44%) are within 100 base pairs (bp) of the 5’ end of a GENCODE-annotated transcript or previously reported full-length messenger RNA.

The remaining regions predominantly lie across exons and 3’ untranslated regions (UTRs), and some exhibit cell-type-restricted expression; these may represent the start sites of novel, cell-type-specific transcripts

B- Transcripts annotation



Long RNA expression landscape

Detection of annotated and novel transcripts

The GENCODE gene (Supplementary Fig. 3a) and transcript (Supplementary Fig. 3b) reference annotation[8] captures our current understanding of the polyadenylated human transcriptome. In the samples interrogated here, we cumulatively detected 70% of annotated splice junctions, transcripts and genes (Fig. 1 and Table 1a). We also detected approximately 85% of annotated exons with an average coverage by RNA-seq contigs of 96%. The variation in the proportion of detected elements among cell lines was small (Fig. 1, width of box plots). Consistent with earlier studies, most annotated elements are present in both polyadenylated (Supplementary Table 3a) and non-polyadenylated (Supplementary Table 3b) samples[12–15]. Only a small proportion of GENCODE elements (0.4% of exons, 2.8% of splice sites, 3.3% of transcripts and 4.7% of genes) are detected exclusively in the non-polyadenylated RNA fraction.

Wiki

Why do the Authors refer to «**GENCODE**» genes ?

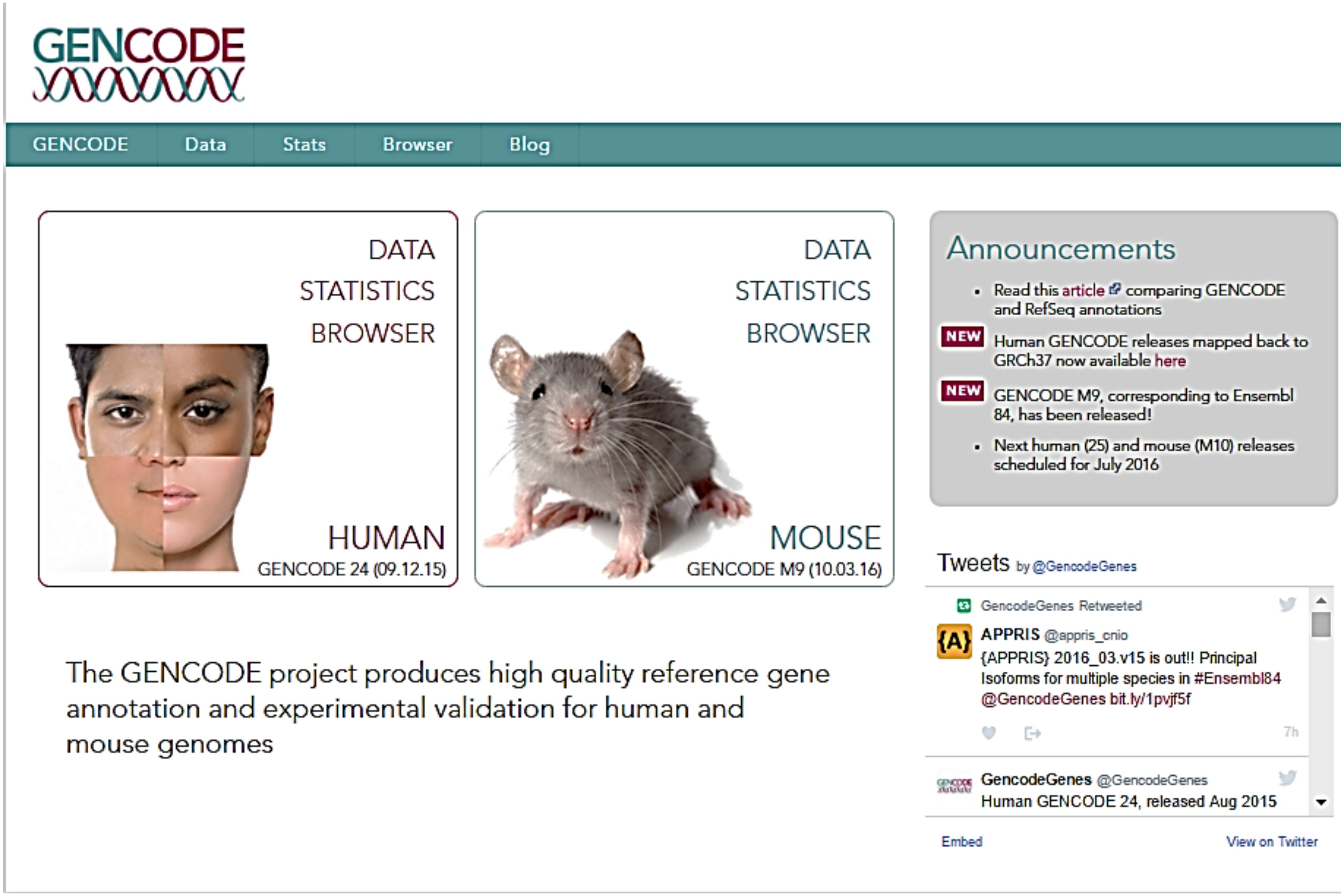
GENCODE is the bioinformatics and database gemini project of ENCODE.

GENCODE collects information from ENCODE and other projects in Human and Mouse

Defining regions of the genome that are transcribed: mapping transcripts and other annotation, in order to define coding and noncoding regions.

To define genes, GENECODE integrates RNA-Seq data with other available transcriptomics data and with chromatin and epigenetic data that reinforced the concept of transcribed regions.

Extensive comparison with other databases that used different automathic or manual annotation algorithms (ENSEMBL, Havana).



HUMAN

GENCODE 24 (09.12.15)

MOUSE

GENCODE M9 (10.03.16)

The GENCODE project produces high quality reference gene annotation and experimental validation for human and mouse genomes

Announcements

- Read this article comparing GENCODE and RefSeq annotations
- NEW Human GENCODE releases mapped back to GRCh37 now available here
- NEW GENCODE M9, corresponding to Ensembl 84, has been released!
- Next human (25) and mouse (M10) releases scheduled for July 2016

Tweets by @GencodeGenes

GencodeGenes Retweeted

APPRIS @appris_cnio

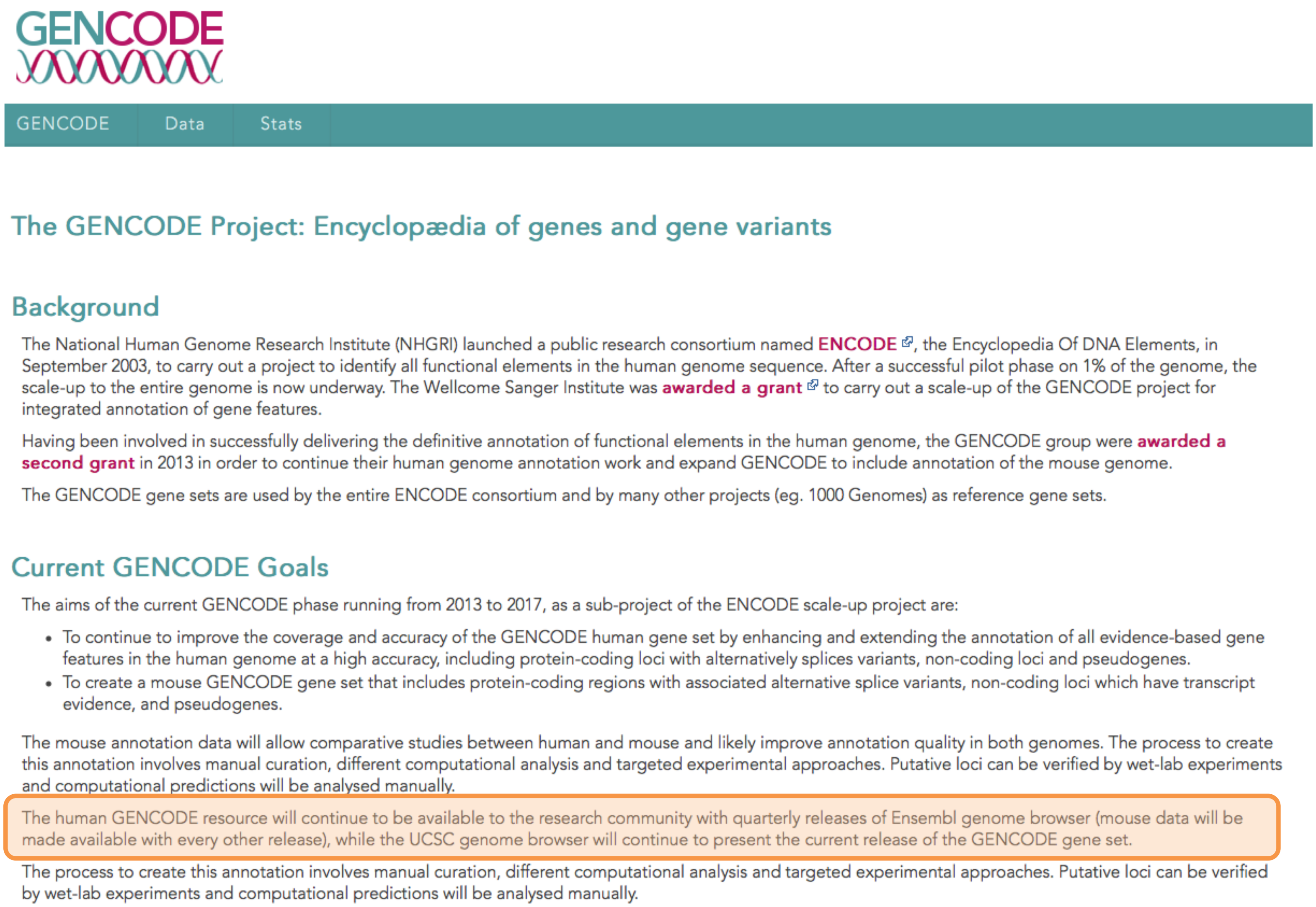
{APPRIS} 2016_03.v15 is out!! Principal Isofoms for multiple species in #Ensembl84 @GencodeGenes bit.ly/1pvjf5f

7h

GencodeGenes @GencodeGenes

Human GENCODE 24, released Aug 2015

http://www.gencodegenes.org/



The GENCODE Project: Encyclopædia of genes and gene variants

Background

The National Human Genome Research Institute (NHGRI) launched a public research consortium named **ENCODE**, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. After a successful pilot phase on 1% of the genome, the scale-up to the entire genome is now underway. The Wellcome Sanger Institute was **awarded a grant** to carry out a scale-up of the GENCODE project for integrated annotation of gene features.

Having been involved in successfully delivering the definitive annotation of functional elements in the human genome, the GENCODE group were **awarded a second grant** in 2013 in order to continue their human genome annotation work and expand GENCODE to include annotation of the mouse genome.

The GENCODE gene sets are used by the entire ENCODE consortium and by many other projects (eg. 1000 Genomes) as reference gene sets.

Current GENCODE Goals

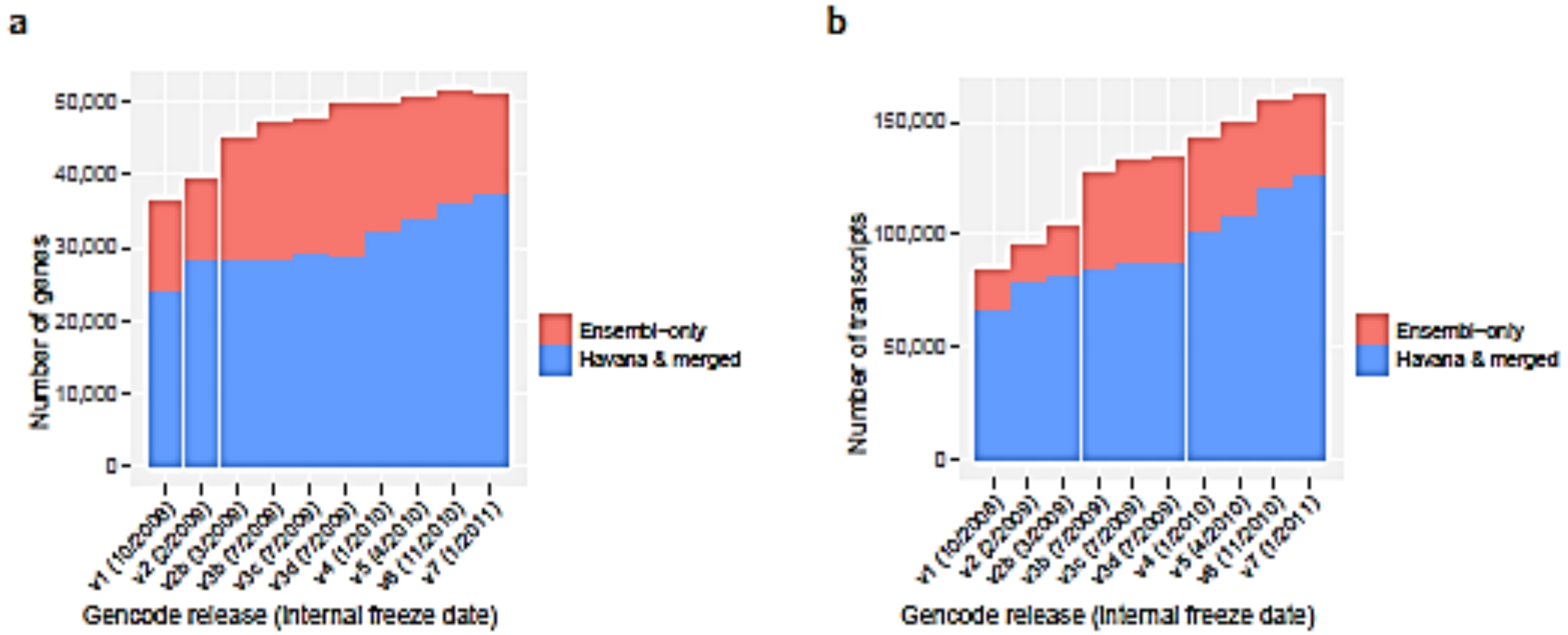
The aims of the current GENCODE phase running from 2013 to 2017, as a sub-project of the ENCODE scale-up project are:

- To continue to improve the coverage and accuracy of the GENCODE human gene set by enhancing and extending the annotation of all evidence-based gene features in the human genome at a high accuracy, including protein-coding loci with alternatively splices variants, non-coding loci and pseudogenes.
- To create a mouse GENCODE gene set that includes protein-coding regions with associated alternative splice variants, non-coding loci which have transcript evidence, and pseudogenes.

The mouse annotation data will allow comparative studies between human and mouse and likely improve annotation quality in both genomes. The process to create this annotation involves manual curation, different computational analysis and targeted experimental approaches. Putative loci can be verified by wet-lab experiments and computational predictions will be analysed manually.

The human GENCODE resource will continue to be available to the research community with quarterly releases of Ensembl genome browser (mouse data will be made available with every other release), while the UCSC genome browser will continue to present the current release of the GENCODE gene set.

The process to create this annotation involves manual curation, different computational analysis and targeted experimental approaches. Putative loci can be verified by wet-lab experiments and computational predictions will be analysed manually.



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.

Other Gene Annotation Projects

There are several large-scale gene annotation projects in progress on the human genome, including RefSeq (Pruitt et al. 2005), GENCODE (Harrow et al. 2012), and UCSC Genes (Dreszer et al. 2012).

In each ''gene set'' or ''genebuild'' produced, the vast majority of models are based upon transcriptomics data.

GENCODE (the gene set of the ENCODE project) represents a merge between manually annotated HAVANA and computationally derived Ensembl models, with annotation taking place on the genome sequence.

RefSeq also combines manual and automated processes, most human annotation takes place on full-length cDNAs that are subsequently linked to the chromosome.

UCSC Genes combine RefSeq models mapped to the genome with additional models from other data sources, for example computational models based on GenBank ESTs.

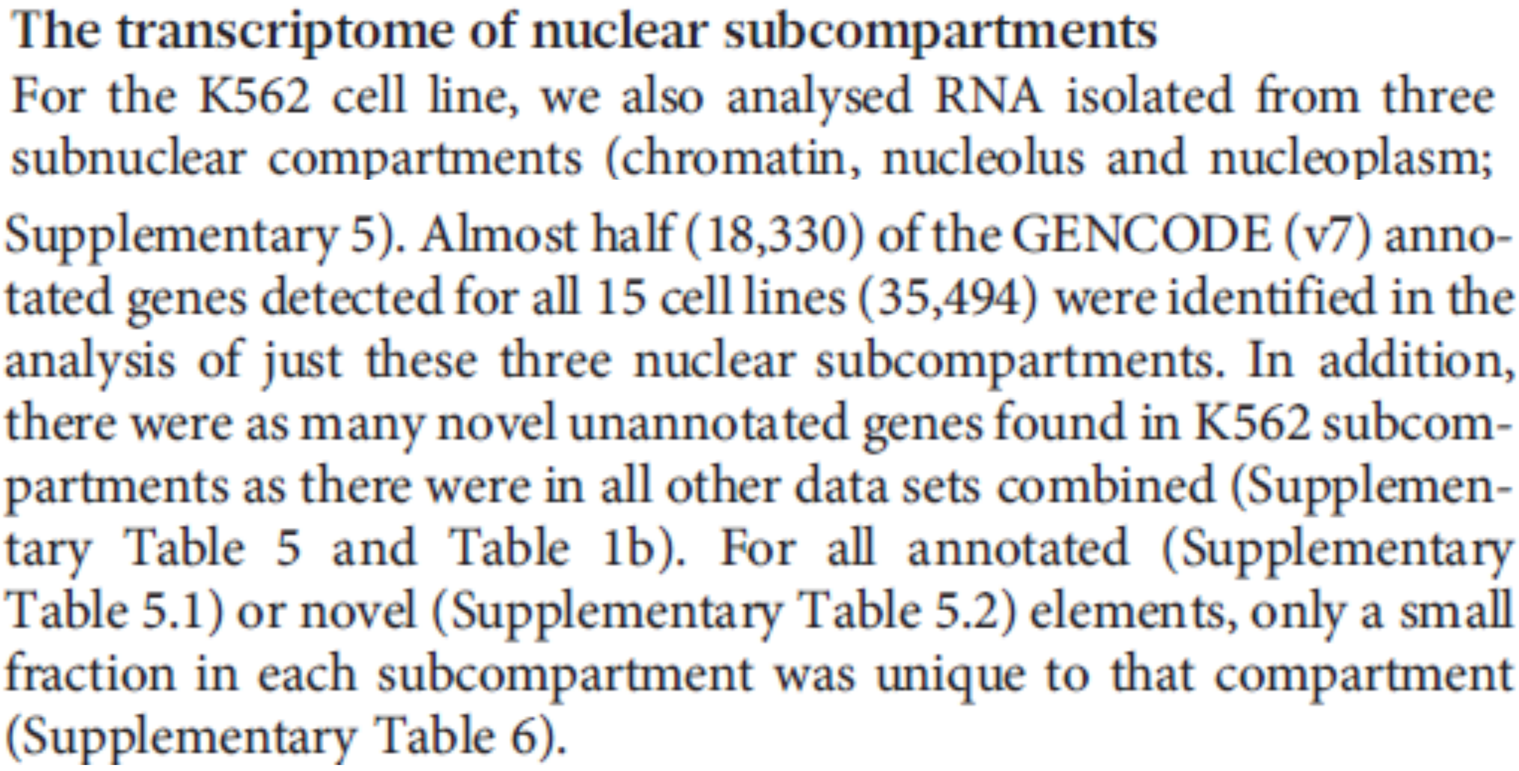
To see GENECODE model, go to http://www.gencodegenes.org/

To see Transcripts, go to: ENCODE data in the UCSC browser (http://genome.ucsc.edu/ENCODE/index.html)

Or other Browsers:

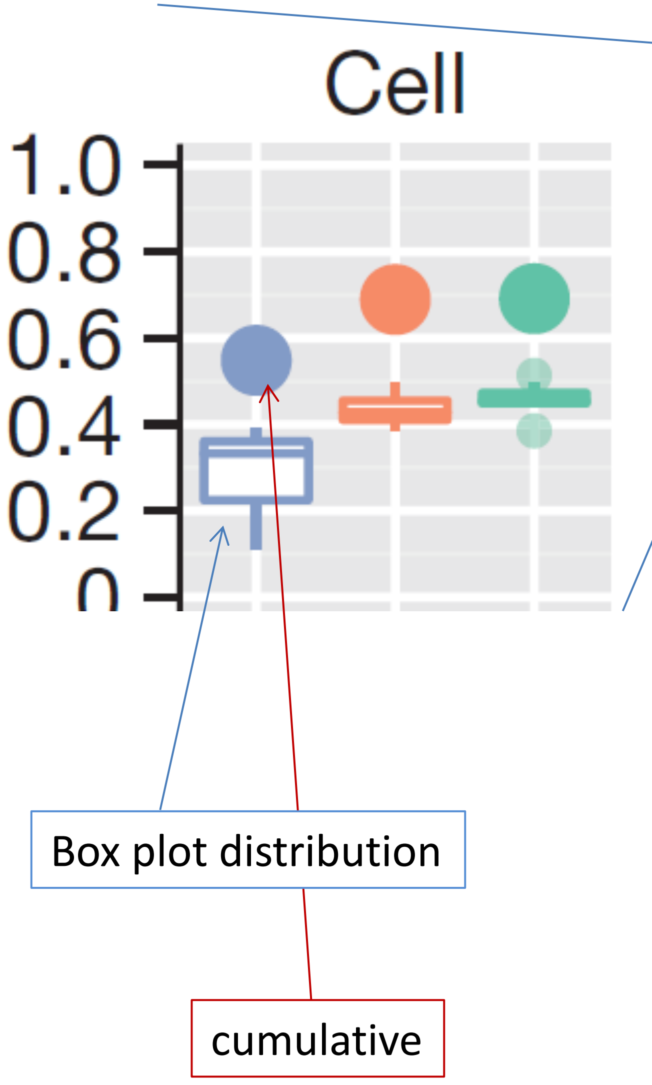
ENSEMBL	(http://www.ensembl.org/index.html)
NCBI	(http://www.ncbi.nlm.nih.gov/gene)
UCSC	(https://genome-euro.ucsc.edu)
WASHU	(http://epigenomegateway.wustl.edu/)

C- Subcellular localization



The transcriptome of nuclear subcompartments

For the K562 cell line, we also analysed RNA isolated from three subnuclear compartments (chromatin, nucleolus and nucleoplasm; Supplementary 5). Almost half (18,330) of the GENCODE (v7) annotated genes detected for all 15 cell lines (35,494) were identified in the analysis of just these three nuclear subcompartments. In addition, there were as many novel unannotated genes found in K562 subcompartments as there were in all other data sets combined (Supplementary Table 5 and Table 1b). For all annotated (Supplementary Table 5.1) or novel (Supplementary Table 5.2) elements, only a small fraction in each subcompartment was unique to that compartment (Supplementary Table 6).



Box plot distribution

cumulative

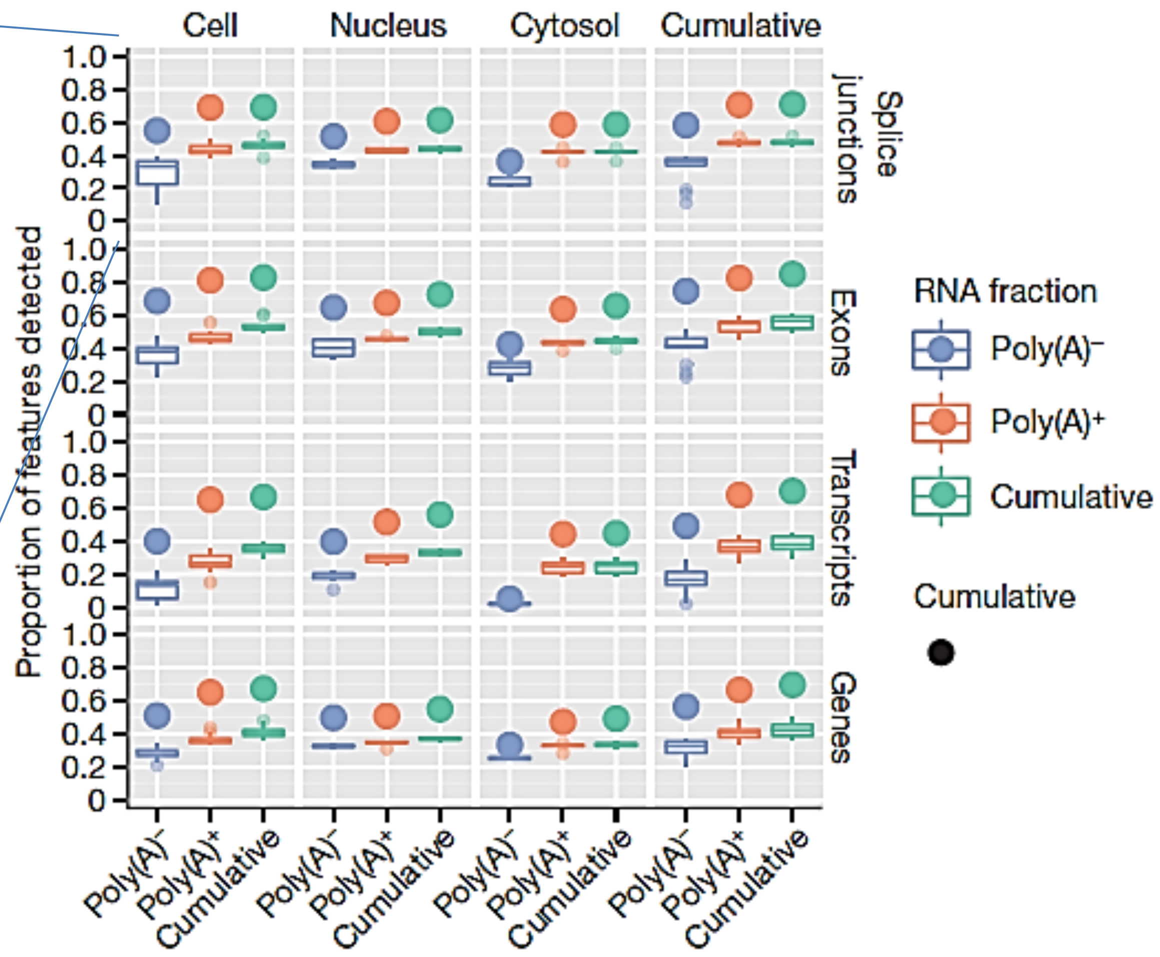
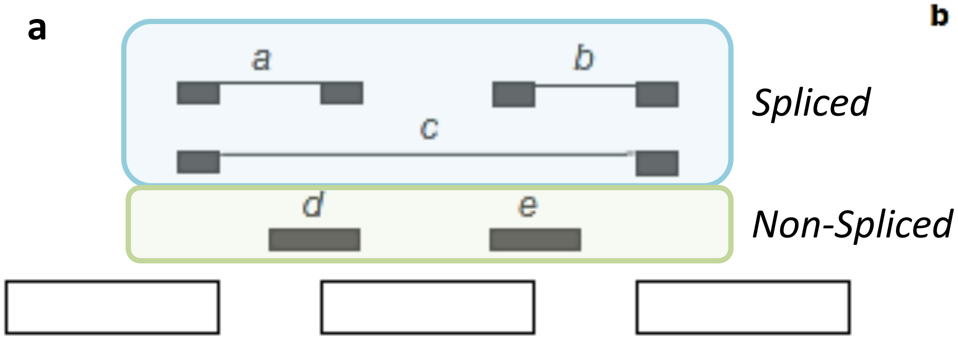


Figure 1 | A large majority of GENCODE elements are detected by RNA-seq data. Shown are GENCODE-detected elements in the polyadenylated and non-polyadenylated fractions of cellular compartments (cumulative counts for both RNA fractions and compartments refer to elements present in any of the fractions or compartments). Each box plot is generated from values across all cell lines, thus capturing the dispersion across cell lines. The largest point shows the cumulative value over all cell lines.



coSI definition as ratio between junction reads and exon-intron reads

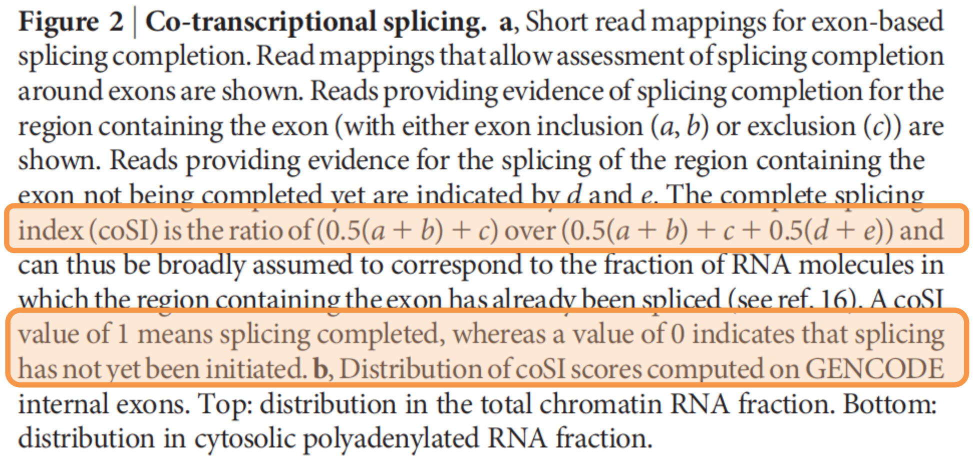
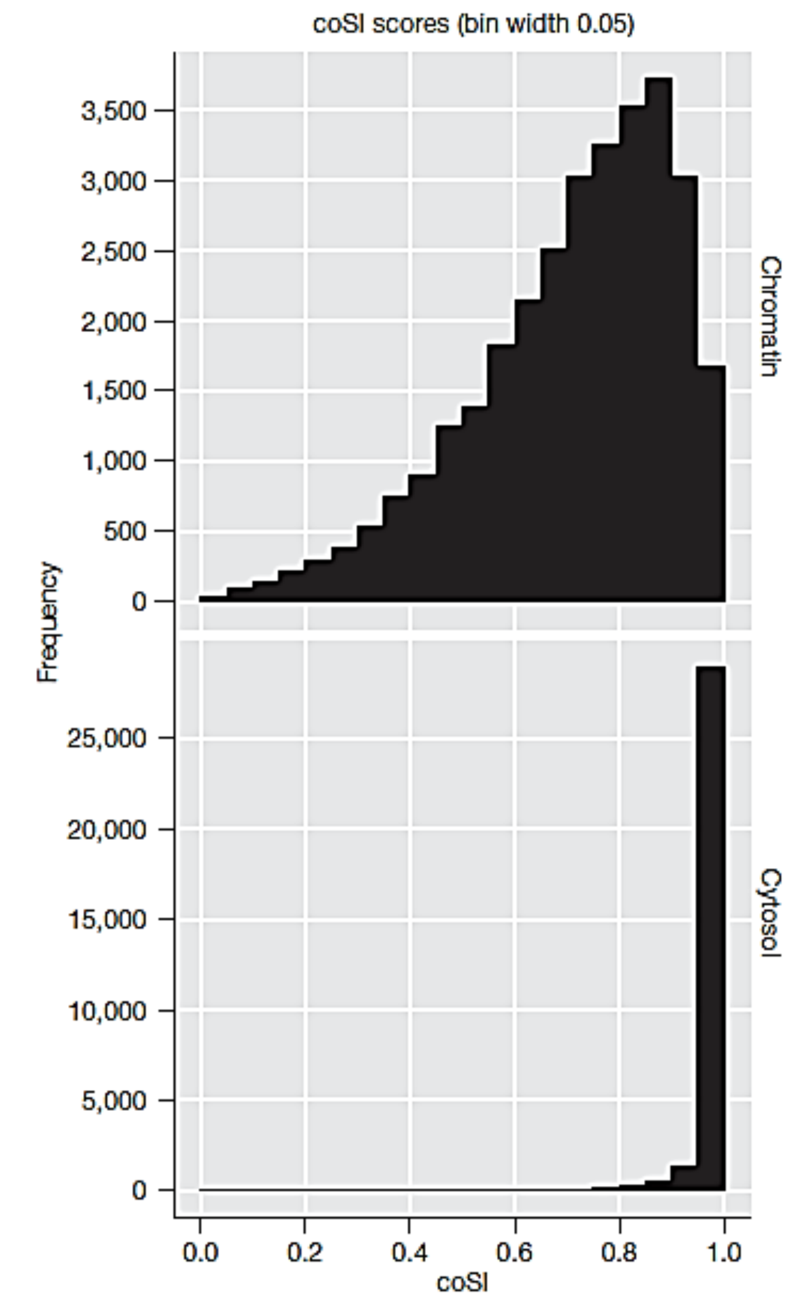


Figure 2 | Co-transcriptional splicing. a, Short read mappings for exon-based splicing completion. Read mappings that allow assessment of splicing completion around exons are shown. Reads providing evidence of splicing completion for the region containing the exon (with either exon inclusion (a, b) or exclusion (c)) are shown. Reads providing evidence for the splicing of the region containing the exon not being completed yet are indicated by d and e. The complete splicing index (coSI) is the ratio of $(0.5(a + b) + c)$ over $(0.5(a + b) + c + 0.5(d + e))$ and can thus be broadly assumed to correspond to the fraction of RNA molecules in which the region containing the exon has already been spliced (see ref. 16). A coSI value of 1 means splicing completed, whereas a value of 0 indicates that splicing has not yet been initiated. **b**, Distribution of coSI scores computed on GENCODE internal exons. Top: distribution in the total chromatin RNA fraction. Bottom: distribution in cytosolic polyadenylated RNA fraction.



D- Transcripts expression

Expression level - quantity

Transcripts range in a 6-order magnitude (poly A+)(10^{-2} to 10^{4} rpkm) or 5 orders of magnitude (poly A-) (10^{-2} to 10^{3} rpkm)

Assuming that 1–4 r.p.k.m. approximates to 1 copy per cell (*Montazavi et al., 2008*):

- one quarter of protein-coding RNAs *and*
- 80% of long noncoding RNAs (lncRNA)

Are expressed at 1 or <1 molecules per cell

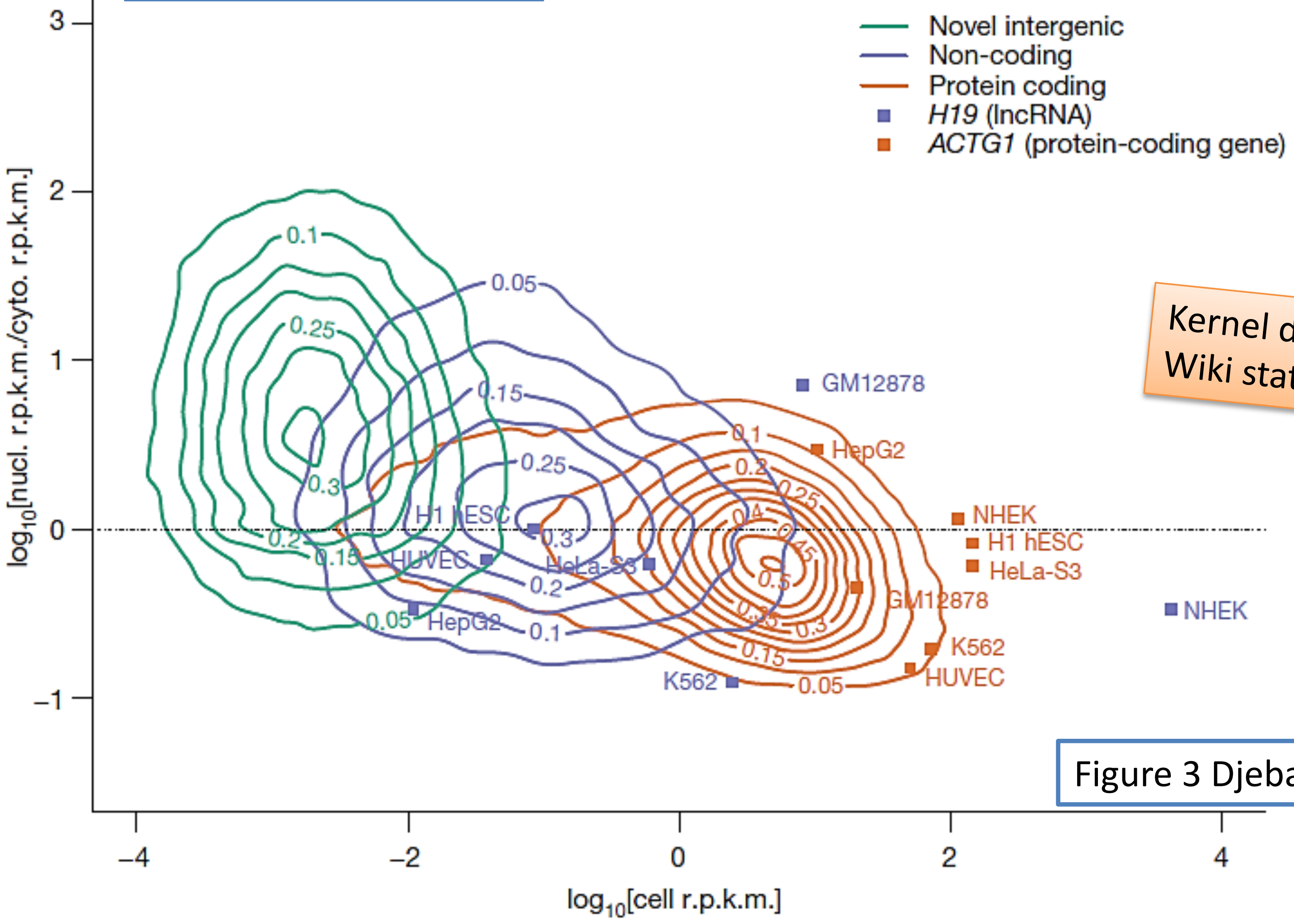
i.e. the majority of lncRNAs are expressed at a very low level

Novel lncRNAs discovered here contains also a class showing *rpkm* from 10^{-4} to 10^{-1}: << extremely low expression >>

Question: what does it mean «less than one molecule per cell ?

Djebali et al., 2012

Expresson level by class



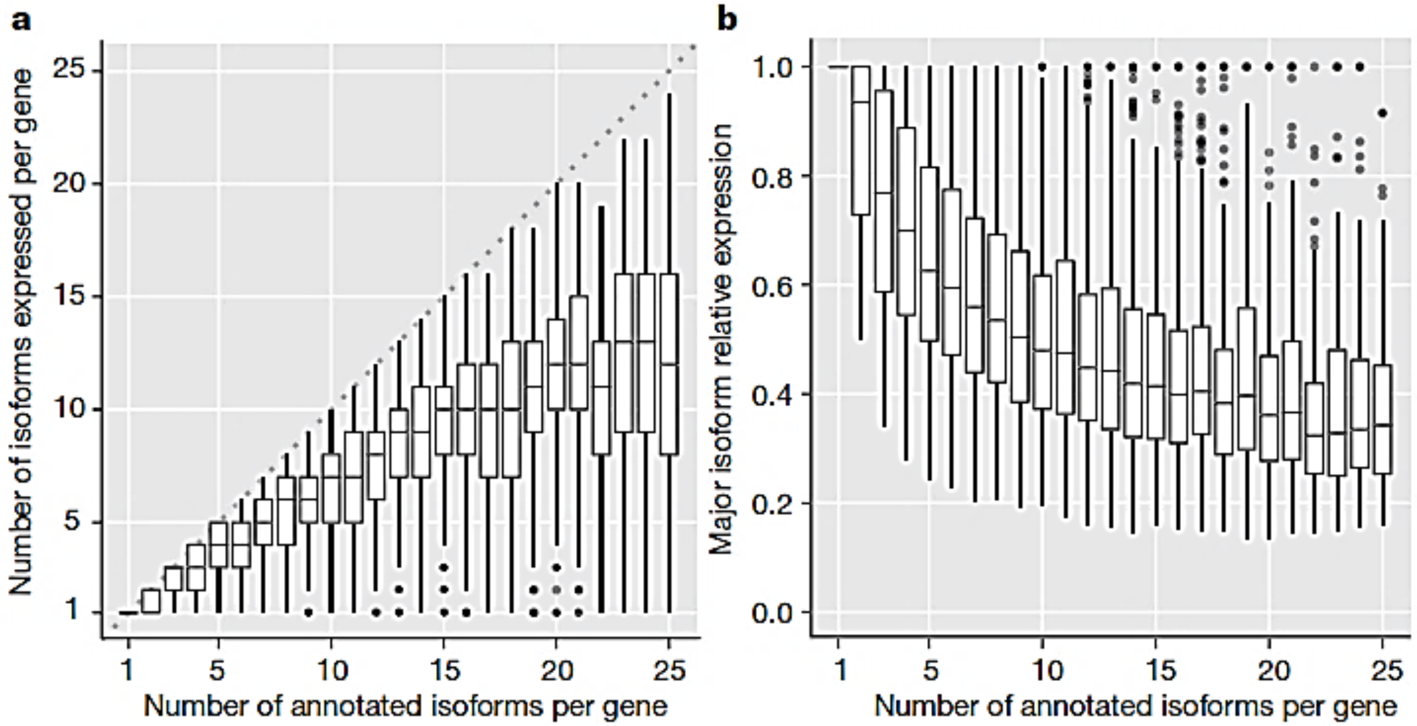
Kernel density
Wiki statistics

Figure 3 Djebali et al. 2012

Protein coding transcripts are the only class that is enriched in the cytoplasm

E- Splicing Patterns

Djebali 2012, Figure 4 - Isoforms (alternative splicing)



Box plot

Wiki statistics

a) Number of expressed isoforms per gene per cell line. A plateau is evident between 10 and 12

b) Relative expression of the most abundant isoform per gene per cell line.

F- TSS/TESes

Alternative transcription initiation and termination.

On the basis of RNA-seq analysis of polyadenylated RNAs, a total of 128,021 TSSs were detected across all cell lines, of which 97,778 were previously annotated and 30,243 were novel intergenic/antisense TSSs.

CAGE tags…. identified a total of 82,783 nonredundant TSSs

Approximately 48% of the CAGE-identified TSSs are located within 500 base pairs (bp) of an annotated RNA-seq-detected GENCODE TSS, whereas an additional 3% are within 500 bp of a novel TSS.

Integration of data from ChIP-Seq of histone PTMs demonstrated that most TSS do not have the characteristic chromatin features associated to transcription. This suggests alternative modality of transcriptional regulation at promoters.

« Low concordance between models »

Djebali et al., 2012

Other topics

G- Short RNA

H- RNA Editing and Repeats

Brief discussion tomorrow

I- Enhancer RNA

eRNA will be discussed in Part 4

Djebali et al., 2012

ENCODE - Transcriptome

Djebali et al., 2012

General conclusions:

- 62.1% of genome covered by processed transcripts; 74.7% by unprocessed transcripts.
  -> New gene definition
- Novel elements cover 78% of intronic nucleotides and 34% of intergenic sequences.
- Multiple isoforms per gene expressed simultaneously, with a plateau at 10-12 isoforms per gene per cell line.
- eRNA – transcripts starting from enhancers
- 6% of coding and noncoding overlap with small RNA (probably precursors)

Question: is this feature «conclusive» ?