

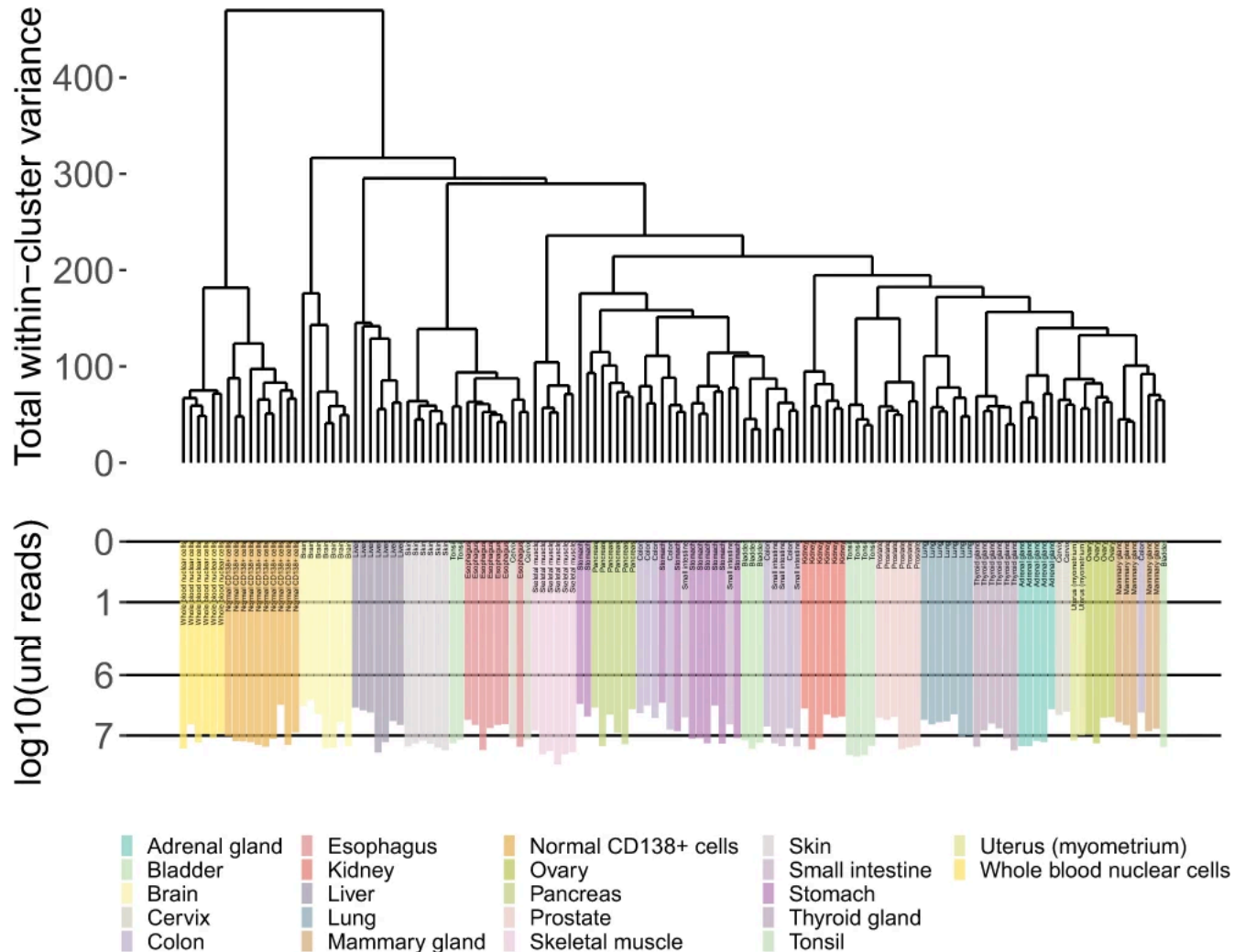
# RNA introduction

---

RNA-seq data analysis

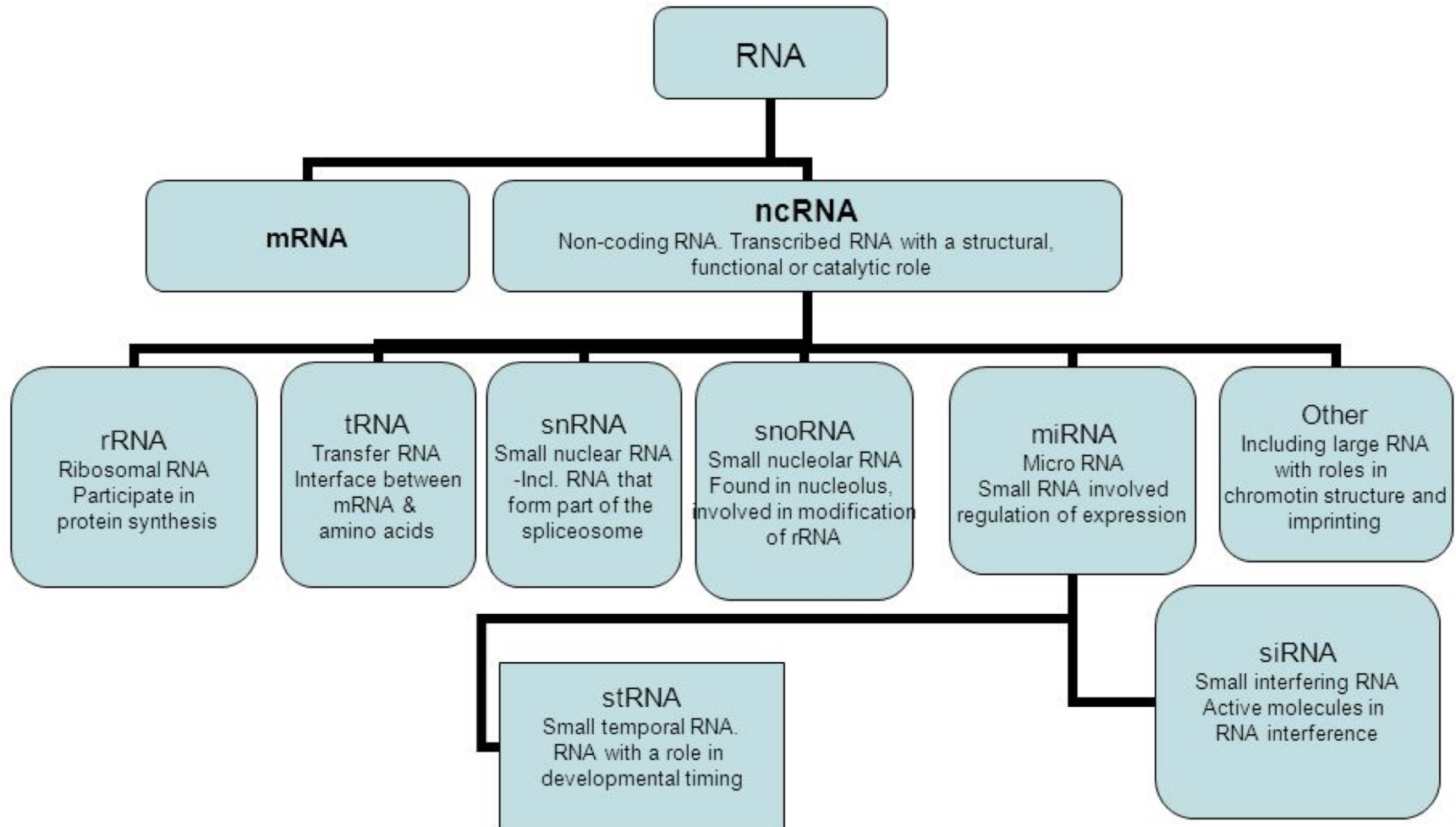
Johan Reimegård | 30-November-2020

DNA is the same in all cells  
 RNAs are different in all cells



Atlas of RNA sequencing profiles for normal human tissues (Scientific data 2019)

# There is a wide variety of different functional RNAs

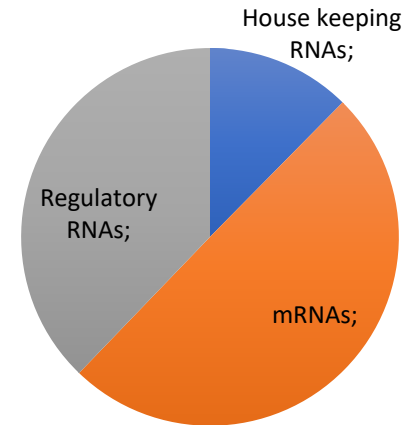


# RNA flavors

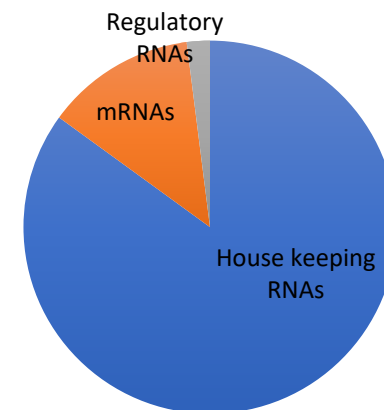
- now

- House keeping RNAs
  - rRNAs, tRNAs, snoRNAs, snRNAs, SRP RNAs, Catalytic RNAs (RNase E)
- Protein coding RNAs
  - 1 coding gene – many mRNAs)
- Regulatory RNAs
  - sRNAs, CRIPSR, miRNAs, piRNAs, lincRNAs, Riboswitches ....

**Variants**



**Abundance**



**Landscape of transcription in human cells, S Djebali et al. *Nature* 2012**



The screenshot displays the ENCODE Explorer website interface. The top navigation bar includes the 'nature' logo, the 'ENCODE' title, and a search bar with the text 'Search ENCODE' and a 'Go' button. Below the navigation bar, there are links for 'Home', 'Research', 'Threads', 'Additional Research', 'News and Comment', 'About', and 'Sponsor'. The main content area is divided into two panels. The left panel, titled 'nature ENCODE explorer' and 'THREADS', features a circular sequence of 13 numbered nodes (01-13) in various colors. The right panel, titled 'PAPERS' and 'PRODUCED WITH SUPPORT FROM illumina', shows a circular arrangement of document icons with the letter 'n' on them. The background of the main content area is a dark grid with a colorful circular genomic map.

ENCODE, the Encyclopedia of DNA Elements, is a project funded by the National Human Genome Research Institute to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence.

# ENCyclopedia Of Dna Elements

## 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

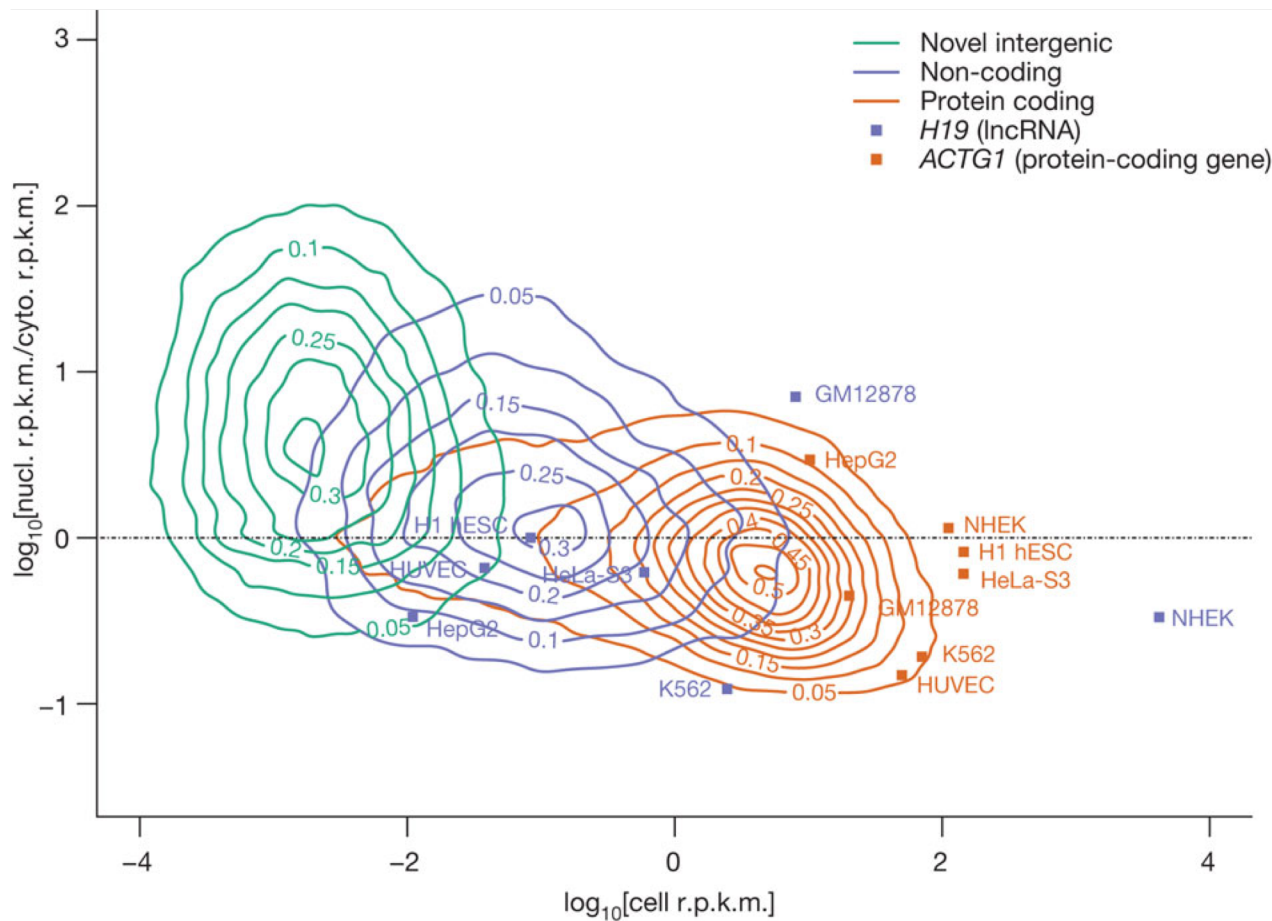
# ENCyclopedia Of Dna Elements

## 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

Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts, respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines.

# Coding genes are more highly expressed than non-coding



# RNA flavors

OPEN ACCESS Freely available online

PLoS BIOLOGY

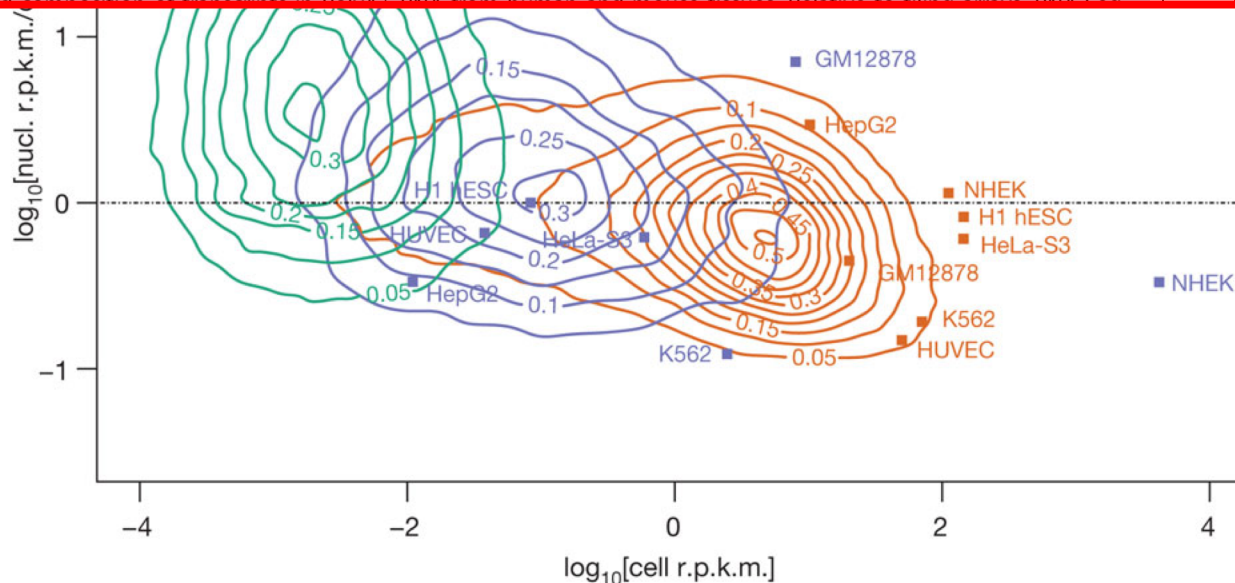
## Most “Dark Matter” Transcripts Are Associated With Known Genes

Harm van Bakel<sup>1</sup>, Corey Nislow<sup>1,2</sup>, Benjamin J. Blencowe<sup>1,2</sup>, Timothy R. Hughes<sup>1,2\*</sup>

<sup>1</sup> Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, <sup>2</sup> Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

### Abstract

A series of reports over the last few years have indicated that a much larger portion of the mammalian genome is transcribed than can be accounted for by currently annotated genes, but the quantity and nature of these additional transcripts remains unclear. Here, we have used data from single- and paired-end RNA-Seq and tiling arrays to assess the





# RNA flavors

OPEN ACCESS Freely available online

PLoS BIOLOGY

## Most “Dark Matter” Transcripts Are Associated With

K

Ha

1 Ba

Ont

OPEN ACCESS Freely available online

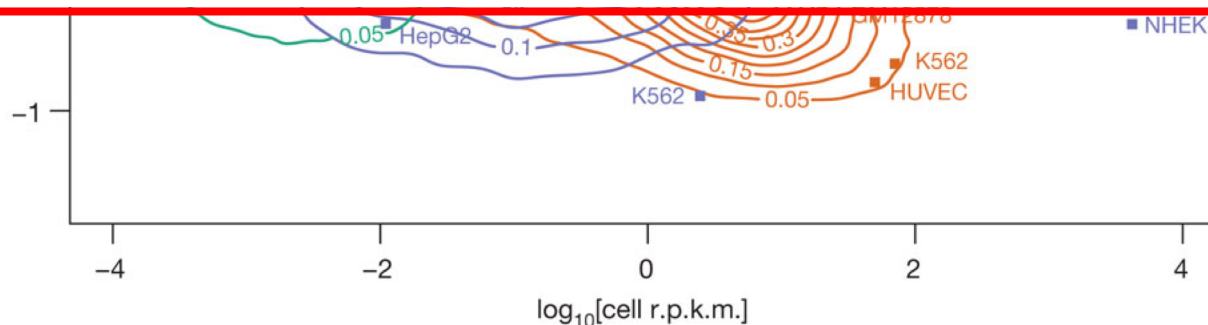
PLoS BIOLOGY

### Perspective

## The Reality of Pervasive Transcription

**Michael B. Clark<sup>1</sup>, Paulo P. Amaral<sup>1,9</sup>, Felix J. Schlesinger<sup>2,9</sup>, Marcel E. Dinger<sup>1</sup>, Ryan J. Taft<sup>1</sup>, John L. Rinn<sup>3</sup>, Chris P. Ponting<sup>4</sup>, Peter F. Stadler<sup>5</sup>, Kevin V. Morris<sup>6</sup>, Antonin Morillon<sup>7</sup>, Joel S. Rozowsky<sup>8</sup>, Mark B. Gerstein<sup>8</sup>, Claes Wahlestedt<sup>9</sup>, Yoshihide Hayashizaki<sup>10</sup>, Piero Carninci<sup>10</sup>, Thomas R. Gingeras<sup>2,\*</sup>, John S. Mattick<sup>1,\*</sup>**

**1** Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia, **2** Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **3** Broad Institute, Cambridge, Massachusetts, United States of America, **4** MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom, **5** Department of Computer Science, University of Leipzig, Leipzig, Germany, **6** Department of Molecular and Experimental Medicine, Scripps Research Institute, La Jolla, California, United States of America, **7** Institut Curie, UMR3244-Pavillon Trouillot Rossignol, Paris, France, **8** Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **9** University of Miami, Miami, Florida, United States of America, **10** Omics Science Center, RIKEN Yokohama Institute, Tsurumi-ku, Yokohama, Kanagawa, Japan



# RNA flavors

OPEN ACCESS Freely available online

PLoS BIOLOGY

## Most “Dark Matter” Transcripts Are Associated With

K

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Ha

1 Ba

Ont

## The Reality of Pervasive Transcription

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Micha

Rinn<sup>3</sup>

Mark

John

1 Institut

Cold Spr

Departm

Germany

Pavillon

Miami, M

## Response to “The Reality of Pervasive Transcription”

Harm van Bakel<sup>1</sup>, Corey Nislow<sup>1,2</sup>, Benjamin J. Blencowe<sup>1,2</sup>, Timothy R. Hughes<sup>1,2\*</sup>

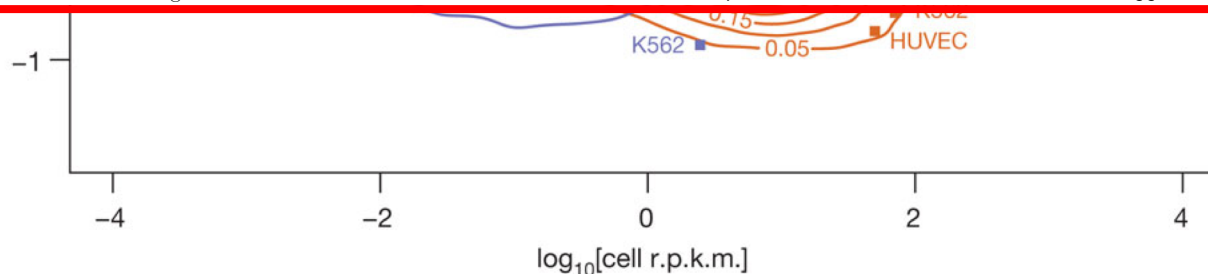
**1** Banting and Best Department of Medical Research and Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

Clark et al. criticize several aspects of our study [1], and specifically challenge our assertion that the degree of pervasive transcription has previously been overstated. We disagree with much of their

“tic” transcripts greatly increases their abundance [7,8].

We acknowledge that the phrase quoted by Clark et al. in our Author Summary should have read “stably transcribed”, or

emphasized the lack of abundant pervasive transcription in our study. Clark et al. cite papers that have previously documented pervasive transcription, and point out that several different approaches have been



# RNA flavors

OPEN ACCESS Freely available online

PLoS BIOLOGY

ance

## Most "Dark Matter" Transcripts Are Associated With

K

Ha

1 Ba

Ont

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

## The Reality of Pervasive Transcription

Micha  
Rinn

Mark  
John

1 Institut  
Cold Spr  
Departm  
Germany  
Pavillon  
Miami, M

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Re

Ha

1 Ba  
Cana

avigation

- Home
- Research
- People
- Publications
- Links
- Internal Home
- Contact

inks

- QuARC
- NRED
- IncRNAdb



mattick

### Comments on van Bakel et al. (2011) Response to "The Reality of Pervasive Transcription"

Comments by [Mike Clark](#)

[Van Bakel et al. 2011](#) (vB 11) have published their reply to our [critique](#) of their paper [van Bakel et al. 2010](#) (vB 10).

Firstly lets briefly review some of our main criticisms of vB 10:

- vB 10 didn't properly consider previous evidence for pervasive transcription (especially that from cDNA analysis in the mouse) when claiming the genome was not as transcribed as previous evidence was unreliable due to false positives.
- vB 10 incorrectly conflated pervasive transcription with the relative abundance of transcripts when the correct (and known) definition was the amount of the genome that was transcribed.
- The tiling arrays vB 10 performed and then used to claim that previous array studies suffered from high false positives were atypical and lacked any validation of the false positive rate.
- The RNA sequencing carried out by vB 10 was severely limited in its ability to address the question of pervasive transcription. The depth of sequencing was too shallow to detect rare transcripts and the assembly of what was found into transcripts was poor. Since it couldn't detect and/or characterize rare transcripts this meant it couldn't even detect the presence of pervasive transcription.
- vB 10 claimed that low level intergenic transcription may be due to "random initiation events" and/or transcriptional "byproducts" (ie: transcription noise), when the limitation was to properly differentiate between this and genuine transcripts under their detection threshold.

Novel intergenic  
transcription  
(coding  
RNA)  
(protein-coding gene)

# Defining functional DNA elements in the human genome

A priori, we should not expect the transcriptome to consist exclusively of functional RNAs.

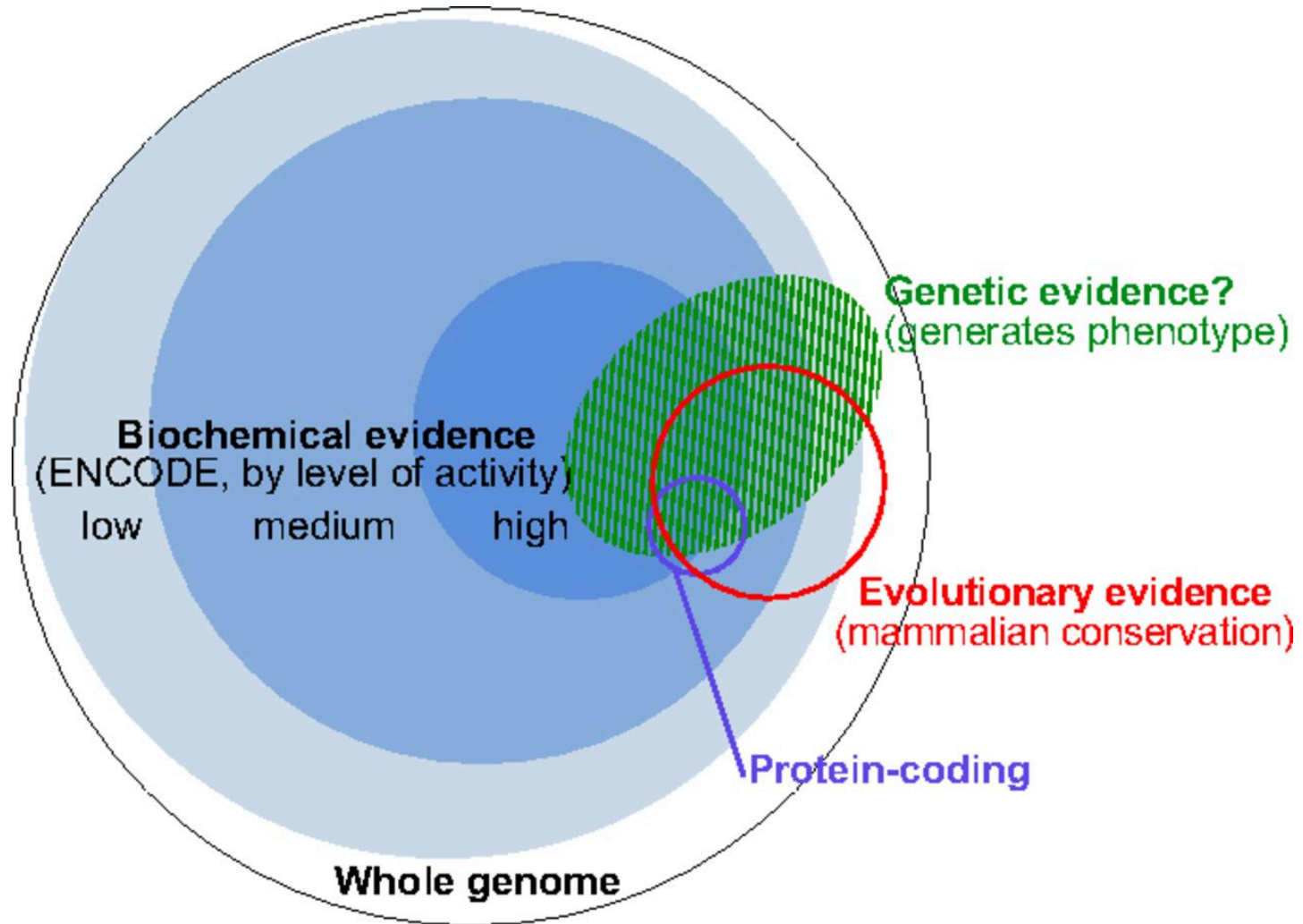
Zero tolerance for errant transcripts would come at high cost in the proofreading machinery needed to perfectly gate RNA polymerase and splicing activities, or to instantly eliminate spurious transcripts.

In general, sequences encoding RNAs transcribed by noisy transcriptional machinery are expected to be less constrained, which is consistent with data shown here for very low abundance RNA

Thus, one should have high confidence that the subset of the genome with large signals for RNA or chromatin signatures coupled with strong conservation is functional and will be supported by appropriate genetic tests.

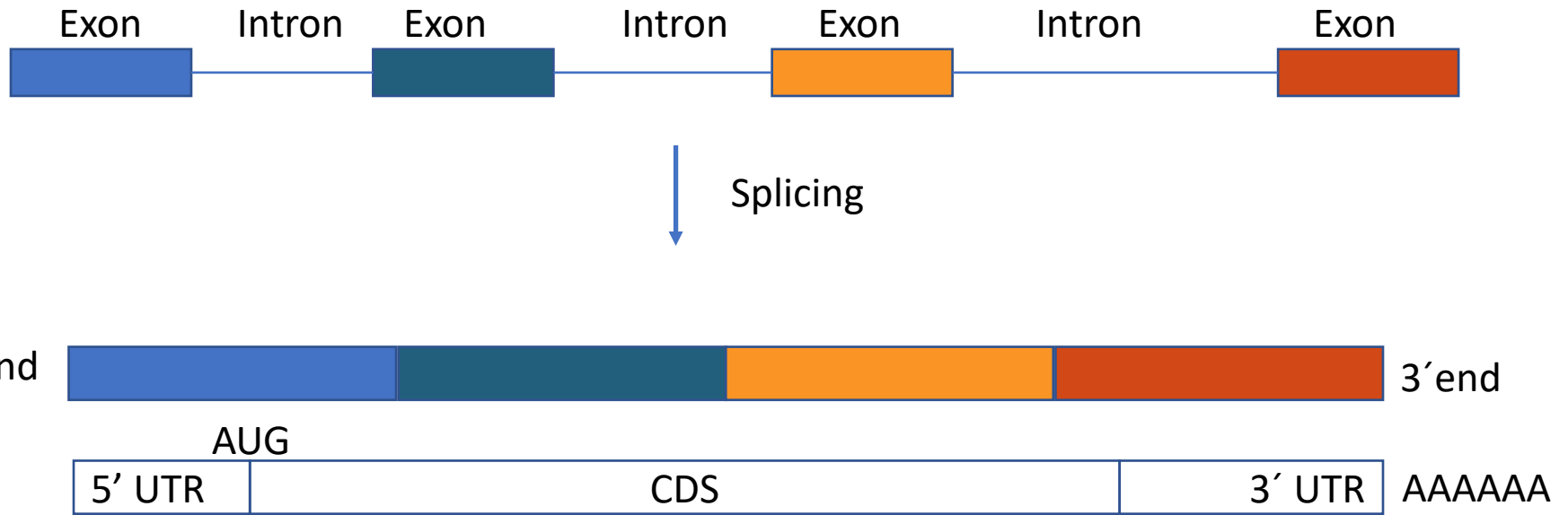
In contrast, the larger proportion of genome with reproducible but low biochemical signal strength and less evolutionary conservation is challenging to parse between specific functions and biological noise.

The complementary nature of evolutionary, biochemical, and genetic evidence.



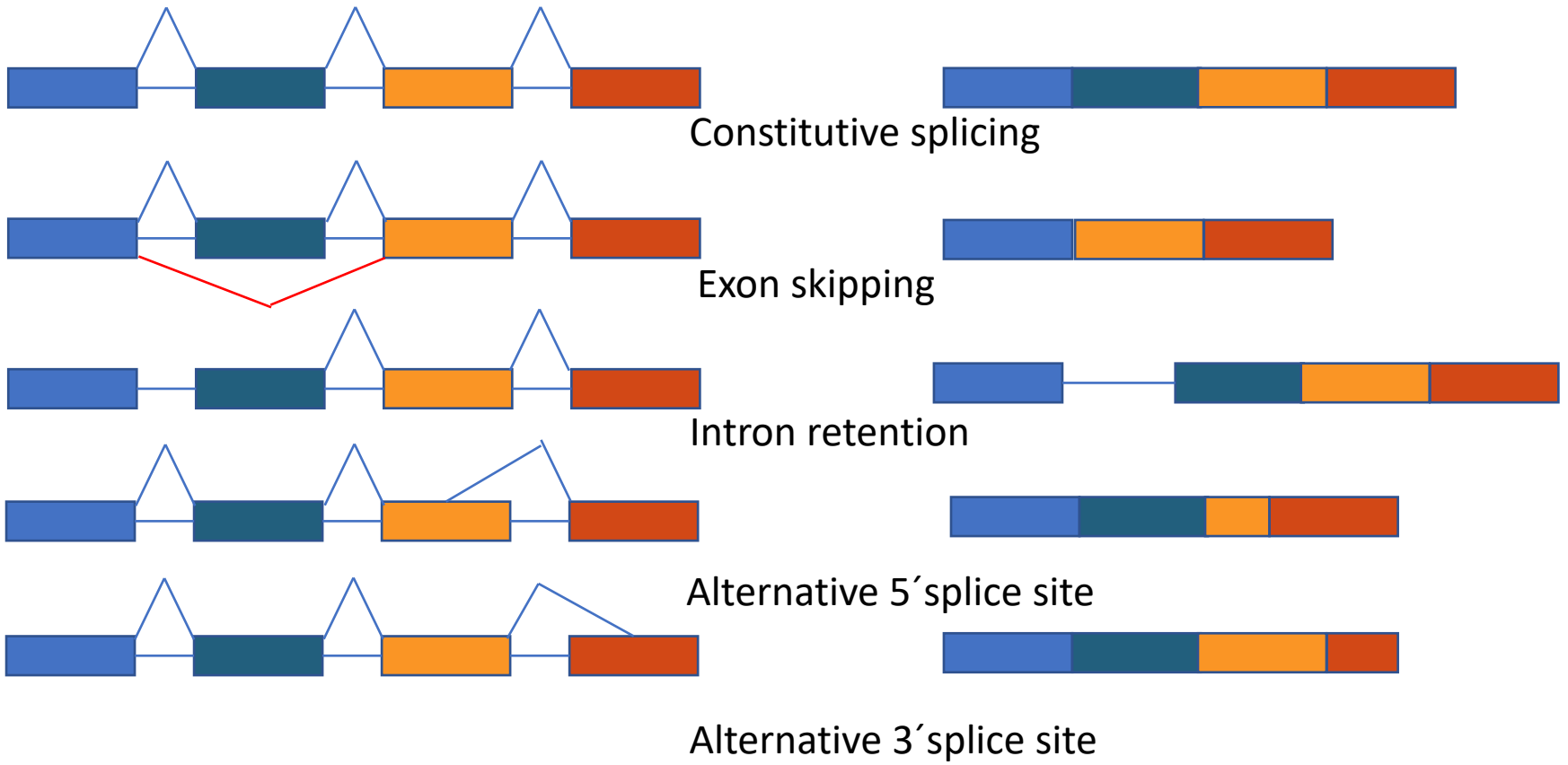


# RNA structure

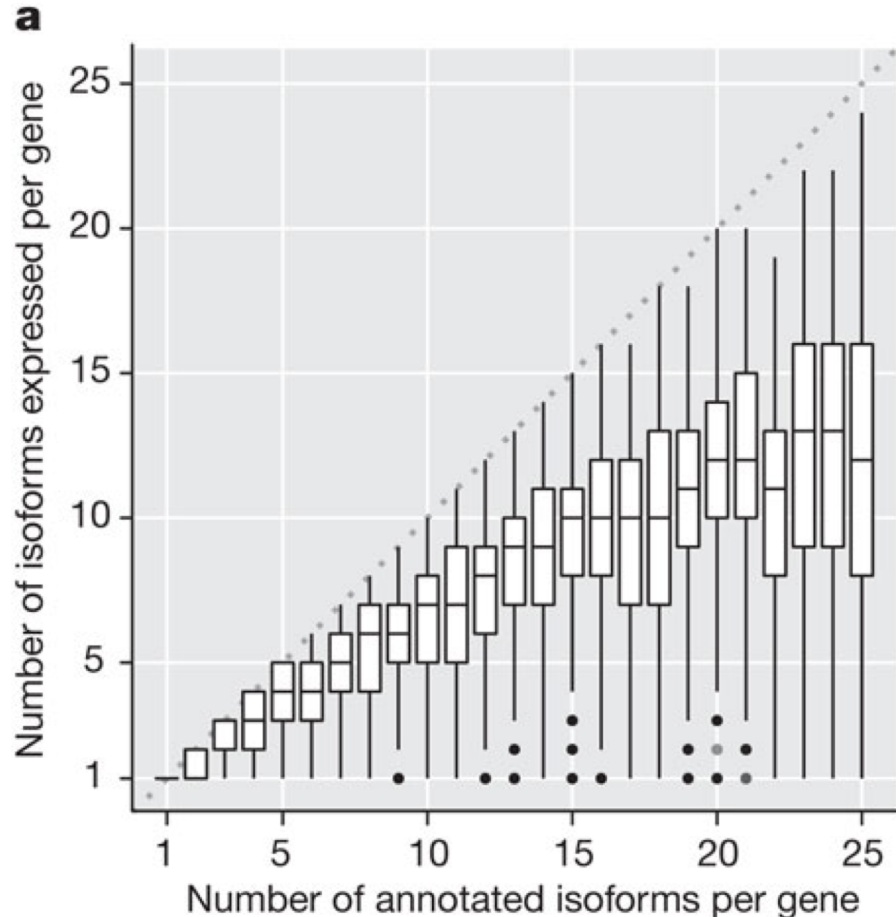


UTR = Untranslated region  
CDS = Coding sequence

# One gene can produce many different isoforms

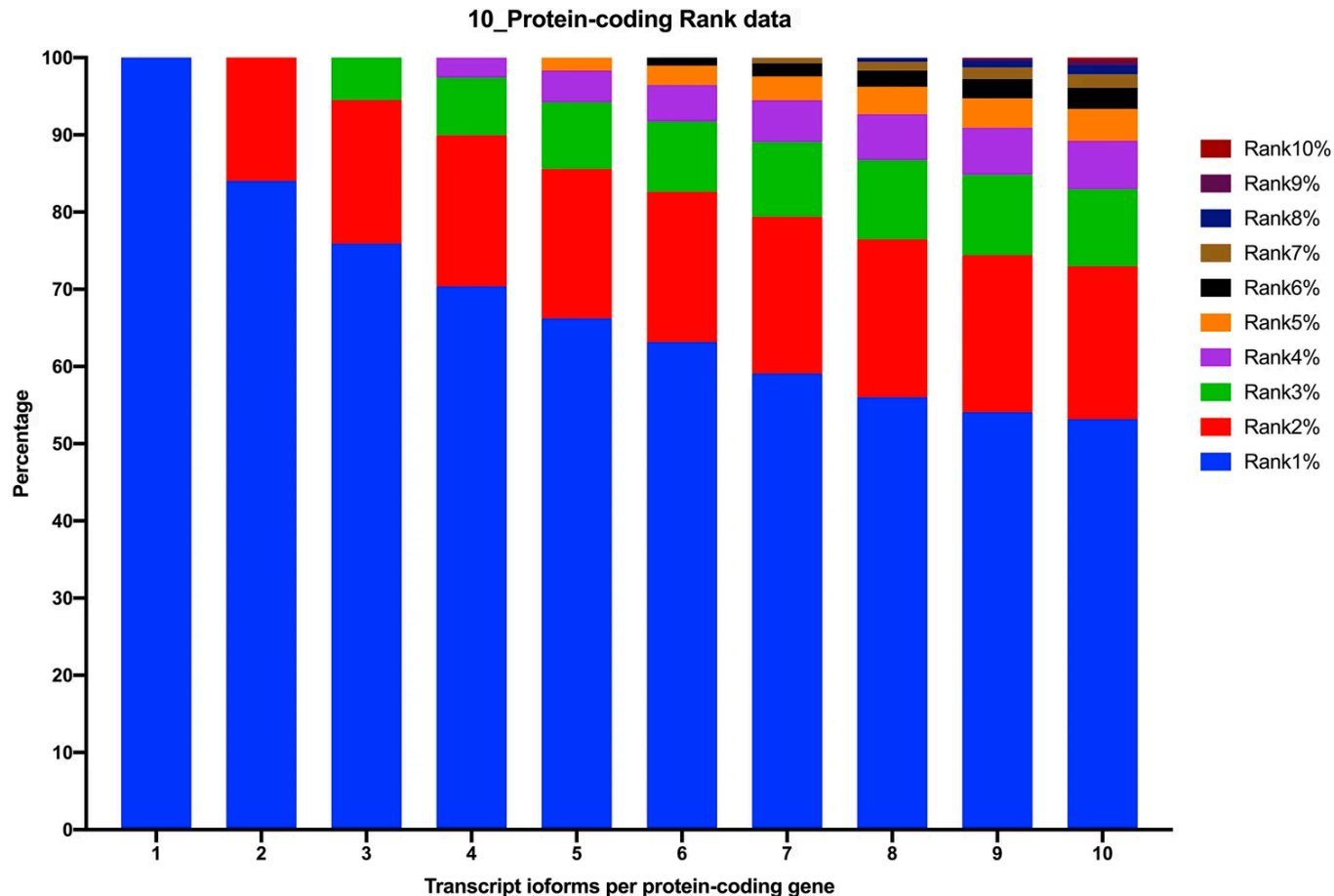


# Encode: most isoforms being transcribed



**Landscape of transcription in human cells, S Djebali *et al.* *Nature* 2012**

# Now: Only a few isoforms being transcribed at a high concentration



Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset, Tung *et al.* Scientific reports. 2020

# Summary

- One gene can produce many isoforms (transcripts)
  - Only a few of those isoforms are likely to be functional
  - Conservation in other species, Functional analysis, coding ability and genetic information can help in identifying which that are important.
- Just because a RNA is differentially expressed between two setting does NOT mean that they are important for the difference between the two settings.





**Thank you.**

---

**Johan Reimegård | 30-November-2020**