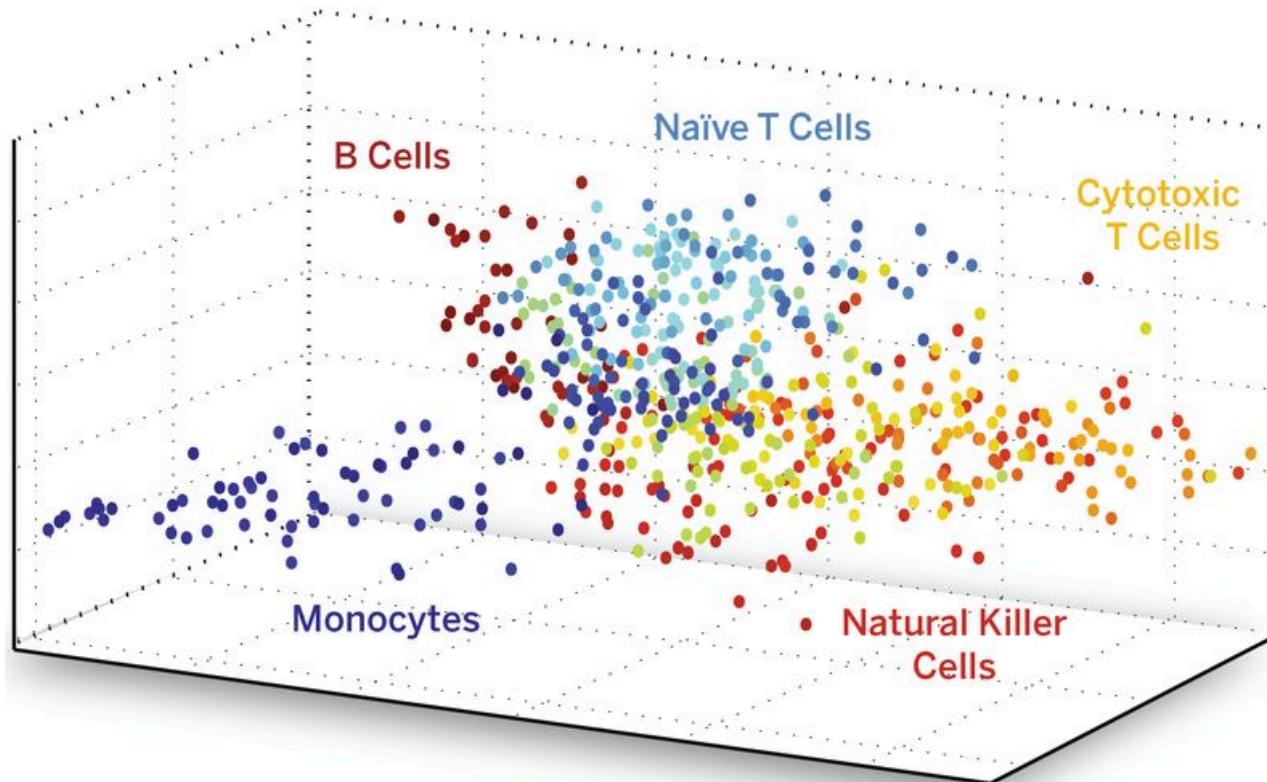


RNA introduction

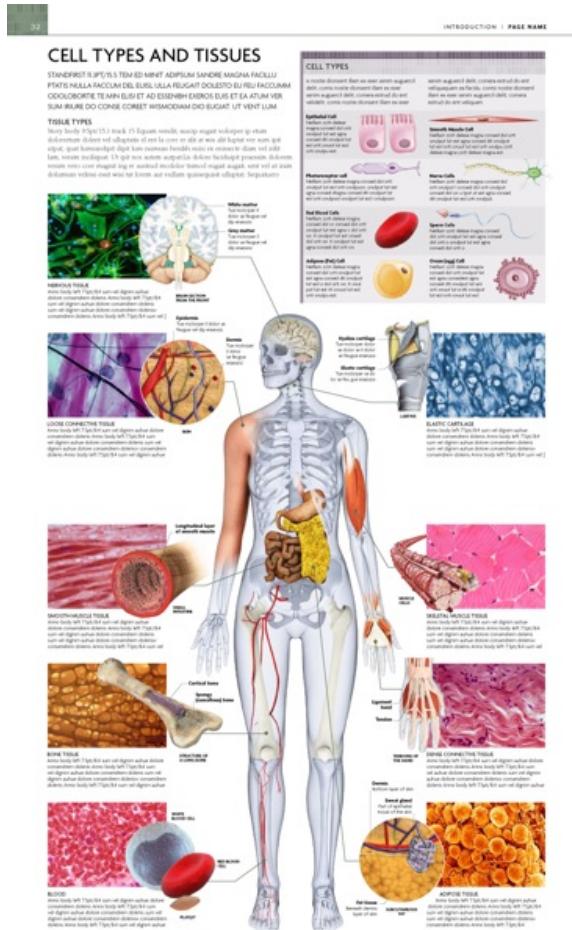
RNA-seq data analysis

Johan Reimegård | 13-May-2019

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



RNA varies between samples



-Tissues

-Cell types

-Cell states

-Individuals

-Cells

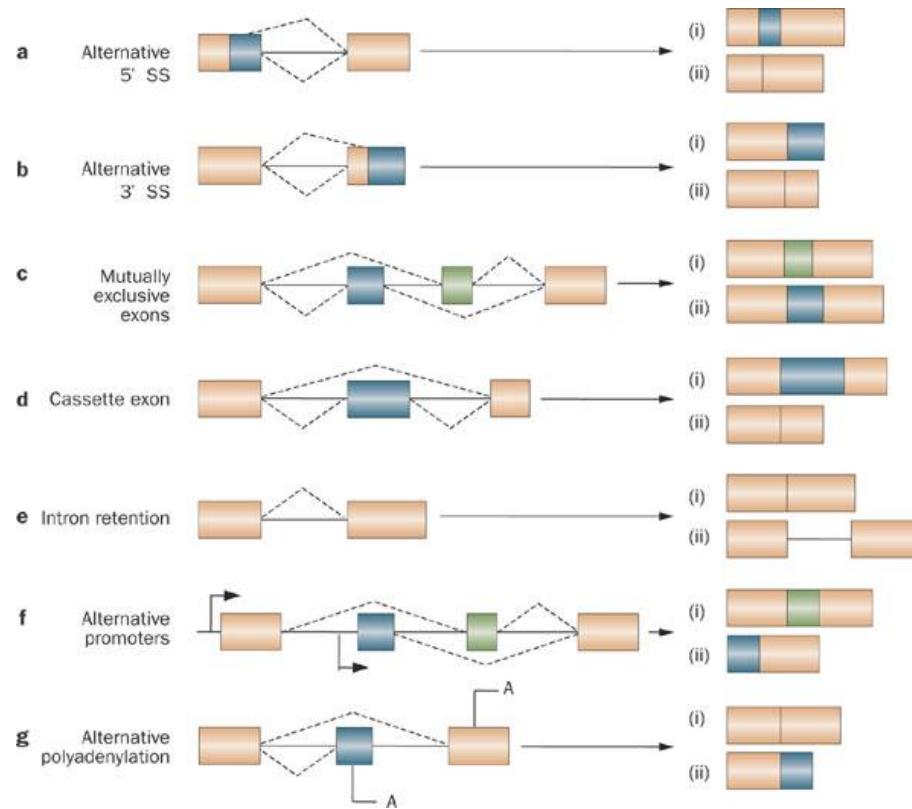
RNA gives information on which genes are expressed



How DNA get transcribed to RNA (and sometimes then translated to proteins) varies between e. g.

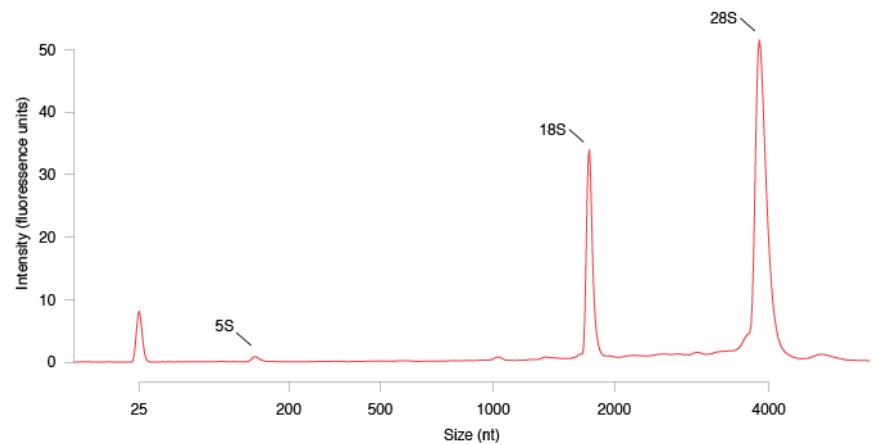
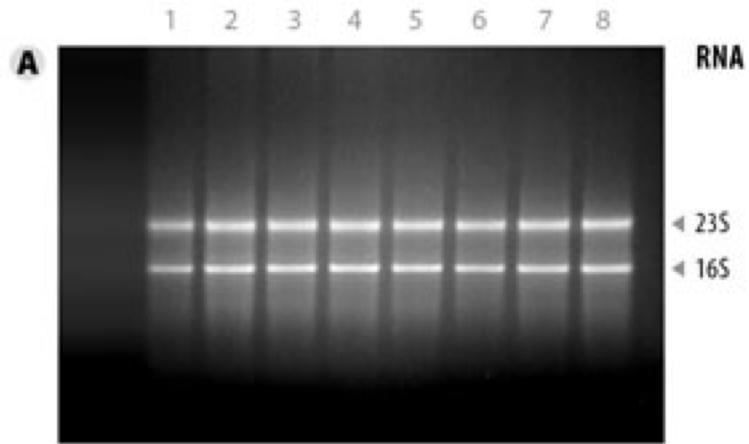
- Tissues
- Cell types
- Cell states
- Individuals

One gene many different isoforms

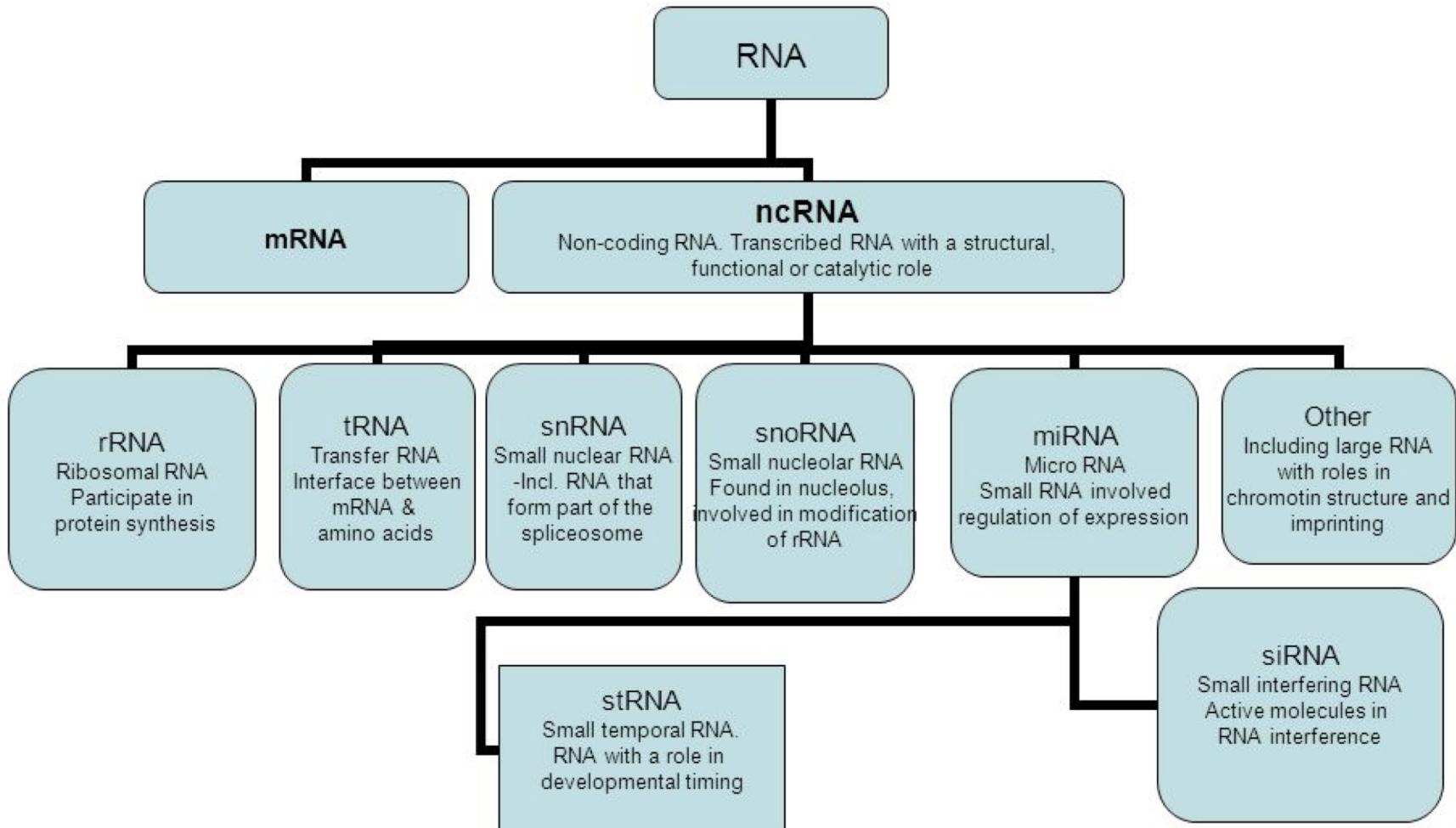


RNA flavors (pre sequencing era)

- House keeping RNAs
 - rRNAs, tRNAs, snoRNAs, snRNAs, SRP RNAs, catalytic RNAs (RNase E)
- Protein coding RNAs
 - (1 coding gene ~ 1 mRNA)
- Regulatory RNAs
 - Few rare examples



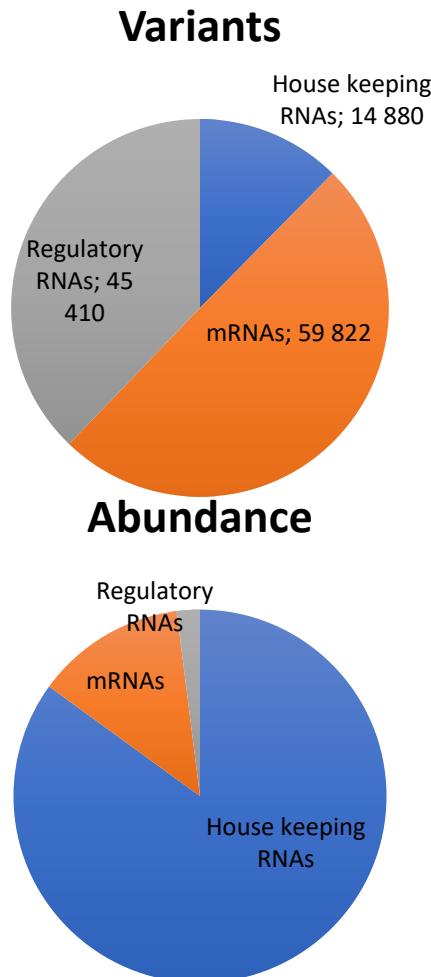
There is a wide variety of different functional RNAs



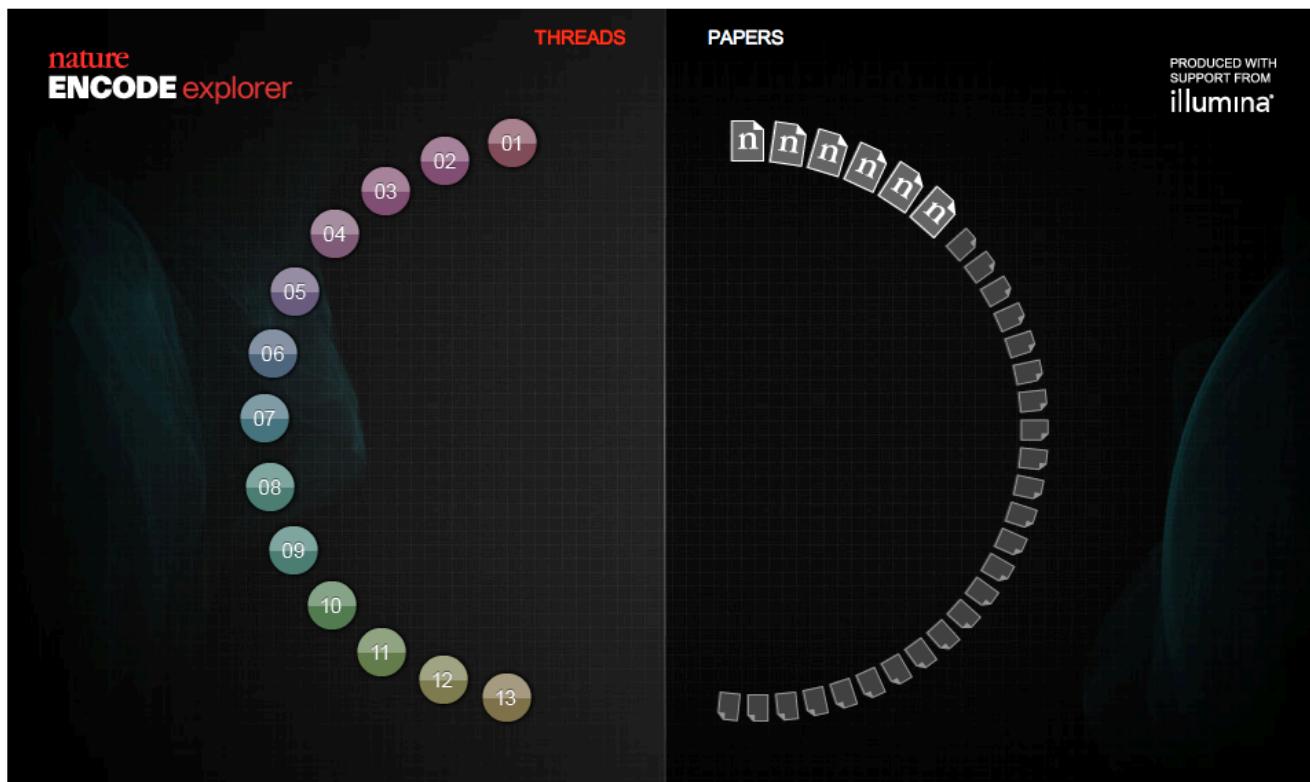
RNA flavors

- 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



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



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.

ENCylopedia 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

ENCylopedia 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

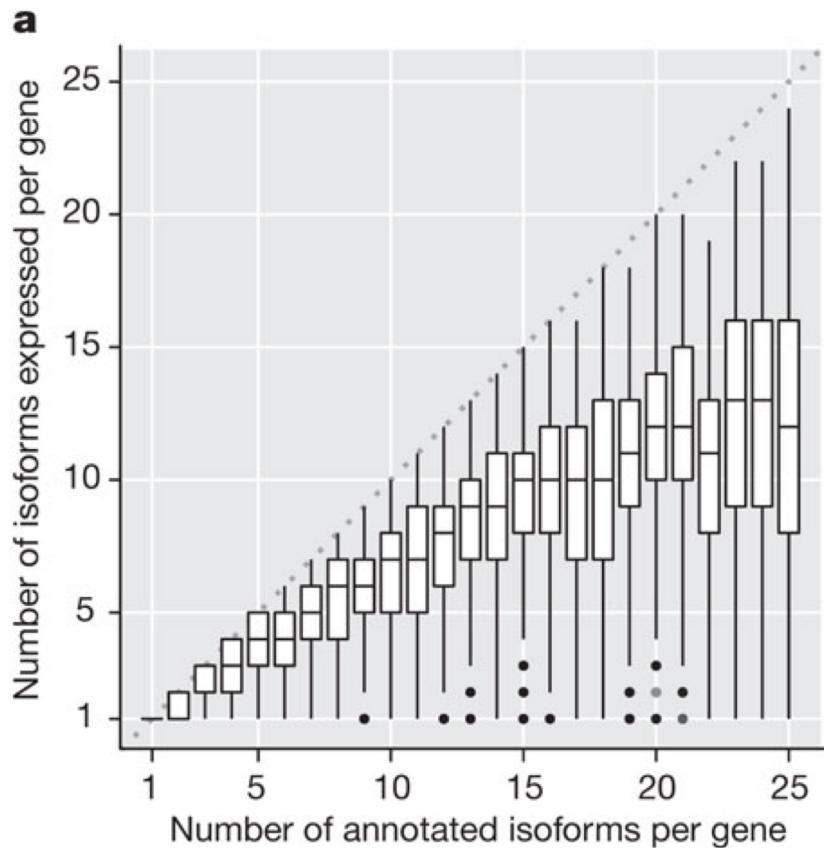
442 researchers

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

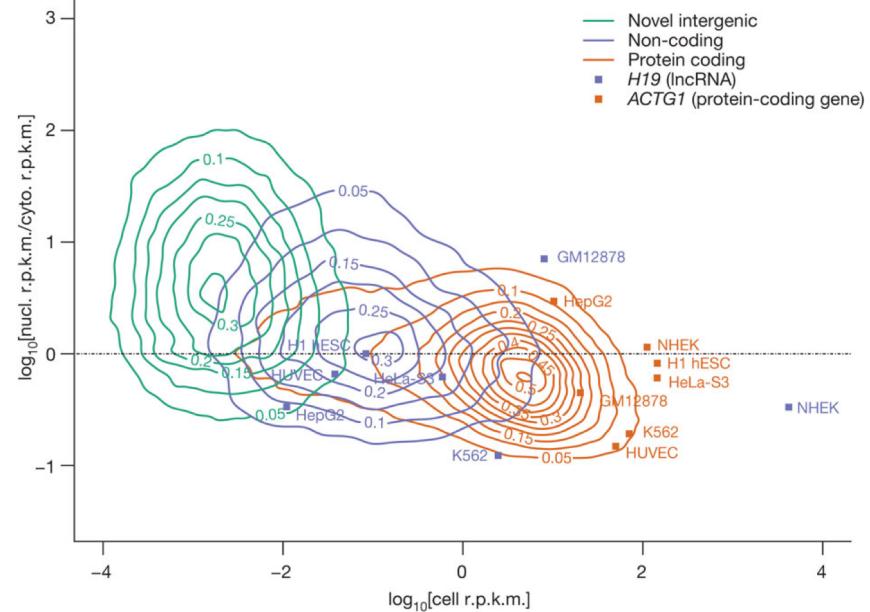
Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts, respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines.

RNA flavors

Variants



Abundance



RNA flavors

OPEN  ACCESS Freely available online

PLOS BIOLOGY

ance

Most “Dark Matter” Transcripts Are Associated With Known Genes

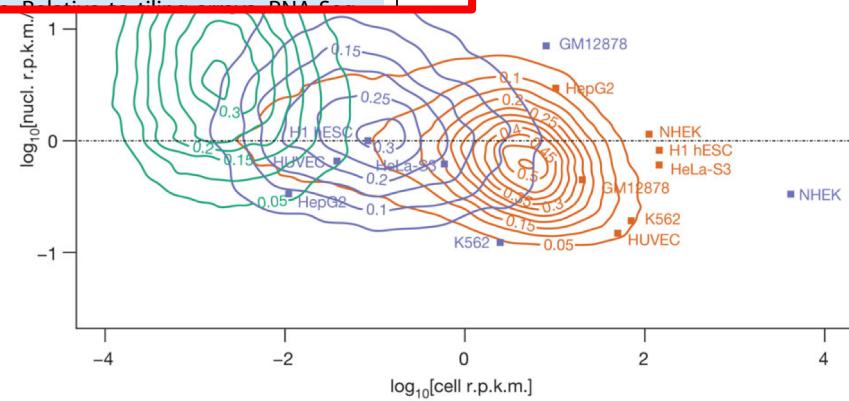
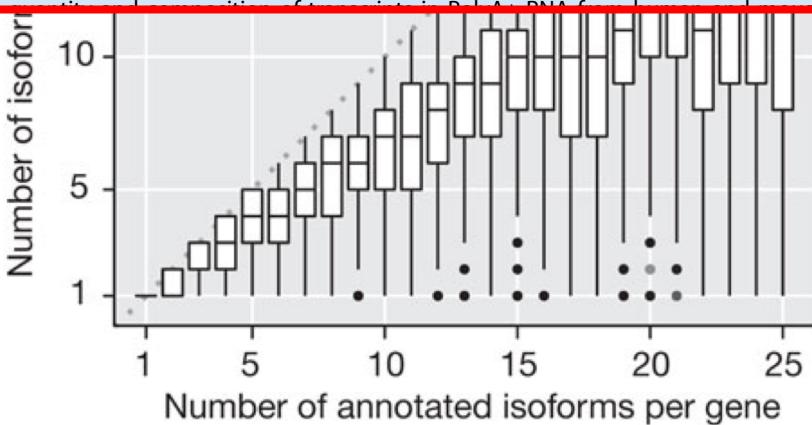
Harm van Bakel¹, Corey Nislow^{1,2}, Benjamin J. Blencowe^{1,2}, Timothy R. Hughes^{1,2*}

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

- Novel intergenic
- Non-coding
- Protein coding
- H19 (lncRNA)
- ACTG1 (protein-coding gene)

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

Balance

Most “Dark Matter” Transcripts Are Associated With

K

Ha

1 Ba
Ont

OPEN ACCESS Freely available online

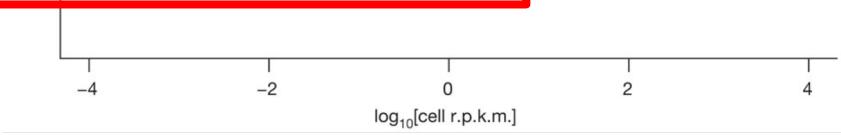
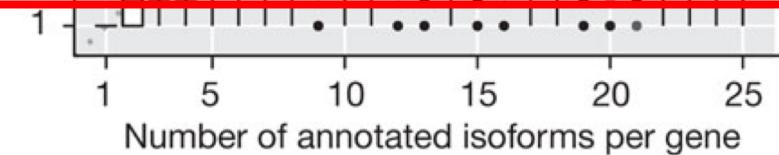
PLOS BIOLOGY

Perspective

The Reality of Pervasive Transcription

Michael B. Clark¹, Paulo P. Amaral^{1*}, Felix J. Schlesinger^{2*}, Marcel E. Dinger¹, Ryan J. Taft¹, John L. Rinn³, Chris P. Ponting⁴, Peter F. Stadler⁵, Kevin V. Morris⁶, Antonin Morillon⁷, Joel S. Rozowsky⁸, Mark B. Gerstein⁸, Claes Wahlestedt⁹, Yoshihide Hayashizaki¹⁰, Piero Carninci¹⁰, Thomas R. Gingeras^{2*}, John S. Mattick^{1*}

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 Trouillet 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



- Novel intergenic
 - Non-coding
 - Protein coding
 - $H19$ (lncRNA)
 - $ACTG1$ (protein-coding gene)

RNA flavors

OPEN  ACCESS Freely available online

Most “Dark Matter” Transcripts Are Associated With

K

H

1 Ba
Ont

Ha
The
Micha
Rinn³
Mark
John

OPEN  ACCESS Freely available online

Perspective

The Reality of Pervasive Transcription

OPEN  ACCESS Freely available online

Perspective

Response to “The Reality of Pervasive Transcription”

Harm van Bakel¹, Corey Nislow^{1,2}, Benjamin J. Blencowe^{1,2}, Timothy R. Hughes^{1,2*}

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

“stably transcribed” 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

Number of annotated isoforms per gene

$\log_{10}[\text{cell r.p.k.m.}]$

1

1

5

10

15

20

25

-4

-2

0

2

4

Novel intergenic
lncRNA
(protein-coding gene)

K
hESC
La-S3

NHEK

RNA flavors

OPEN  ACCESS Freely available online

PLOS BIOLOGY

Most “Dark Matter” Transcripts Are Associated With

K

OPEN  ACCESS Freely available online

PLOS BIOLOGY

H

1 Ba
Ont

Perspective

The Reality of Pervasive Transcription

OPEN  ACCESS Freely available online

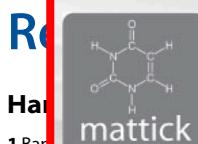
ance

PLOS BIOLOGY

Michael
Rinn³
Mark
John

1 Institute
Cold Spring
Department
Germany
Pavillon
Miami, M

Perspective



Har

1 Bar

Canad

navigation

C

our

our

tran

ed.

1

1

Num

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

Comments by Mike Clark 

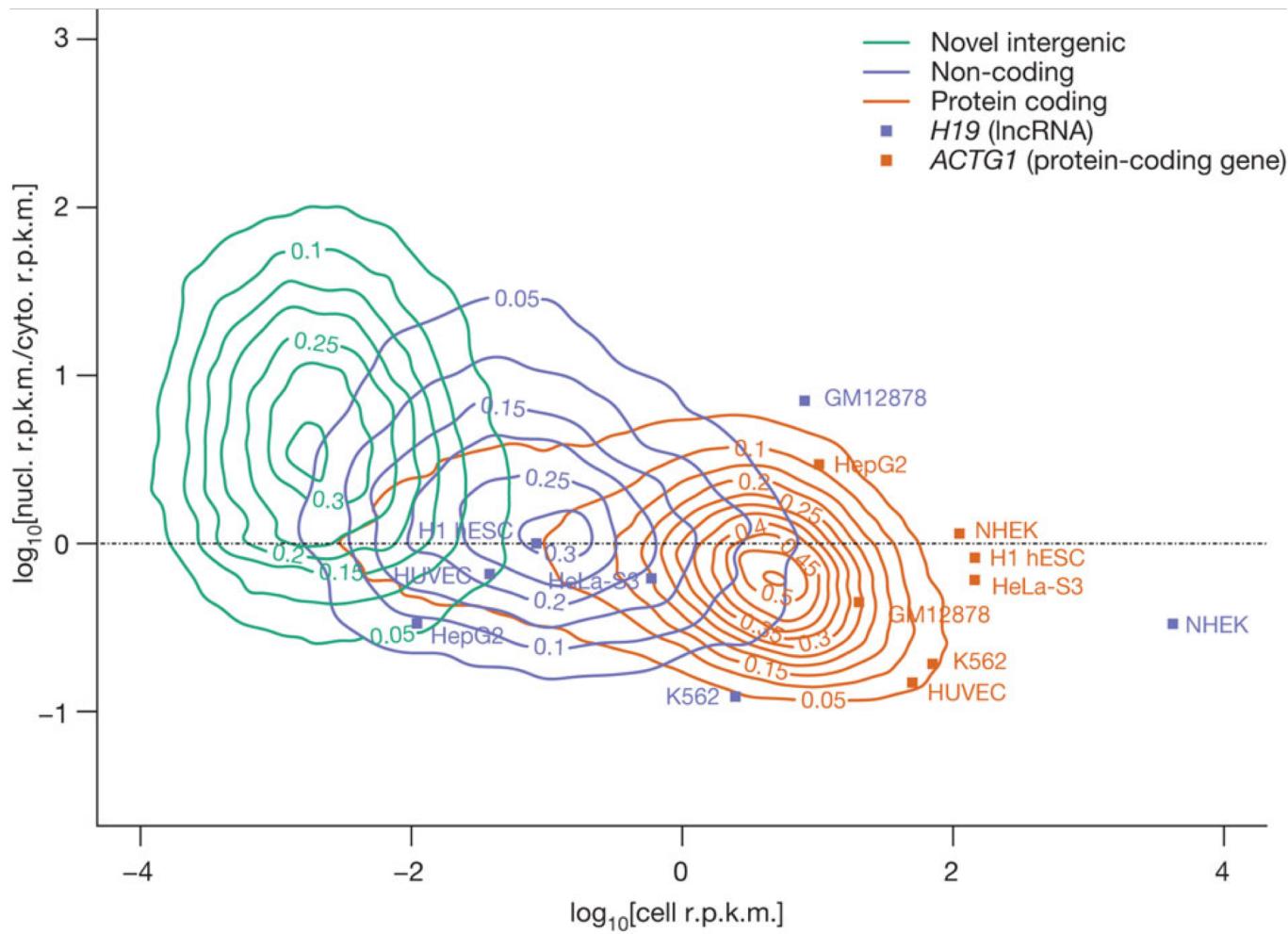
Van Bakel et al. 2011  (vB 11) have published their reply to our critique  of their paper van Bakel et al. 2010  (vB 10).

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

1. 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 previously thought. Previous evidence was unreliable due to false positives.
2. 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.
3. 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 positives.
4. 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 for complex samples and then 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 differentiate properly between this and genuine transcripts under their detection threshold.
5. vB 10 claimed that low level intergenic transcription may be due to “random initiation events” and/or transcriptional “byproducts” (ie: transcription noise), when the limitations of their sequencing and assembly methods made it impossible to distinguish between these and genuine transcripts.

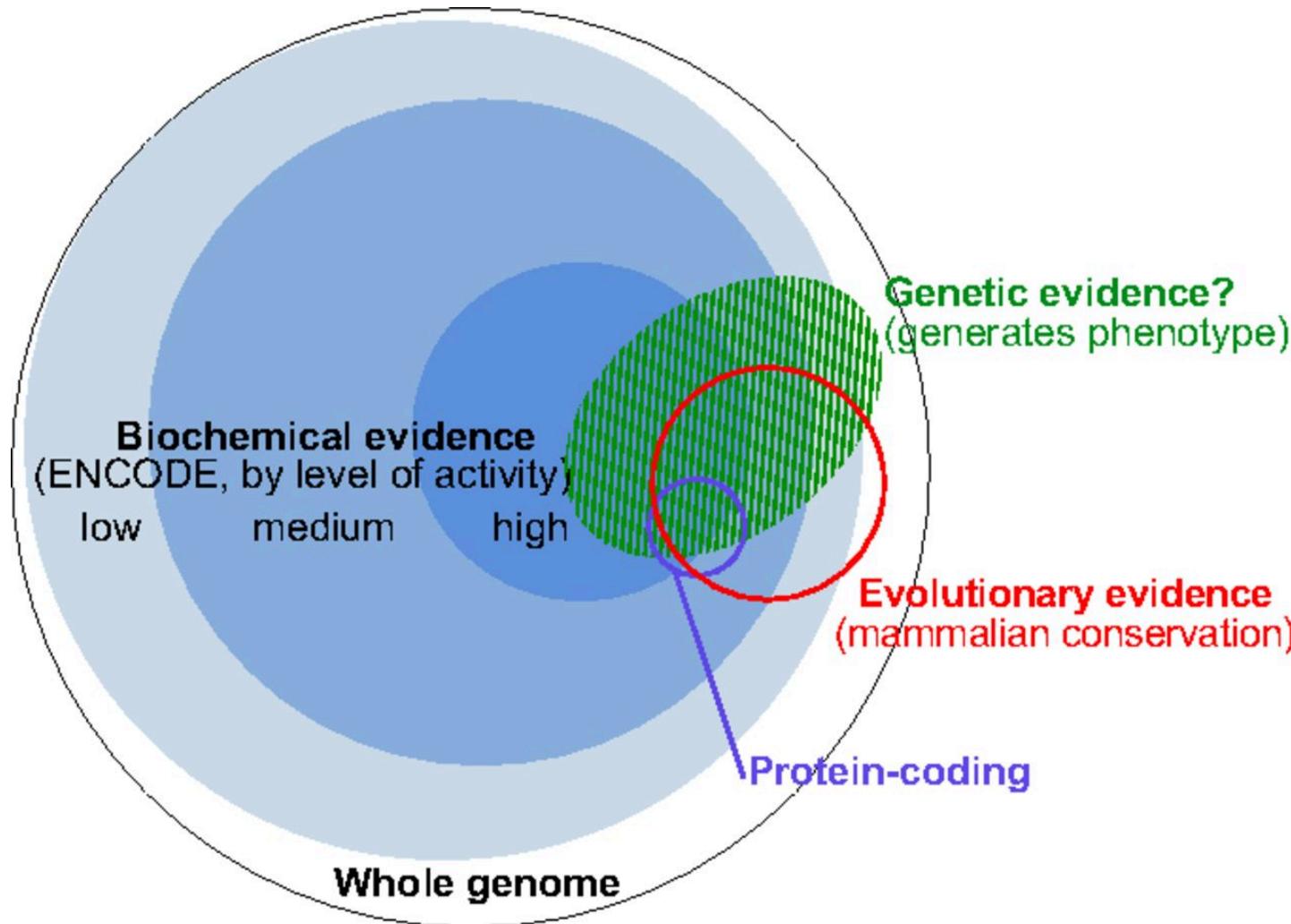
Novel intergenic
transcription
(lncRNA)
(protein-coding gene)

RNA flavors



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

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



Defining functional DNA elements in the human genome
Kellis M et al. PNAS 2014;111:6131-6138

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 background of the slide features a complex, abstract network graph. It consists of numerous small, dark brown dots representing nodes, connected by a dense web of thin, translucent blue lines representing edges. The graph is highly interconnected, with many cycles and dead ends, creating a sense of organic complexity.

Thank you. Questions?

Johan Reimegård | 13-May-2019