Clinical and Research Genomics

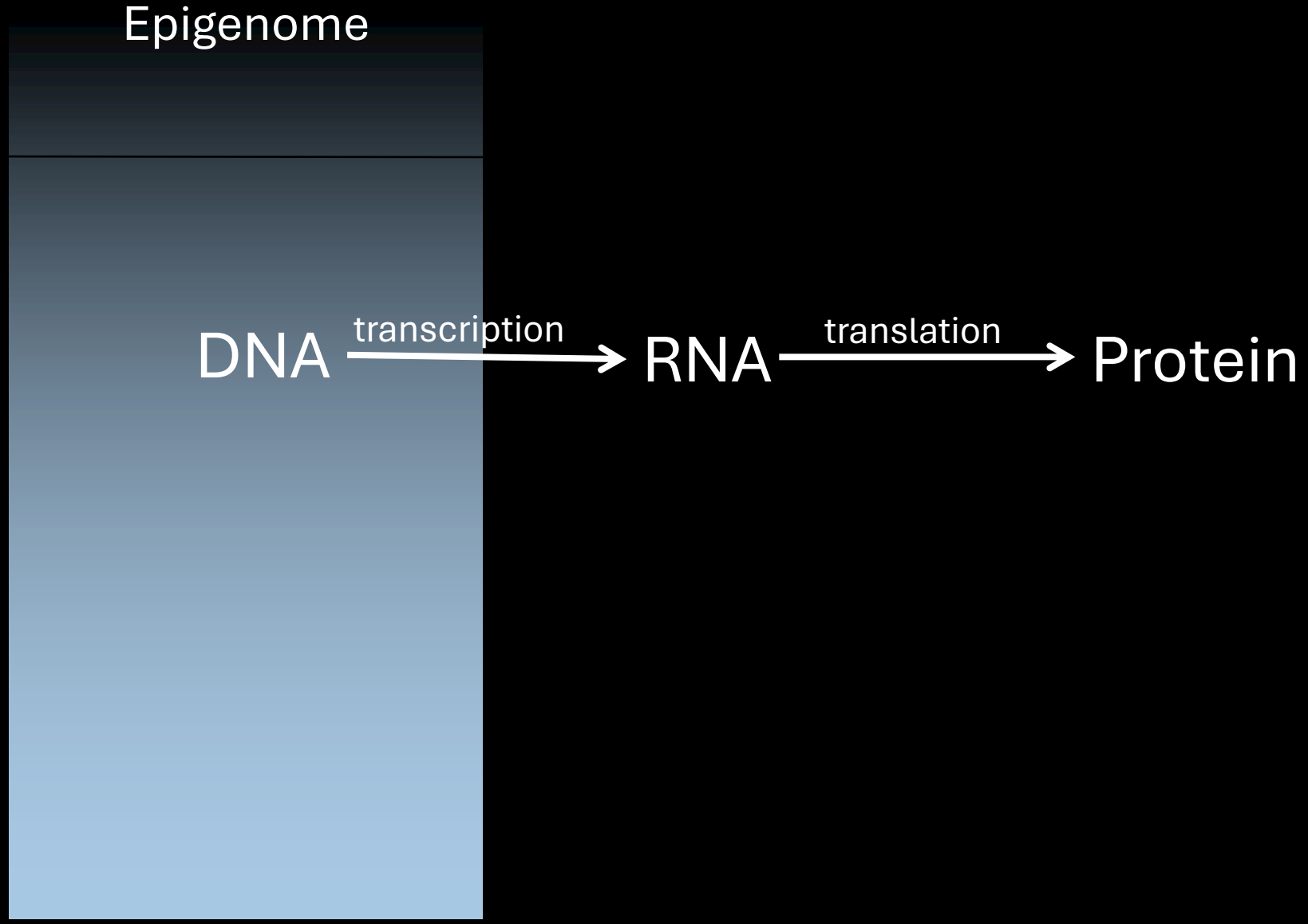
RNA-Sequencing, Epitranscriptomes, and Single Cells

Dr. Christopher E. Mason

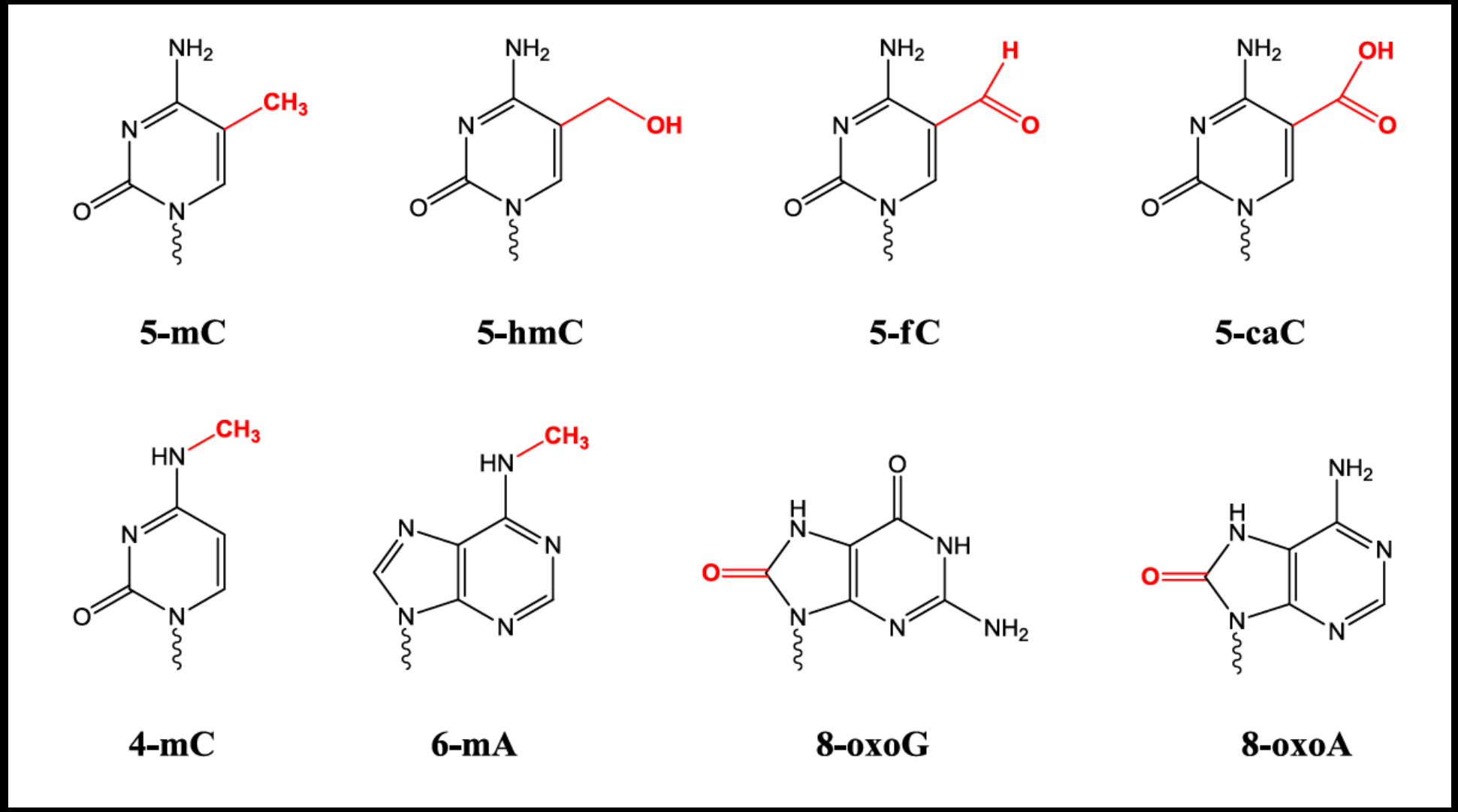
-

April 30, 2024

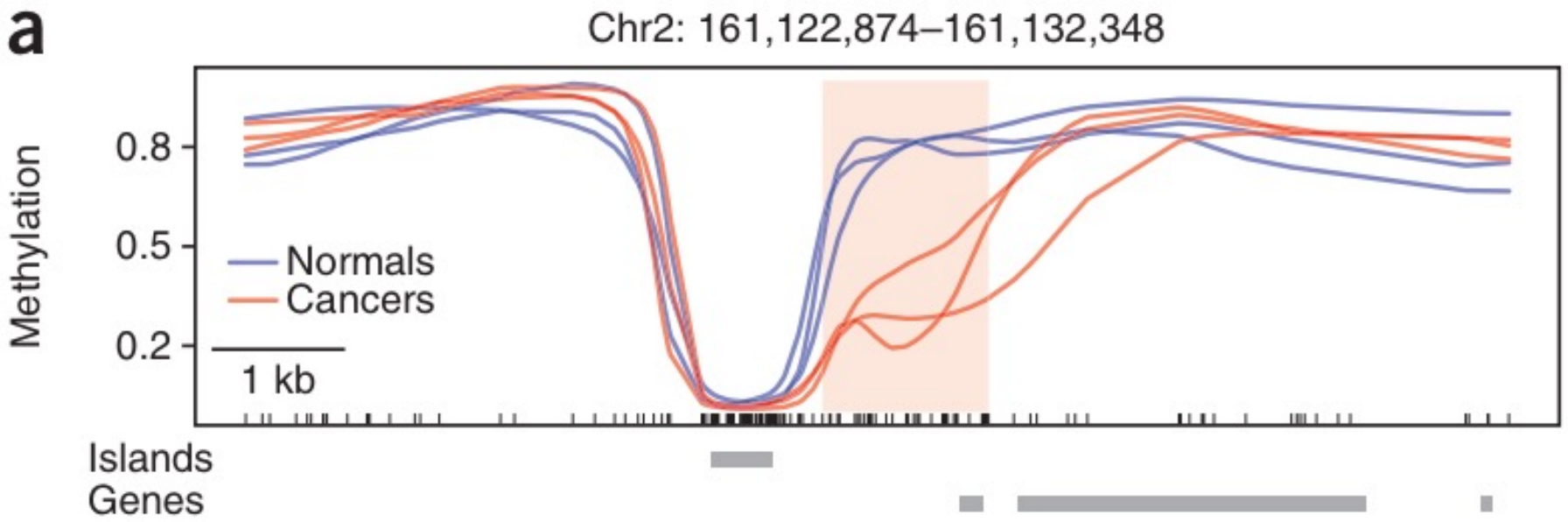
James Watson, 1958


Epigenome
DNA
transcription
RNA
translation
Protein

Small chemical tweaks for regulatory control

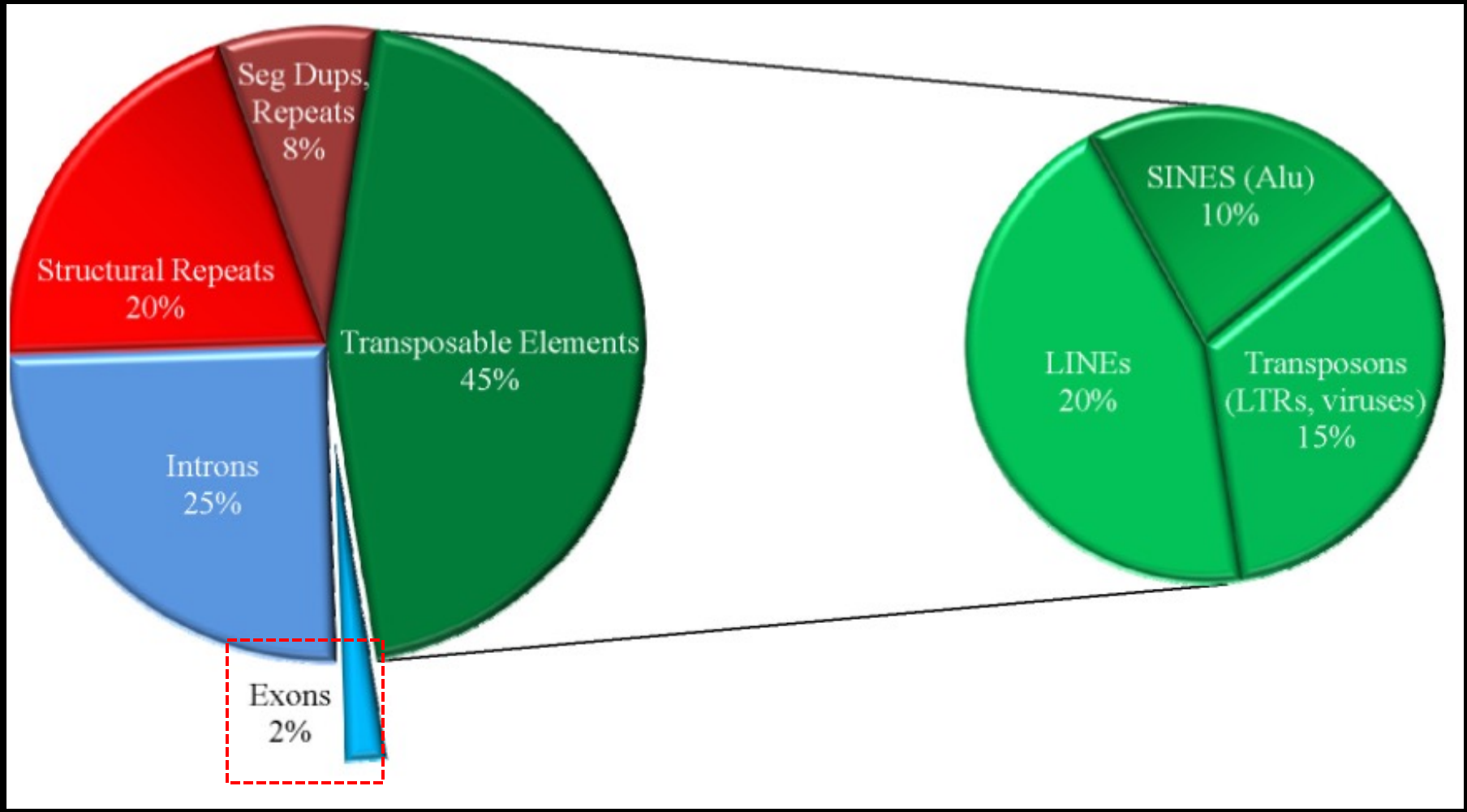


Methylation marks the spot; gene boundaries



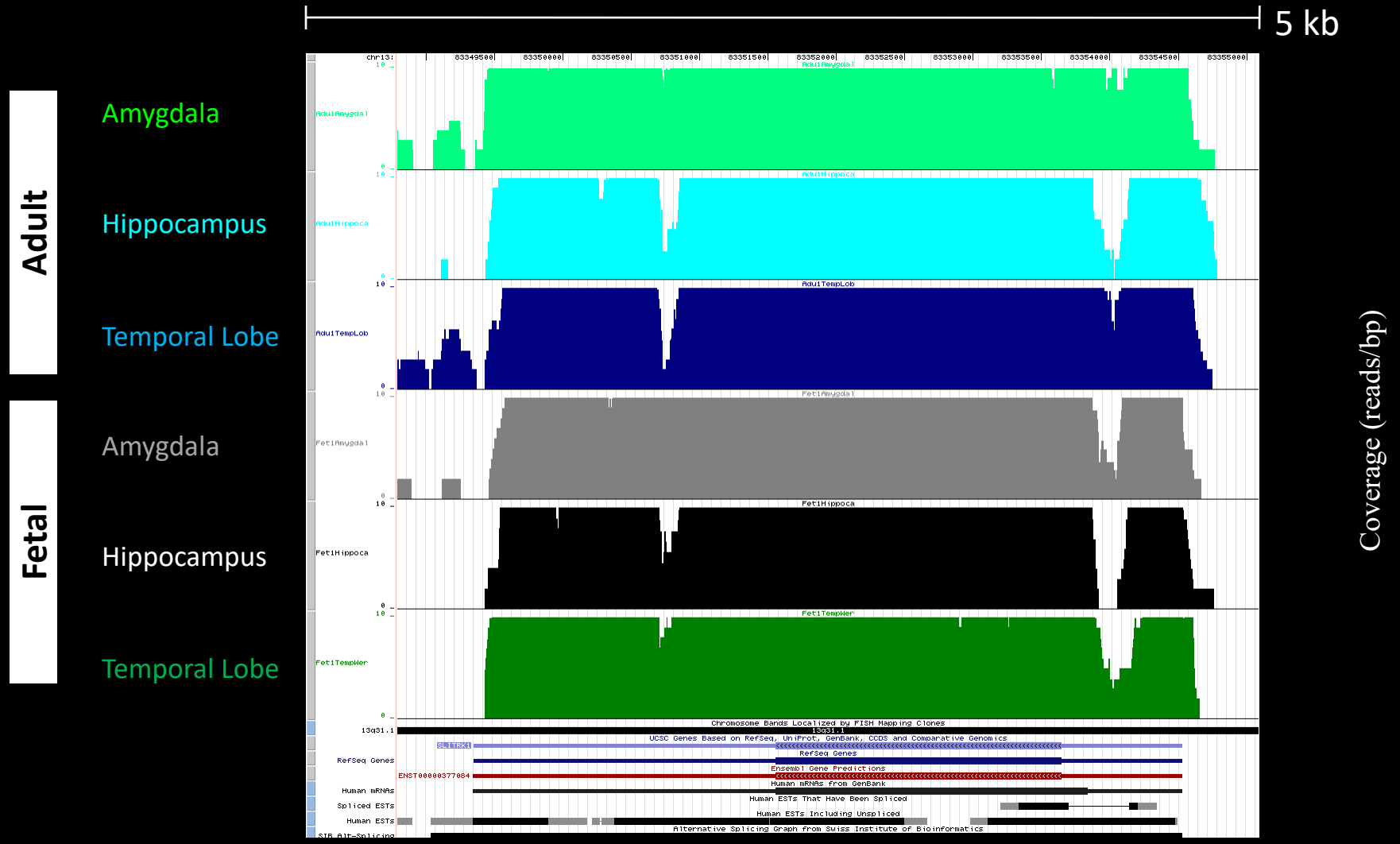
Nat Genet. 2011 Jun 26;43(8):768-75.

But how many genes are there?



from Mason, State, and Moldin, Kaplan & Sadock's Comprehensive Textbook of Psychiatry, 2009

Validation of known Gene Boundaries



Find Longer Isoforms

63 kb

Adult

Amygdala

Hippocampus

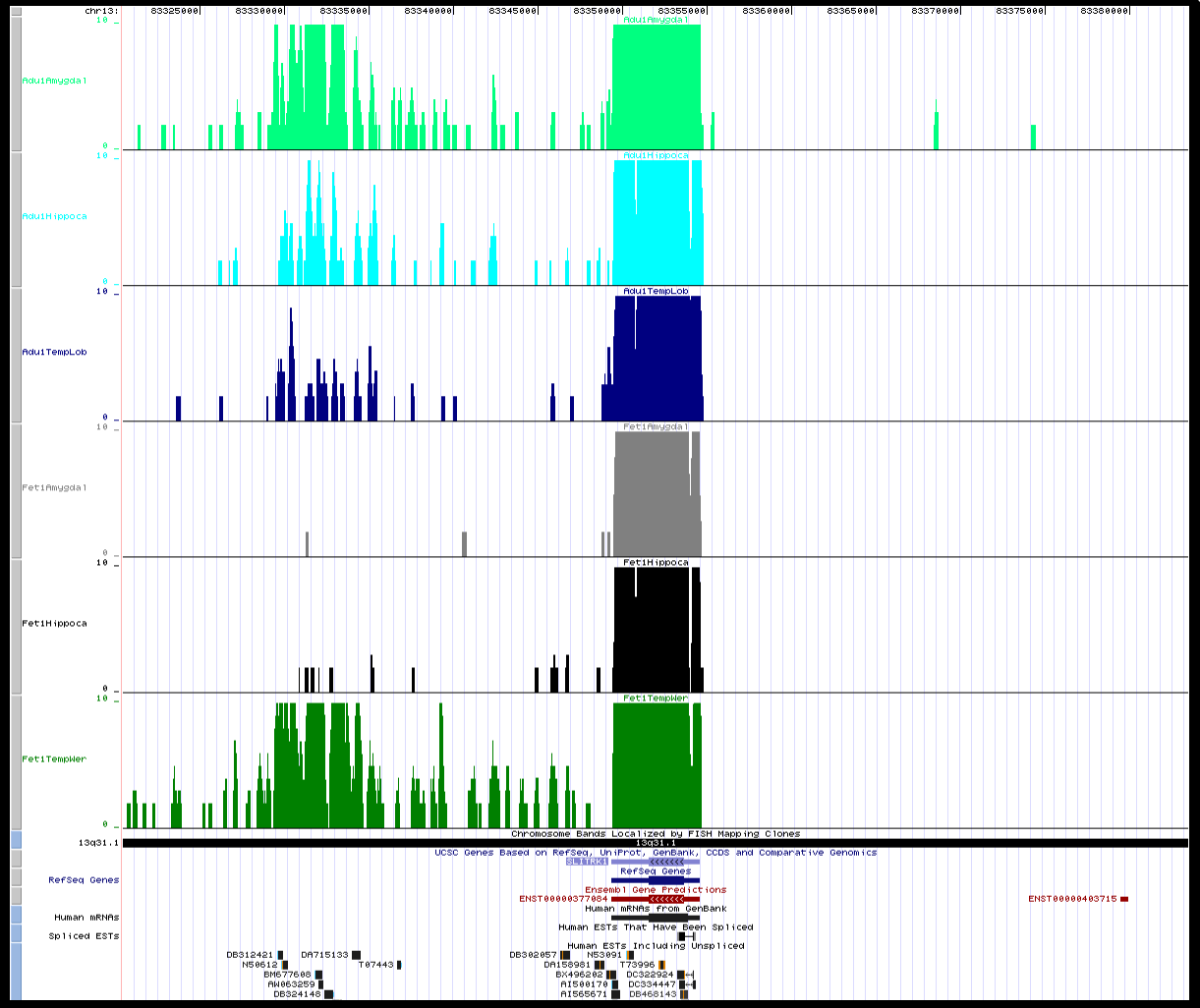
Temporal Lobe

Fetal

Amygdala

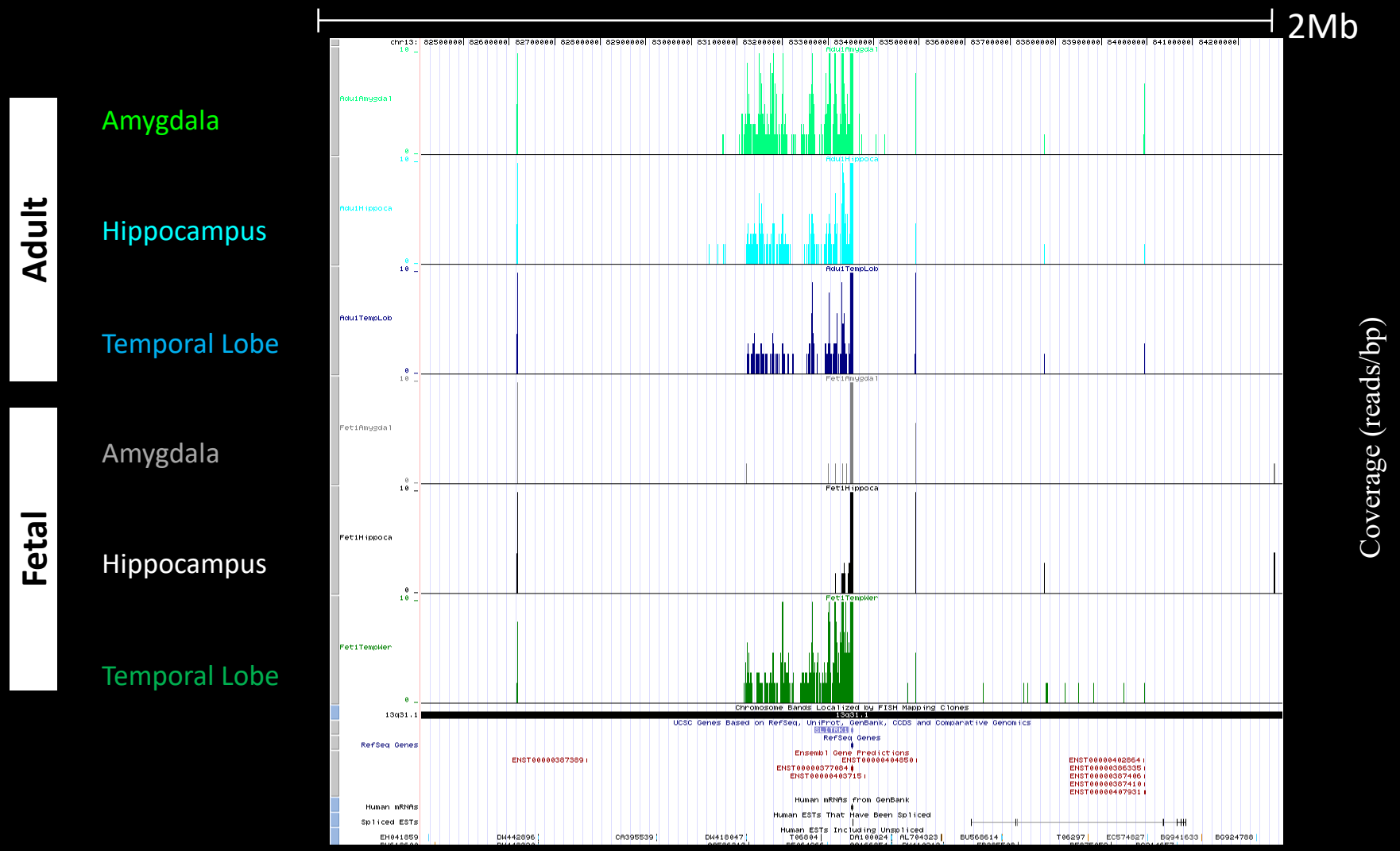
Hippocampus

Temporal Lobe

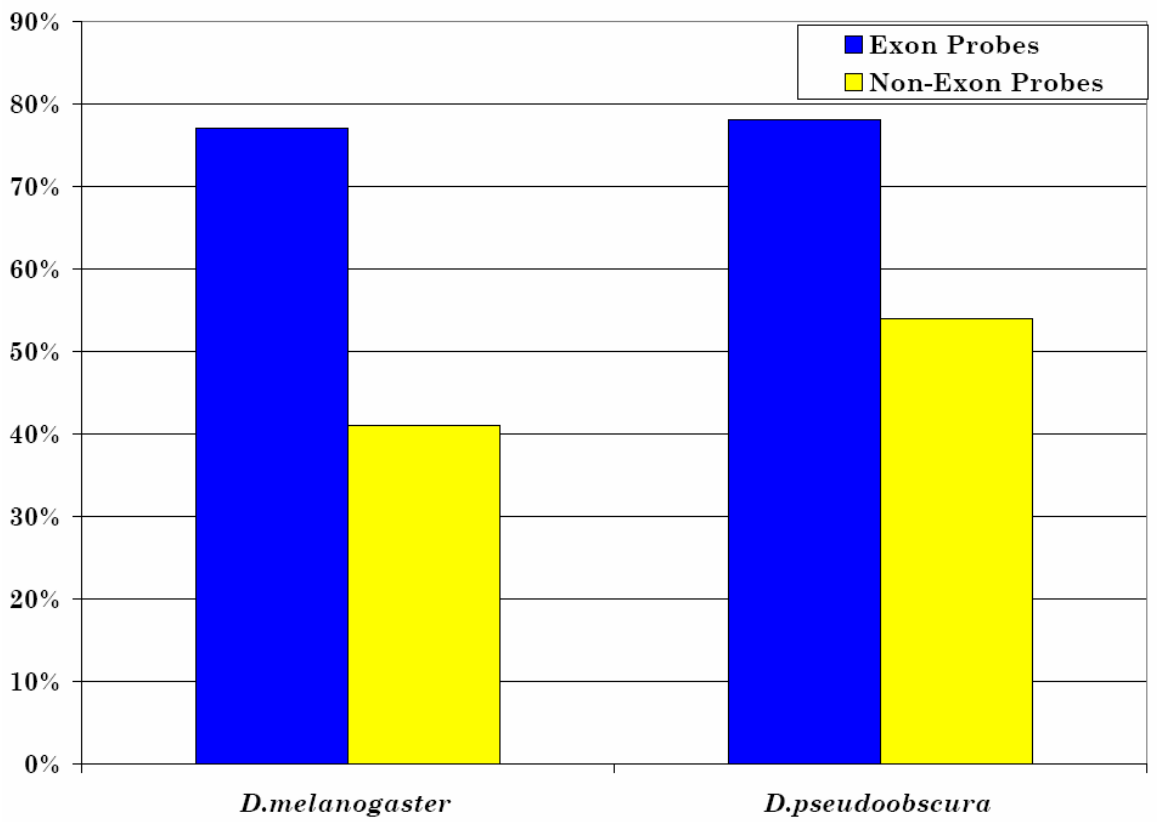


Coverage (reads/bp)

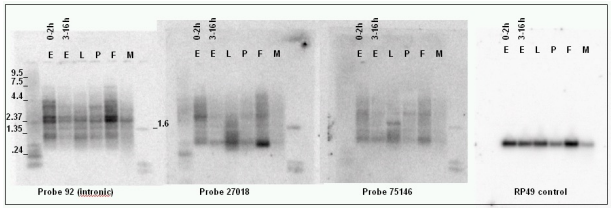
Find New Genes



About Half of the Noncoding Genome is Transcriptionally Active



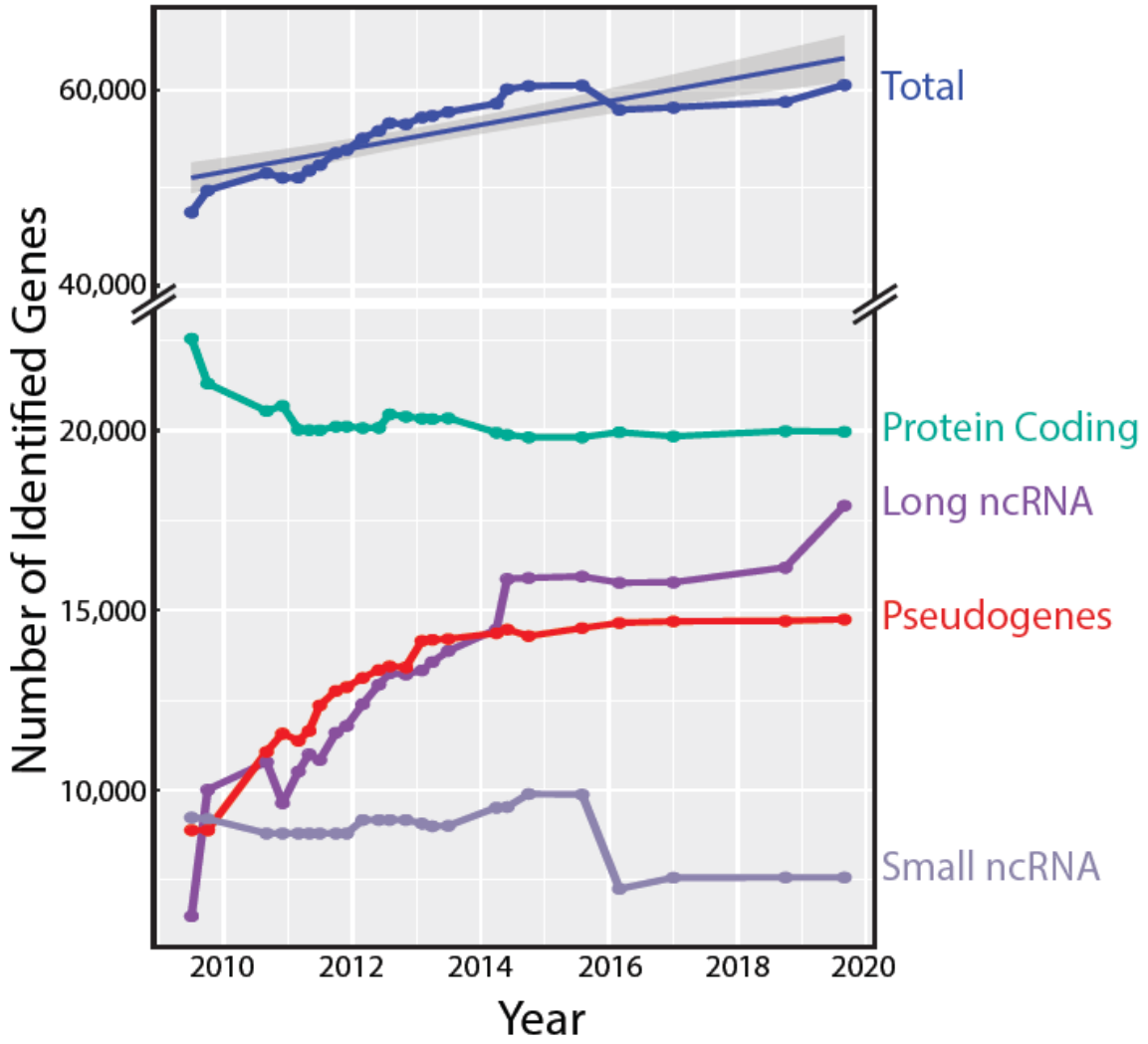
Mason, 2006



Stolc, Gauhar, Mason et al, *Science*, 2004

Humans: 47% (Schadt *et al.*, 2004)
Arabadopsis: 51% (Yamada *et al.*, 2003)

New human genes are still being found



The Next 500 Years
https://www.gencodegenes.org/

Human

Statistics about the current GENCODE Release (version 45)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

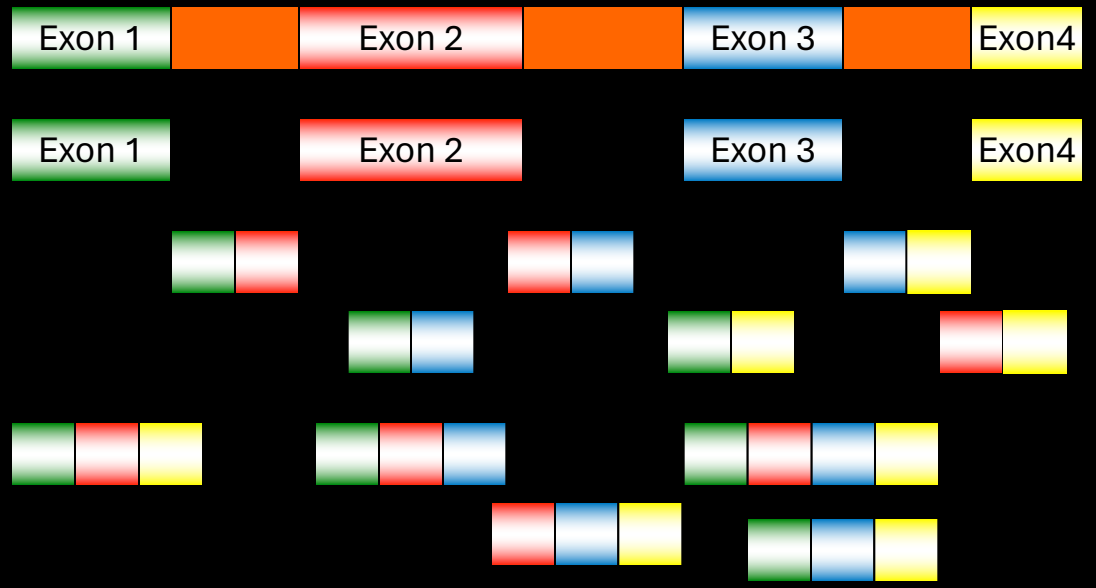
General stats

Total No of Genes	63187	Total No of Transcripts	252930
Protein-coding genes	19395	Protein-coding transcripts	89110
- readthrough genes (not included)	654	- full length protein-coding	64028
Long non-coding RNA genes	20424	- partial length protein-coding	25082
Small non-coding RNA genes	7565	Nonsense mediated decay transcripts	21427
Pseudogenes	14719	Long non-coding RNA loci transcripts	59719
- processed pseudogenes	10658		
- unprocessed pseudogenes	3566		
- unitary pseudogenes	258		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	65357
- protein coding segments	411	Genes that have more than one distinct translations	13600
- pseudogenes	237		

https://www.gencodegenes.org/human/stats.html

The transcriptome's potential complexity is vast

Exons	Variants	Junctions
1	1	0
2	3	1
3	7	3
4	15	6
5	31	10
6	63	15
7	127	21
8	255	28
	2^n-1	$\frac{n(n-1)}{2}$



exon1
exon2
exon3
exon1-exon2
exon1-exon3
exon2-exon3
exon1-exon2-exon3

exon4
exon1- exon4
exon2-exon4
exon3-exon4
exon1-exon2-exon4
exon1-exon3-exon4
exon2-exon3-exon4
exon1-exon2-exon3-exon4

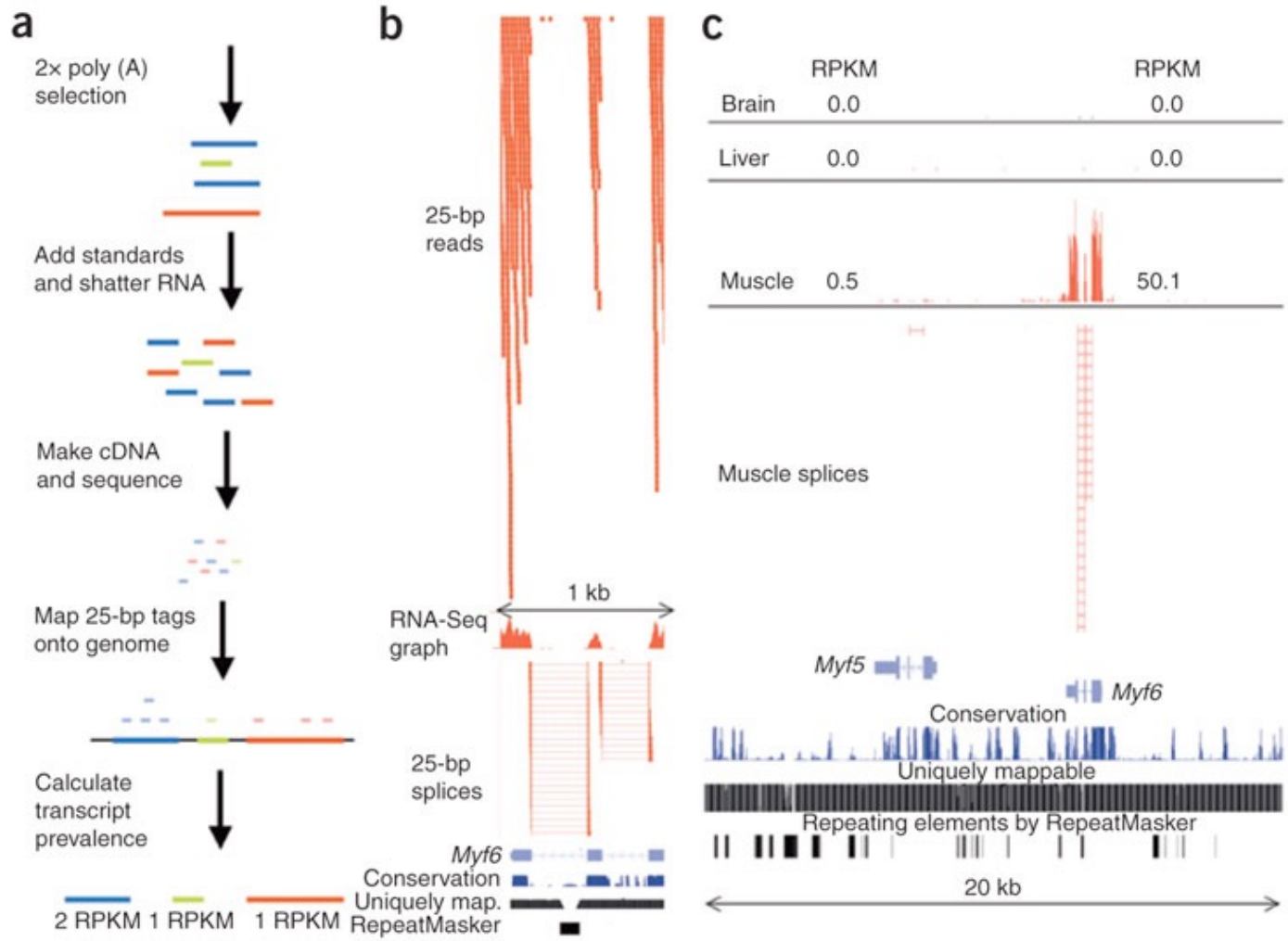
$8x10^{83}$ theoretical transcript combinations
$8x10^{80}$ atoms in the universe
(1^{59} atoms/star, 1^{11} stars/galaxy, 1^{10} galaxies)

Li and Mason, ARGHG, 2014

Need a way to sequence them all

Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]



2008

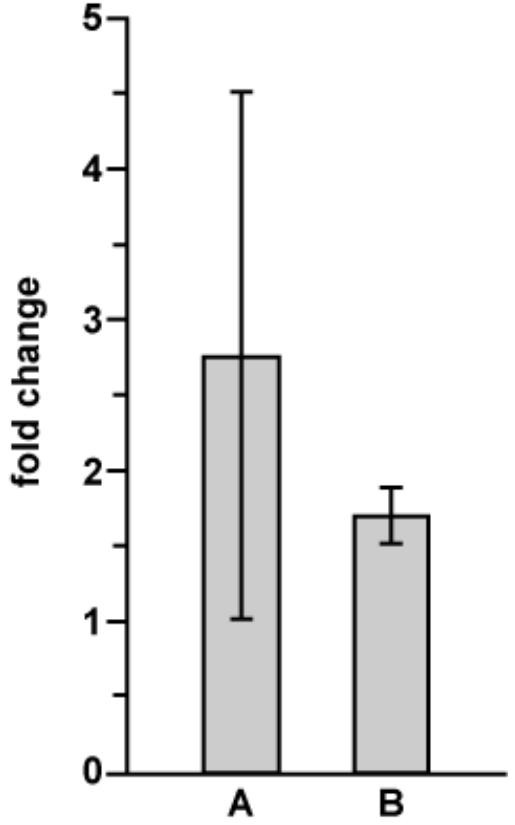
Can RNA-Seq replace microarrays?

RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays

Marioni and Mason *et. al*, Genome Research, 2008

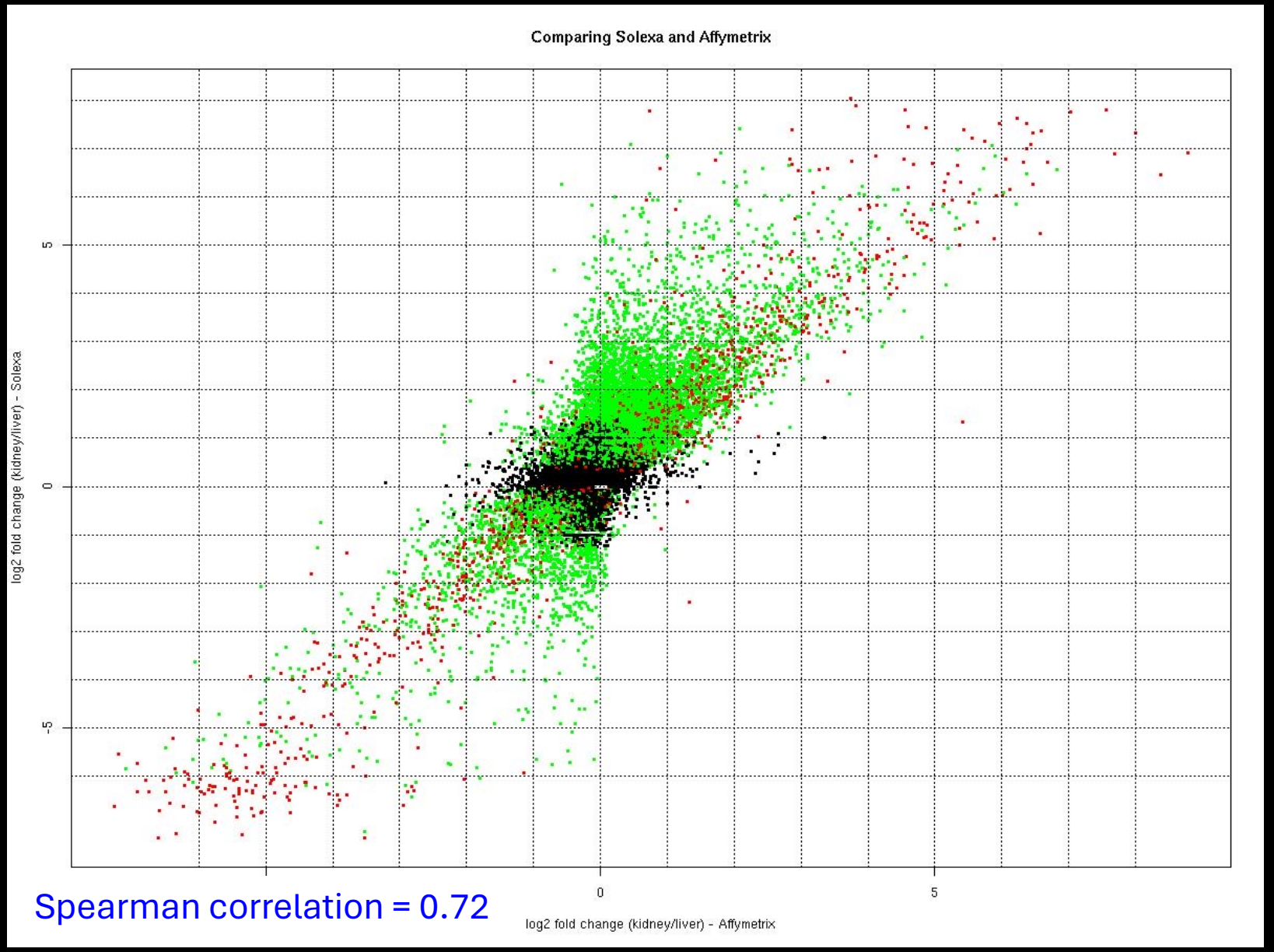
Data Analysis: What genes are differentially expressed?

- Early days—fold change cutoffs (e.g., 2x difference or better)
- not a very satisfying approach:
  - -doesn't take into account variance
  - -misses any small changes



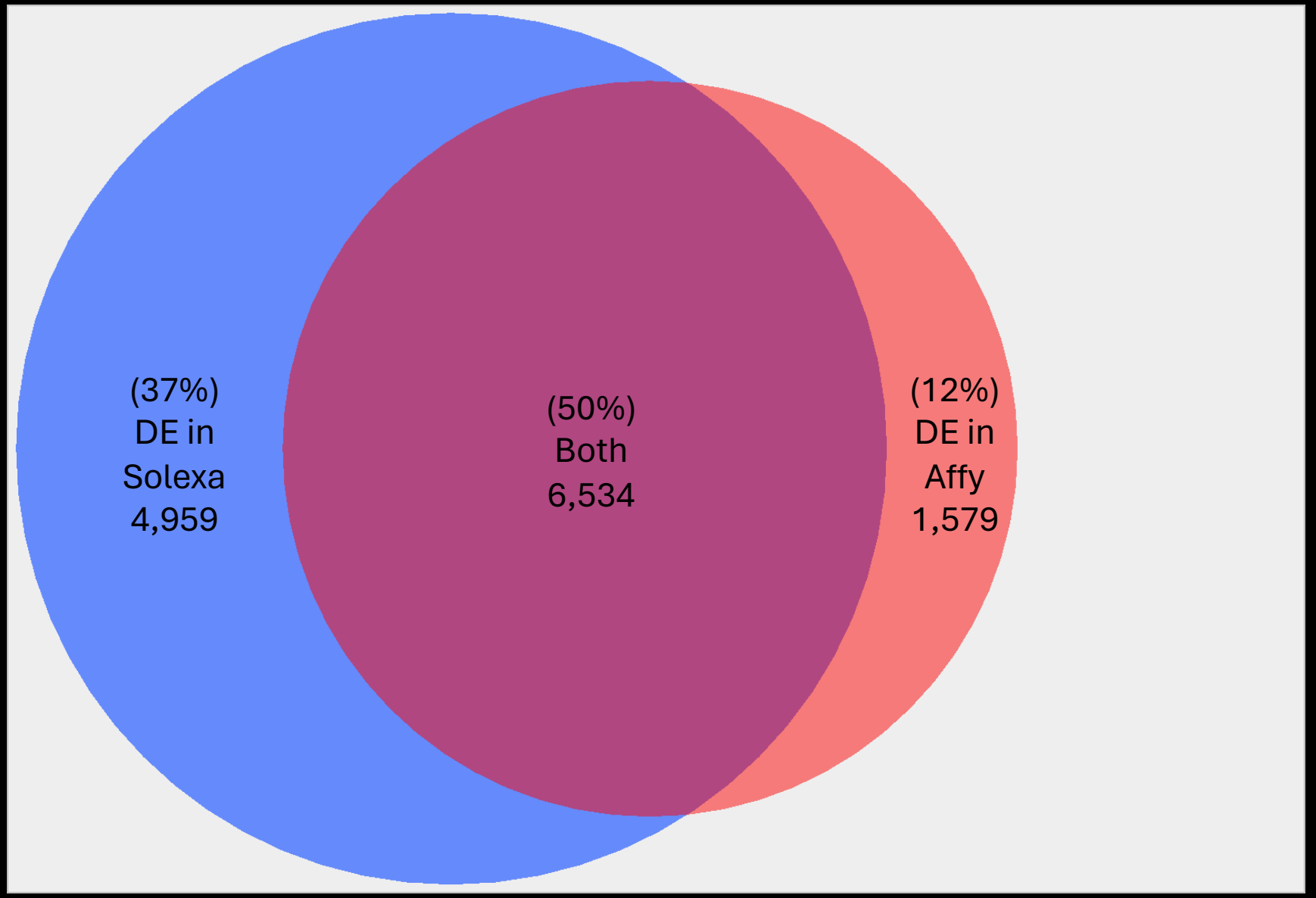
Here, "A" has a fold change >2.5, but varies greatly between replicate experiments. "B" has a fold change of only 1.75, but changes reliably each time the experiment is performed.

Comparing GA and Affy arrays



Spearman correlation = 0.72

13,072 Differentially Expressed (DE) Genes



How many lanes/plates/wells?

<u>Depends on:</u>

1.Read Length

2.Size of Transcriptome

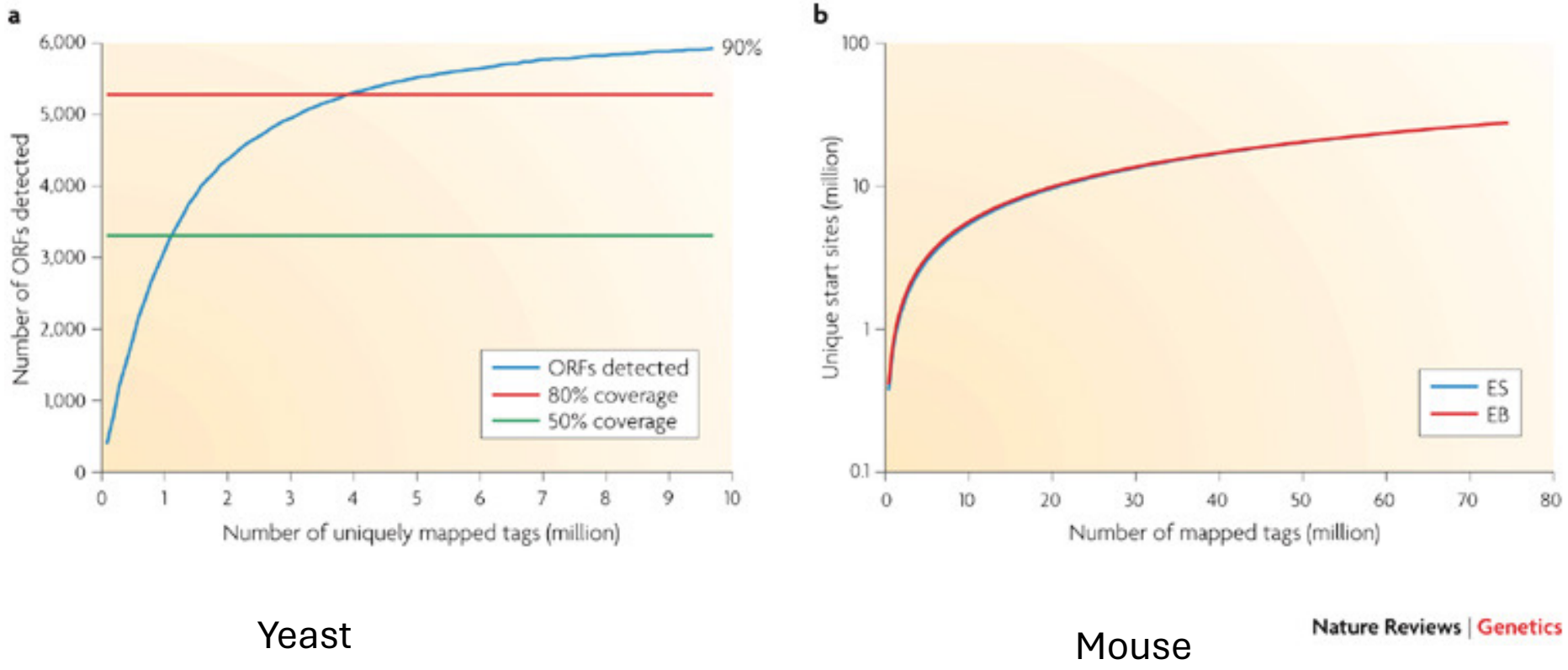
3.Complexity of Transcriptome

4.Cellular Heterogeneity of Tissue

5.Biological Variance

6.Errors (random and systematic)

But, coverage Requirements depend on your species



Yeast

Mouse

Wang, Gerstein, and Snyder, 2008

Metric for RNA-Seq Expression

RPKM:

Reads per Kilobase per Million Reads

Normalizes for (1) gene size and (2) sequencing depth

(~0.1-1 transcript/cell)

$$\text{RPKM} = \frac{N\ reads}{1\ gene} \times \frac{1\ \text{gene}}{B\ bp} \times \frac{1000\ \text{bp}}{1Kb} \times \frac{1\ Million\ reads}{Y\ total\ reads}$$

Y = (exons, introns, intergenic reads)

FPKM=fragments-PKM
is for paired-end data

Mortazavi, Williams, *et al.*
Nature Methods, 2008

RPKM, FPKM, TPM

RPKM:

1.Count up the total reads in a sample and divide that number by 1,000,000 – this is our "per million" scaling factor.
2.Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
3.Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

TPM:

1.Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2.Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3.Divide the RPK values by the "per million" scaling factor. This gives you TPM.

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

TPM normalizes data across replicates better

RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

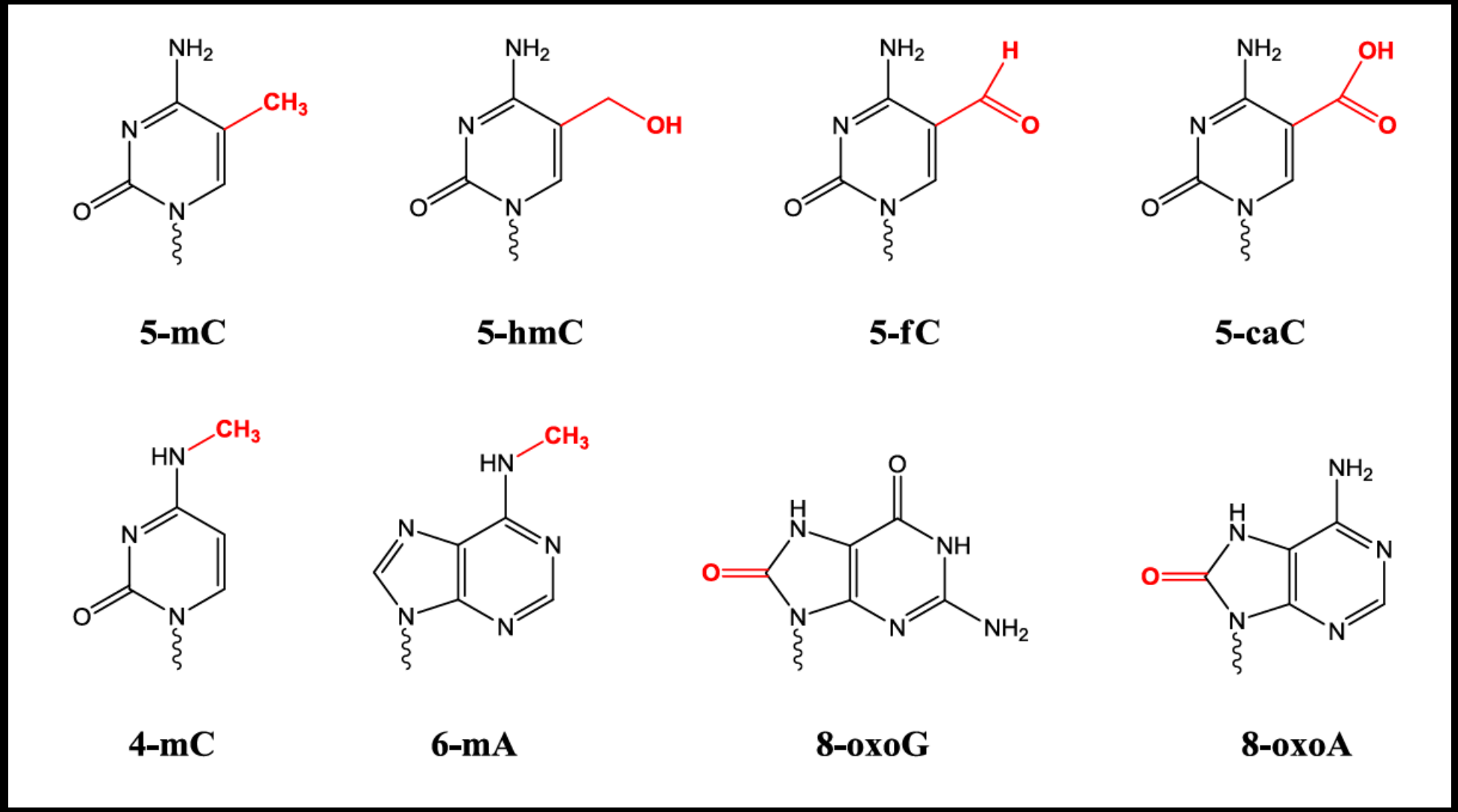
... the sums of each column are very different.

TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

Epitranscriptome

The four-base genome is just the beginning



There are many RNA-mods as well:

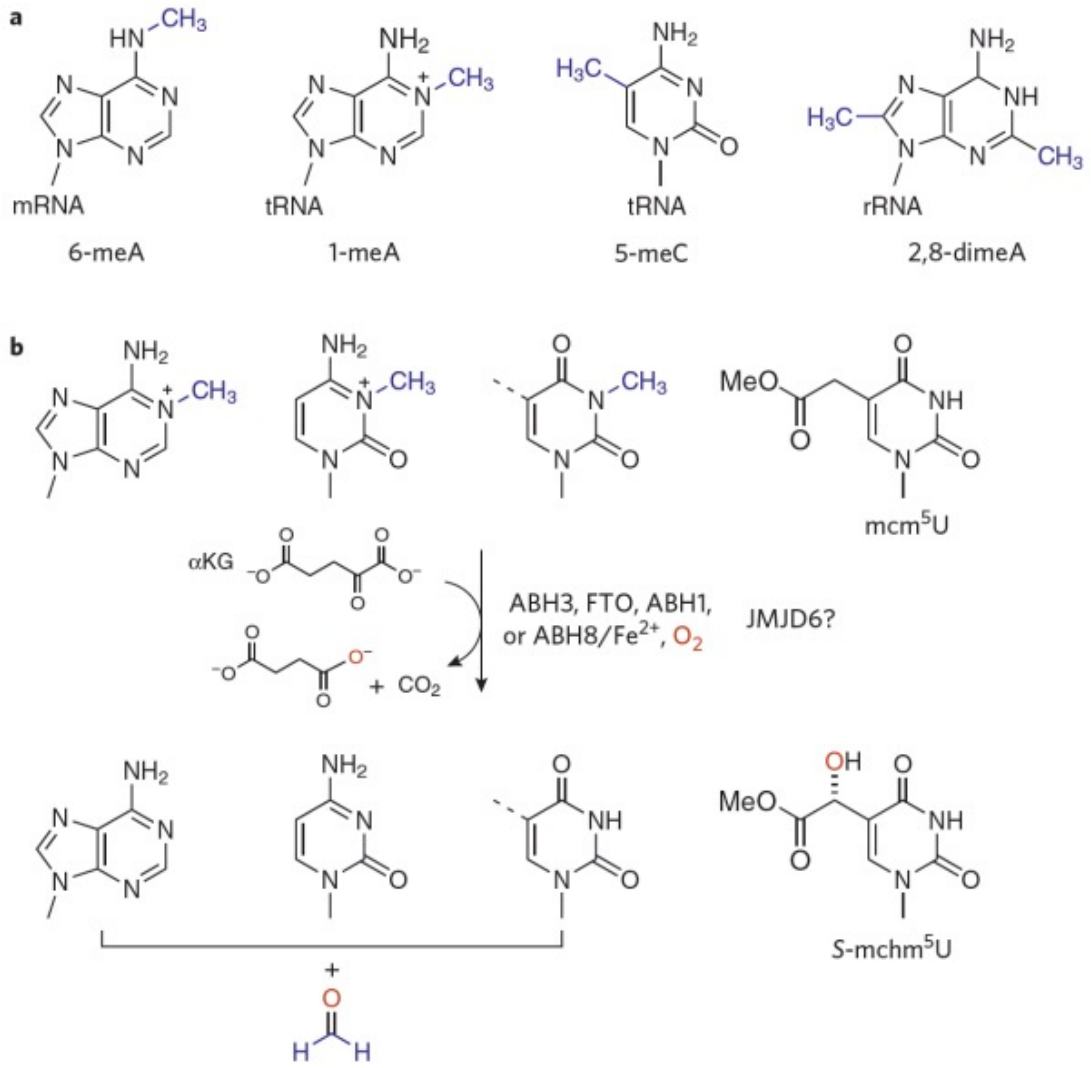
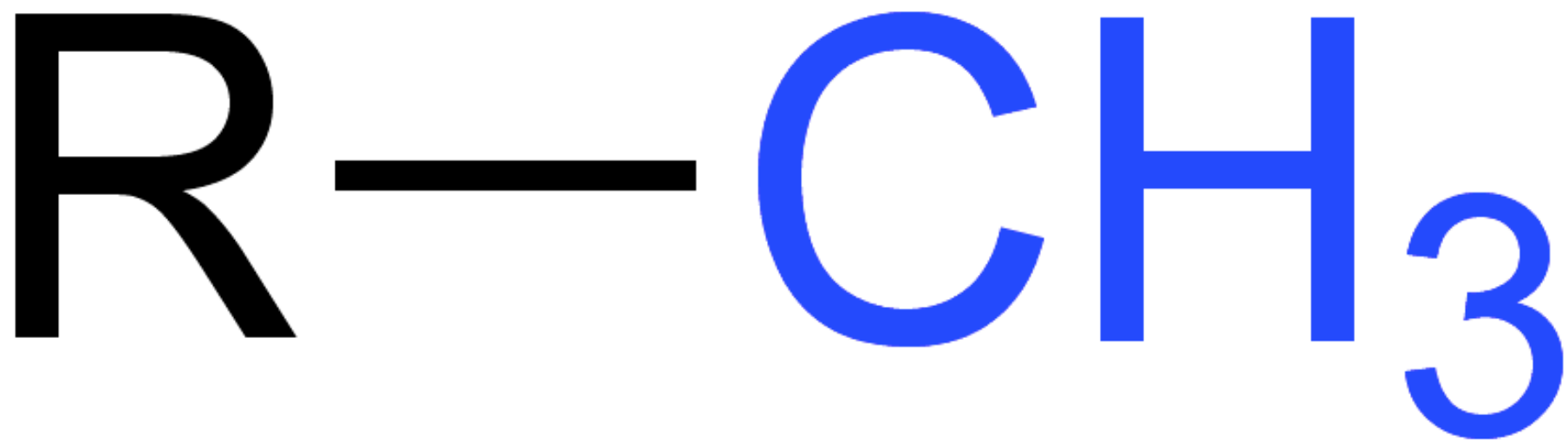
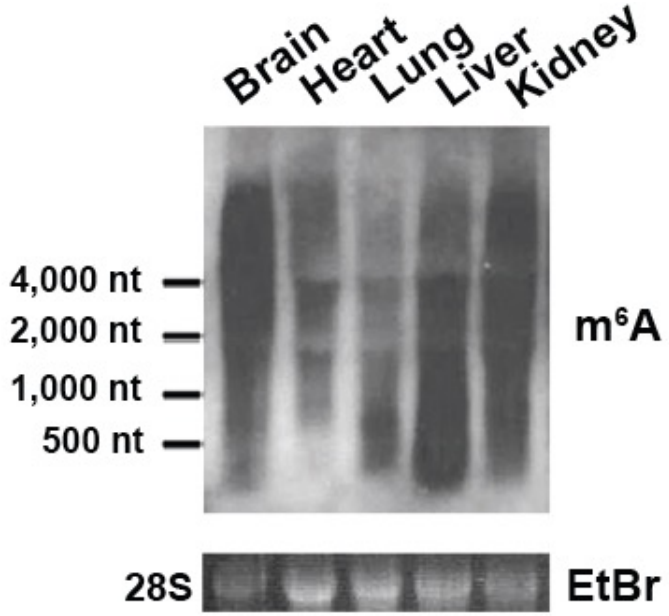


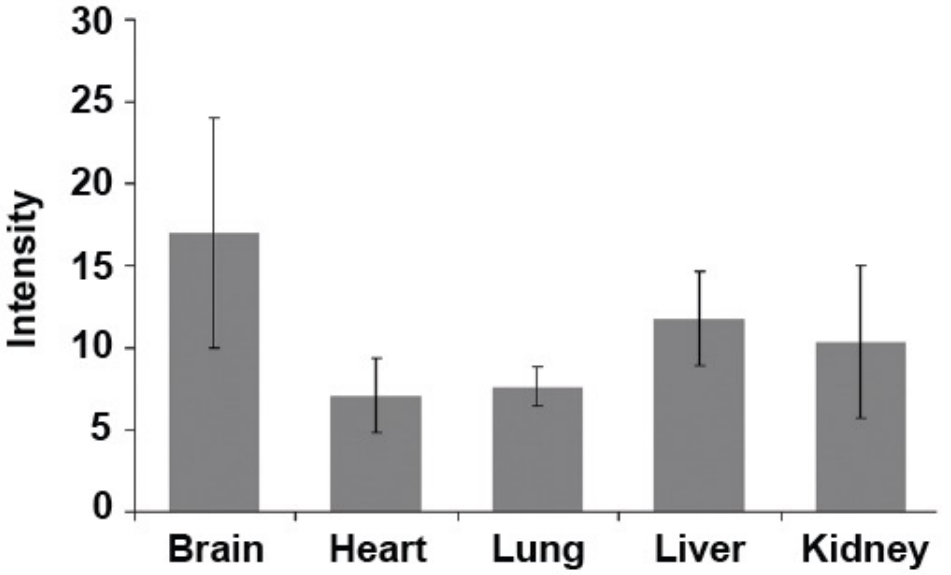
Figure 1 | Examples of RNA modification and demodification that may impact biological regulation. (**a**) Selected examples of RNA base methylation. (**b**) A group of dioxygenases that use iron, α-ketoglutarate and dioxygen to perform oxidation of modified RNA bases for demethylation or hypermodification.

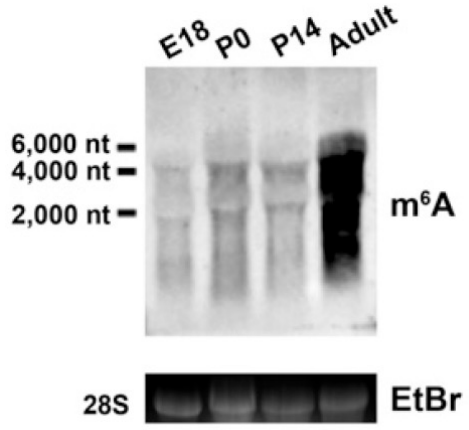
Chuan He

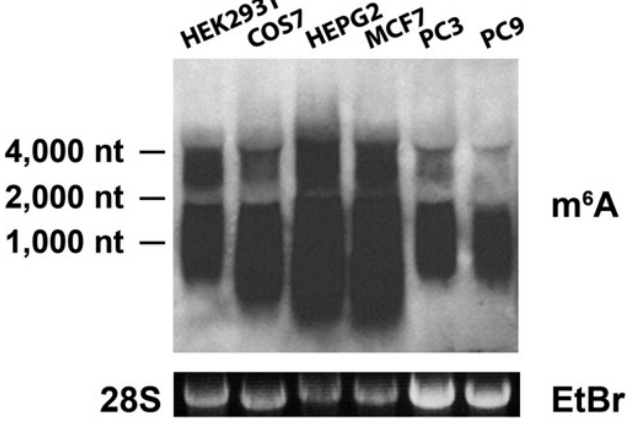

R—CH3

Methylation is important for methyl-6 adenosine (m^6A) in RNA, and is more prominent in brain & adults



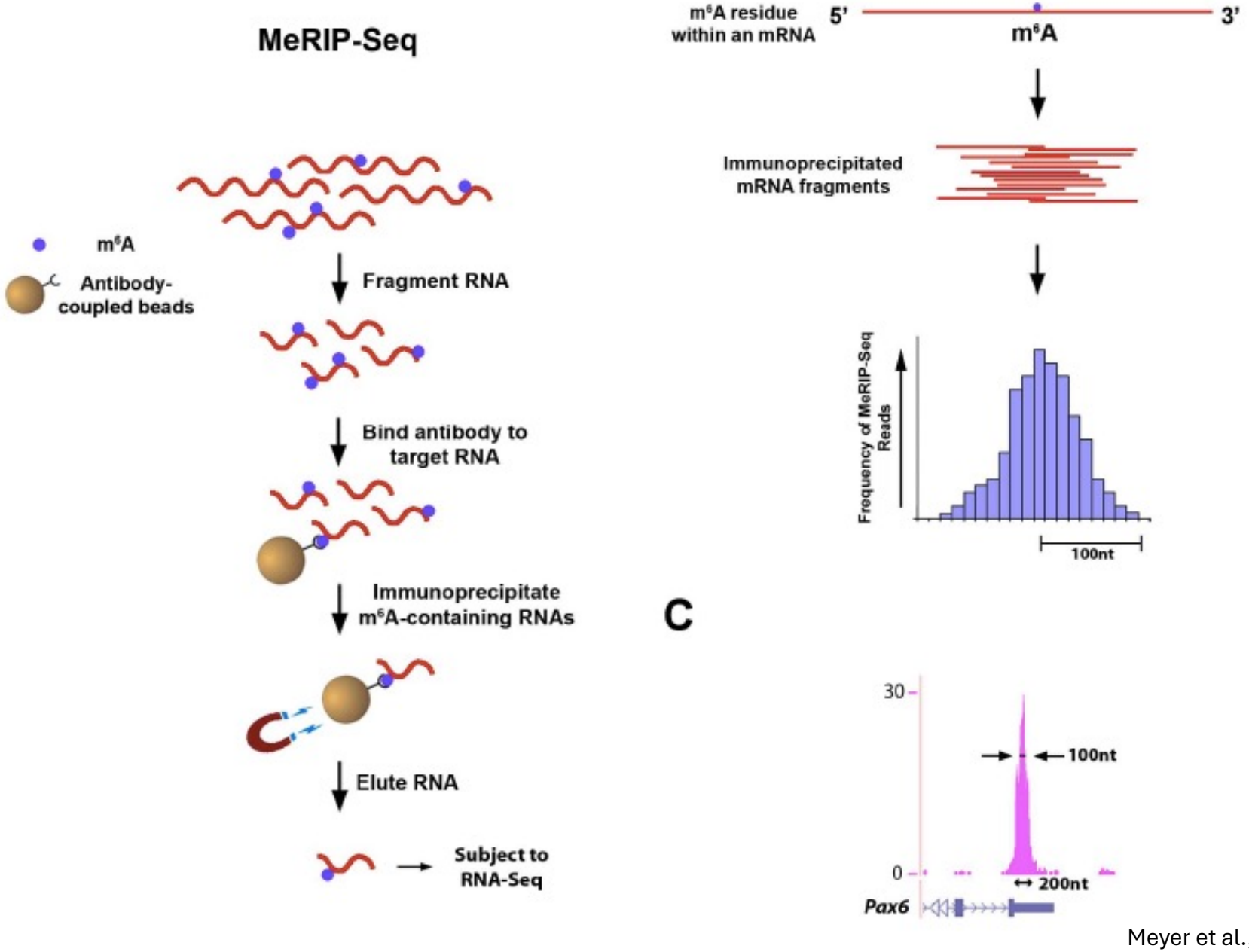






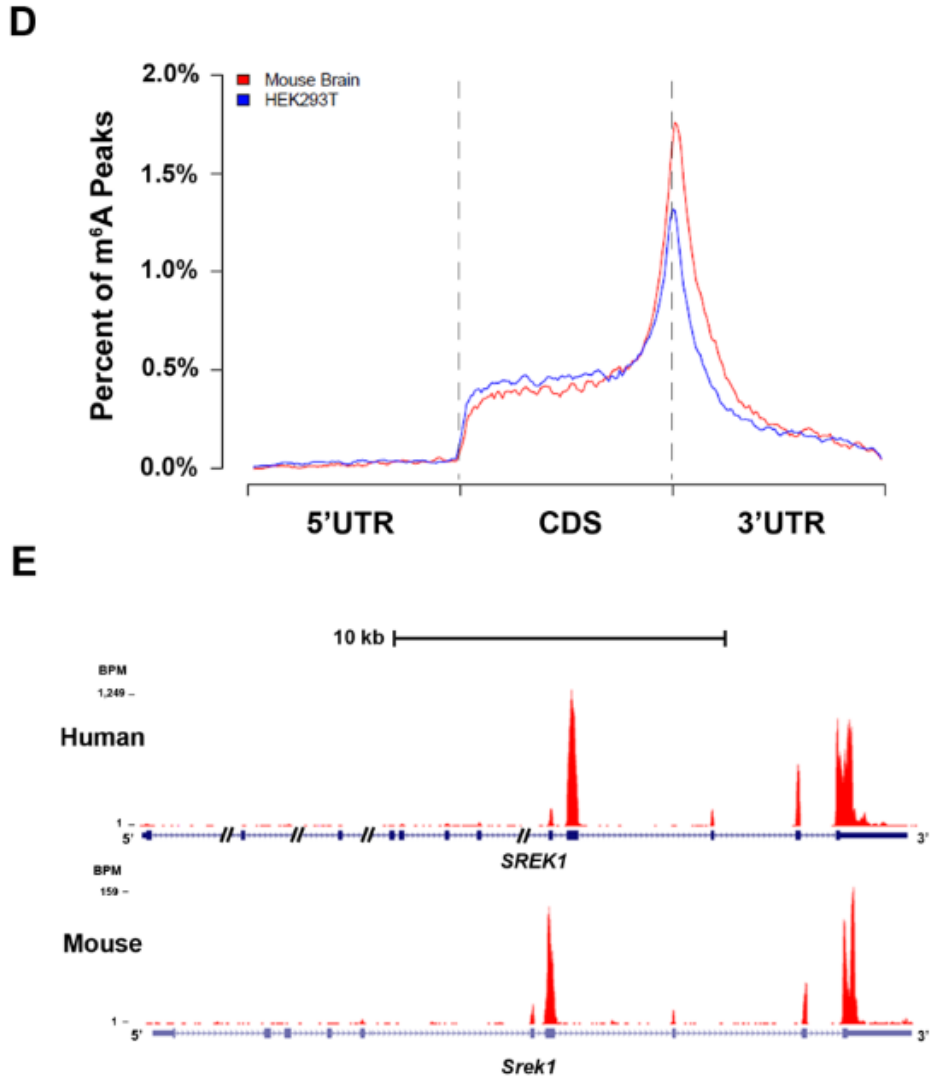
Meyer et al., Cell, 2012

A new method: MeRIP-Seq



Meyer et al., Cell, 2012

Conservations of signal and sites in >10,000 orthologous genes



Meyer et al., Cell, 2012

Global mapping of RNA modifications started in 2012

The birth of the Epitranscriptome: deciphering the function of RNA modifications

Yogesh Saletore[1,2,3], Kate Meyer[4], Jonas Korlach[5], Igor D Vilfan[5], Samie Jaffrey[4] and Christopher E Mason[1,2,*]

Abstract

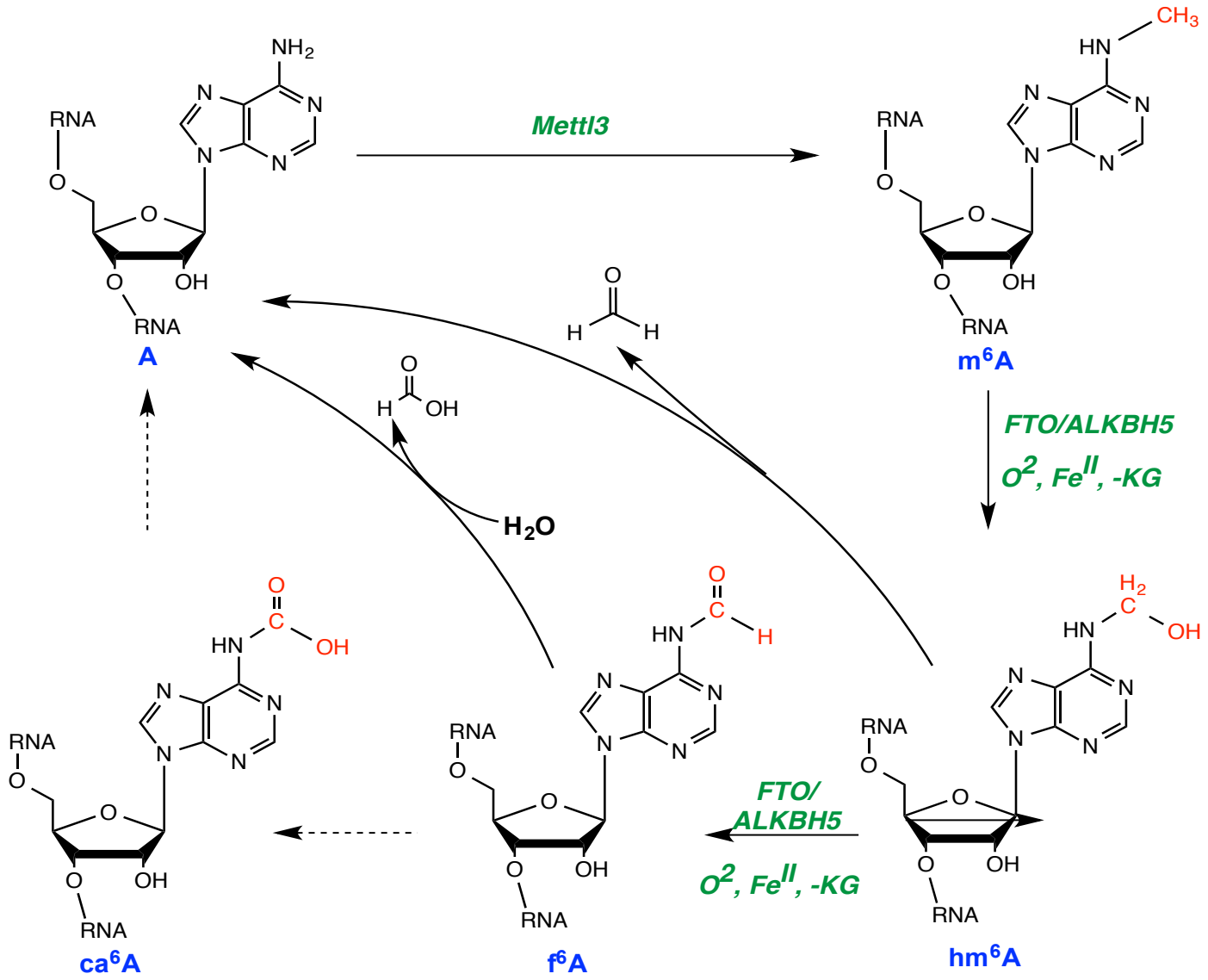
Recent studies have found methyl-6-adenosine in thousands of mammalian genes, and this modification is most pronounced near the beginning of the 3′ UTR. We present a perspective on current work and new single-molecule sequencing methods for detecting RNA base modifications.

Keywords epigenetics, epigenomics, epitranscriptome, m^{6}A, methyl-6-adenosine, methyladenosine, *N*6-methyladenosine, RNA modifications

Project [10]. Similarly, cell-specific, post-translational modifications of proteins, sometimes referred to collectively as the 'epiproteome' [11], are essential mechanisms necessary for the regulation of protein activity, folding, stability and binding partners. Elucidating the roles of protein and DNA modifications has had a major impact on our understanding of cellular signaling, gene regulation and cancer biology [12].

However, our understanding of an additional regulatory layer of biology that rests between DNA and proteins is still in its infancy; namely, the multitude of RNA modifications that together constitute the 'Epitranscriptome.' There are currently 107 known RNA base modifications, with the majority of these having been reported in tRNAs

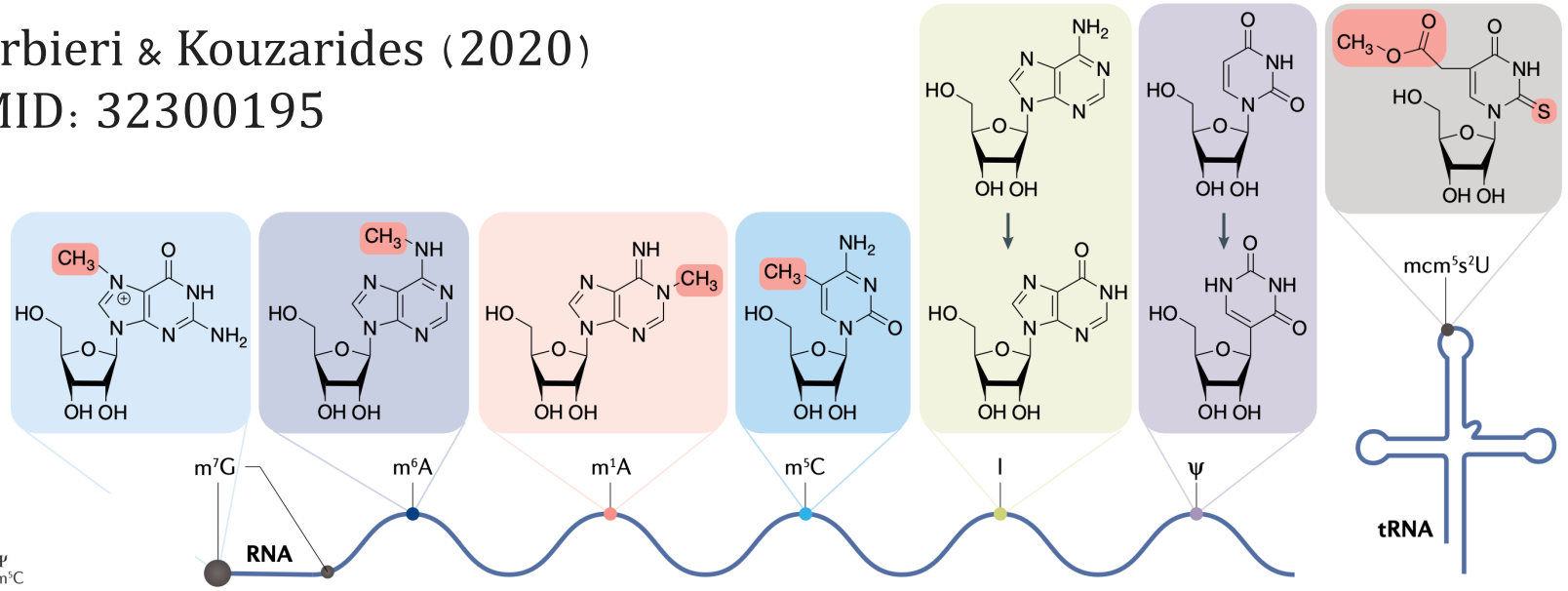
RNA modifications give a new layer of cellular regulation



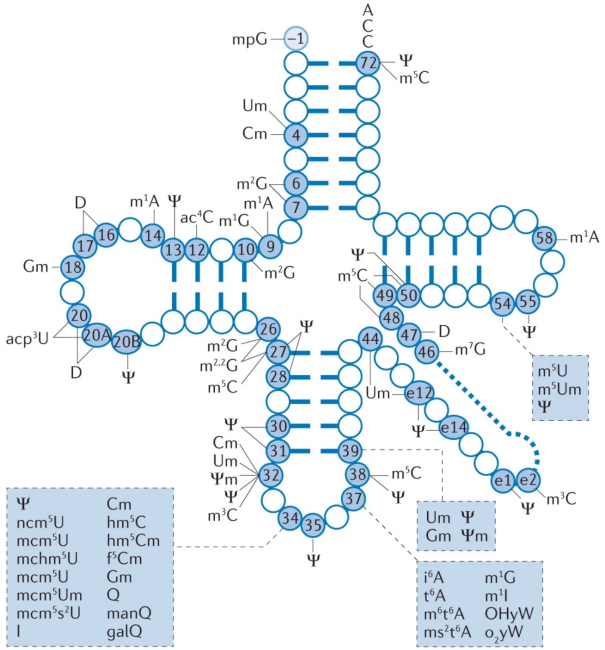
Li and Mason, ARGHG, 2014

Over 400 RNA modifications known today

Barbieri & Kouzarides (2020)
PMID: 32300195



Syzuki (2021)
PMID: 33658722

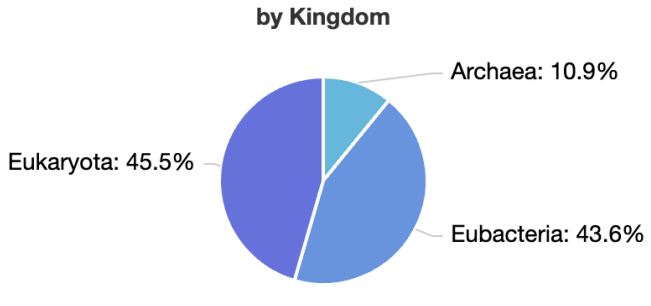


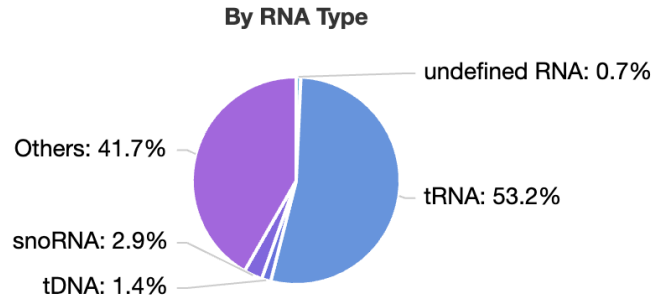
- 25% of all nucleotides in human tRNA are modified, but it can function without modifications
- Bacterial tRNA and rRNA contain significantly less modifications than human tRNA and rRNA and bacterial mRNA has no modifications
- RNA mods evolved to distinguish bacterial from mammalian RNA

MODomics DB

Modifications

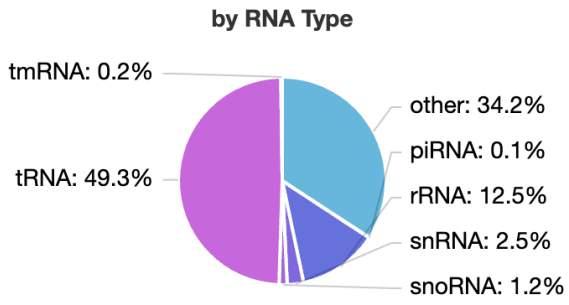
MODOMICS hosts **421** different RNA **modified residues**

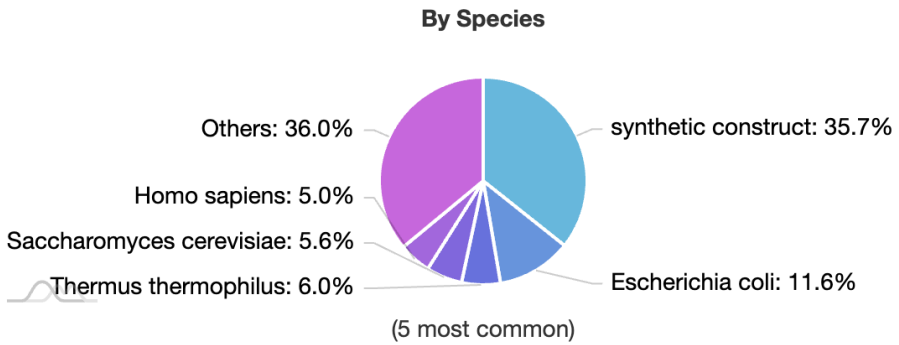




Sequences

MODOMICS hosts **1925** different RNA **Sequences**





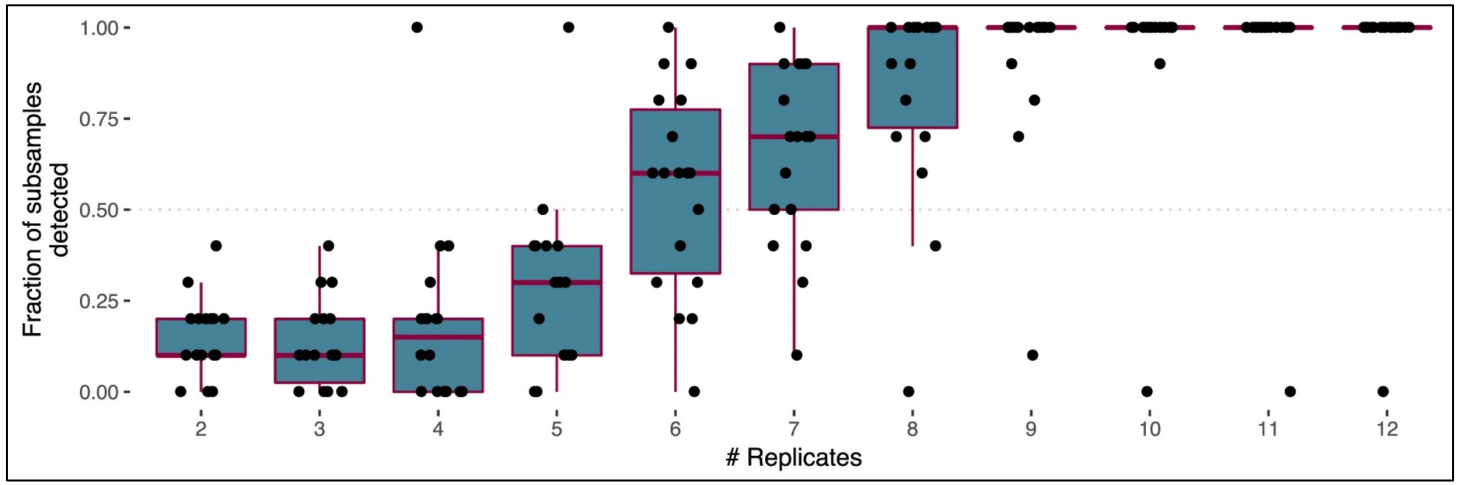
https://genesilico.pl/modomics/

But, MeRIP-seq (m6A-seq) requires many replicates for reliable signal

Limits in the detection of m^{6}A changes using MeRIP/m^{6}A-seq

Alexa B. R. McIntyre, Nandan S. Gokhale, Leandro Cerchietti, Samie R. Jaffrey, Stacy M. Horner & Christopher E. Mason

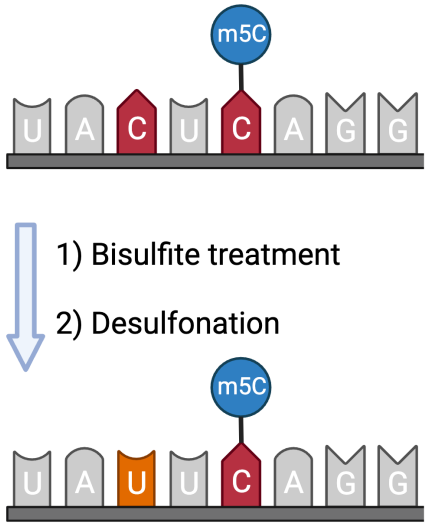
Scientific Reports **10**, Article number: 6590 (2020) | Cite this article



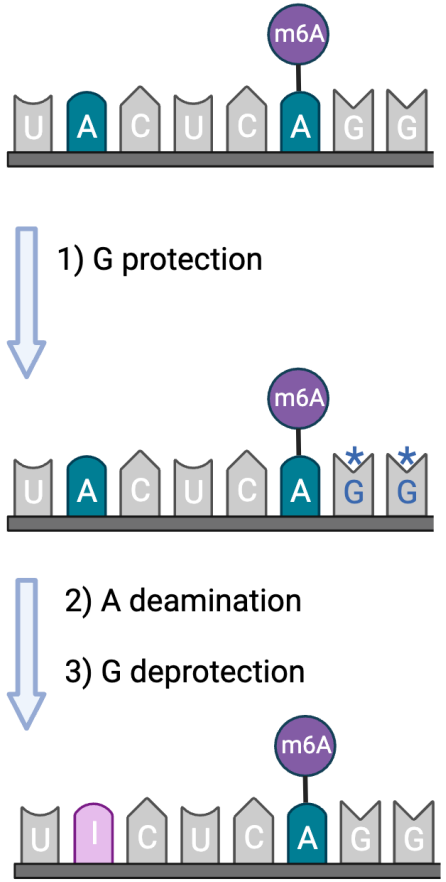
McIntyre, et al., 2020, https://www.nature.com/articles/s41598-020-63355-3

We've now deployed several methods for detecting RNA modifications, but direct RNA-seq with nanopore is the most comprehensive; we get **all mods** & **long reads**

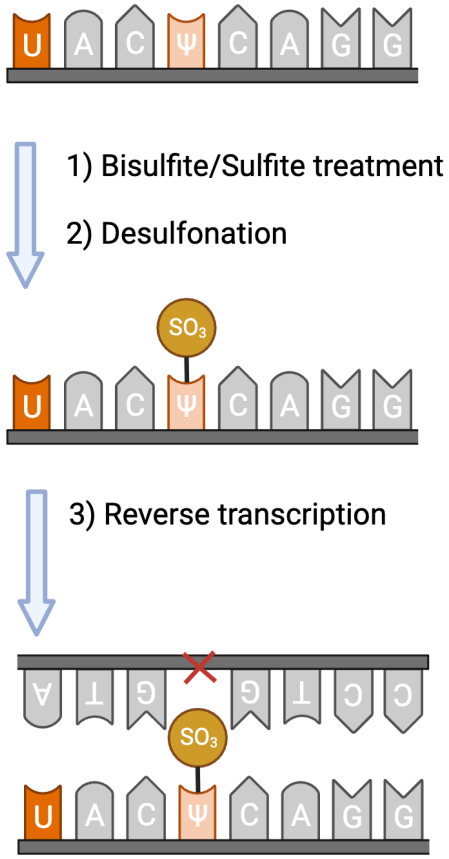
BIS-seq



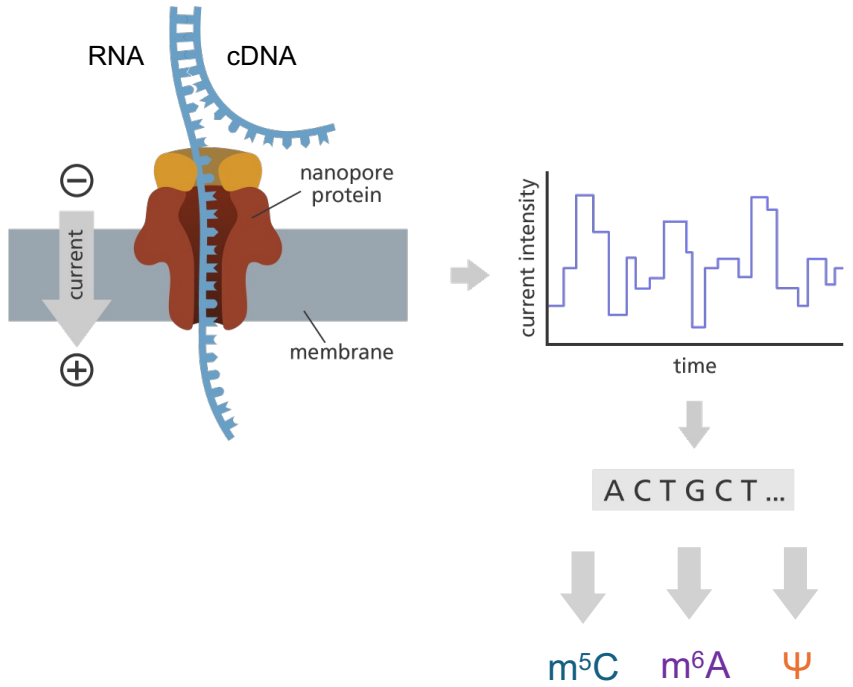
GLORI-seq



PRAISE



Nanopore direct-RNA



nature genetics

Comment | Published: 15 July 2021

A call for direct sequencing of full-length RNAs to identify all modifications

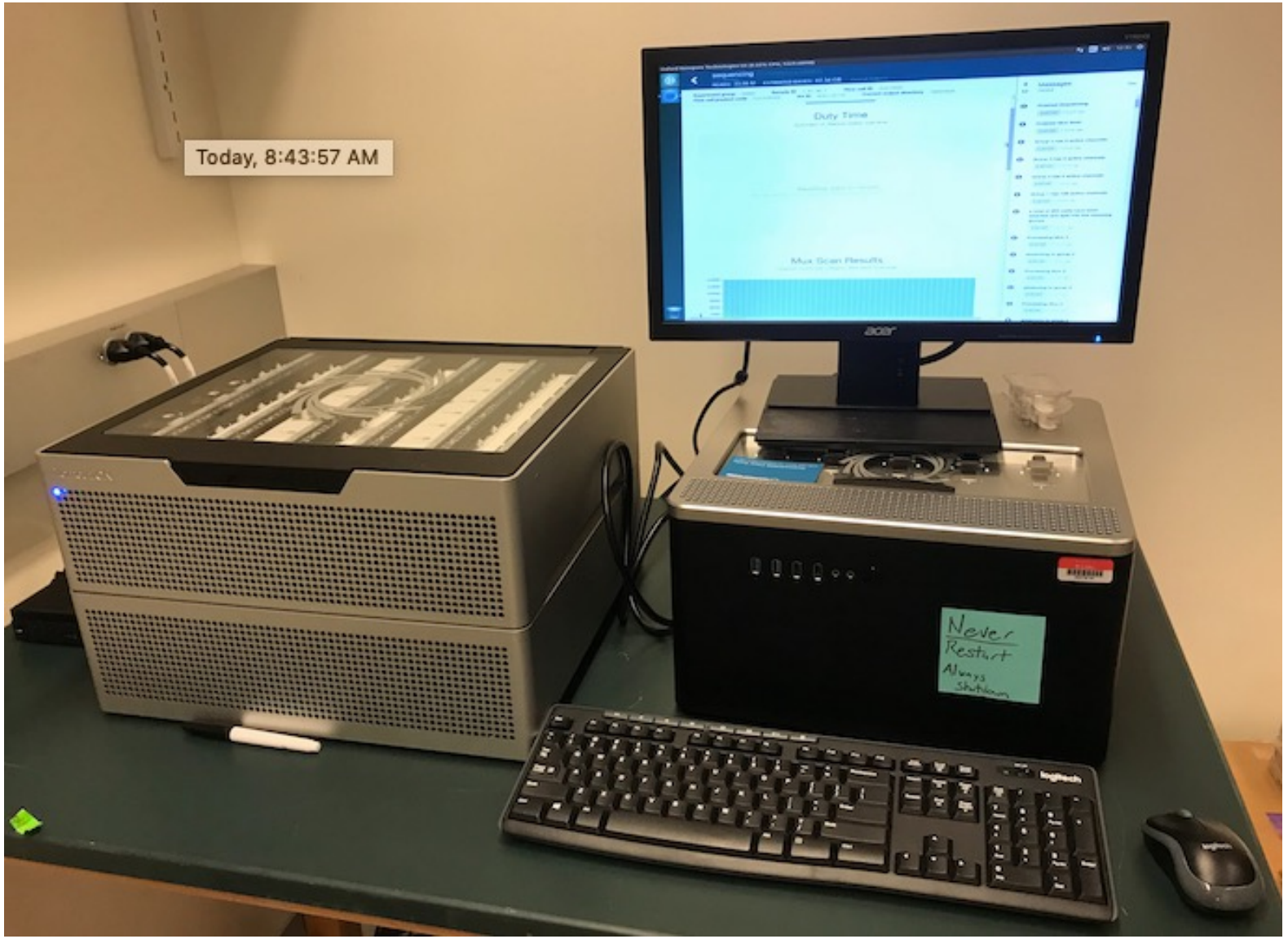
Juan D. Alfonzo, Jessica A. Brown, Peter H. Byers, Vivian G. Cheung, Richard J. Maraia & Robert L. Ross

Nature Genetics **53**, 1113–1116 (2021) | Cite this article

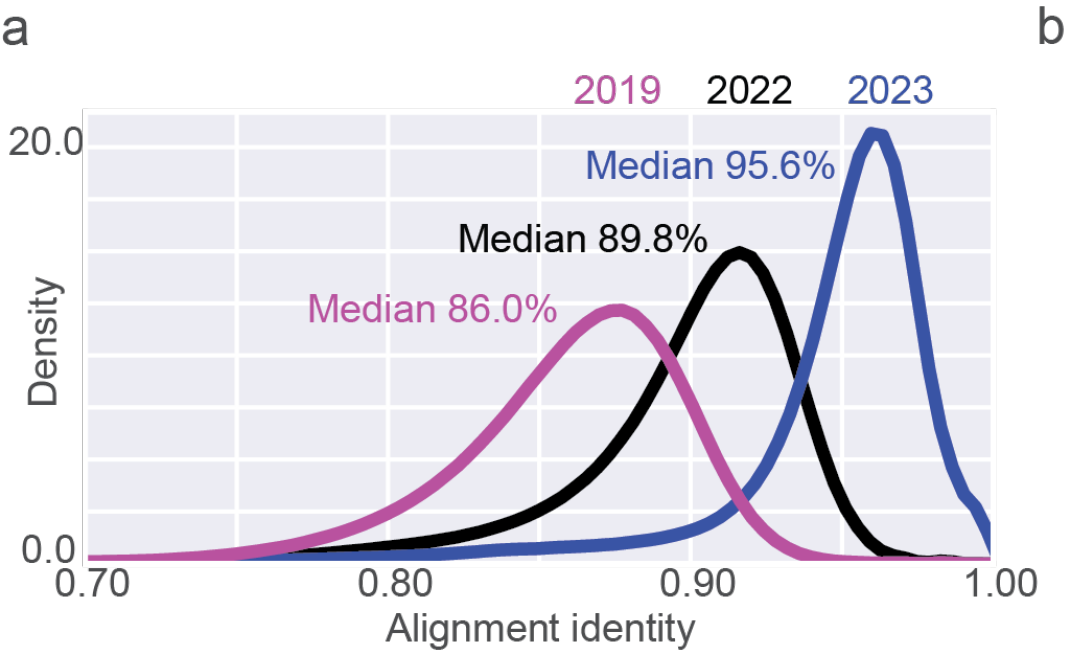
For most organisms, DNA sequences are available, but the complete RNA sequences are not. Here, we call for technologies to sequence full-length RNAs with all their modifications.

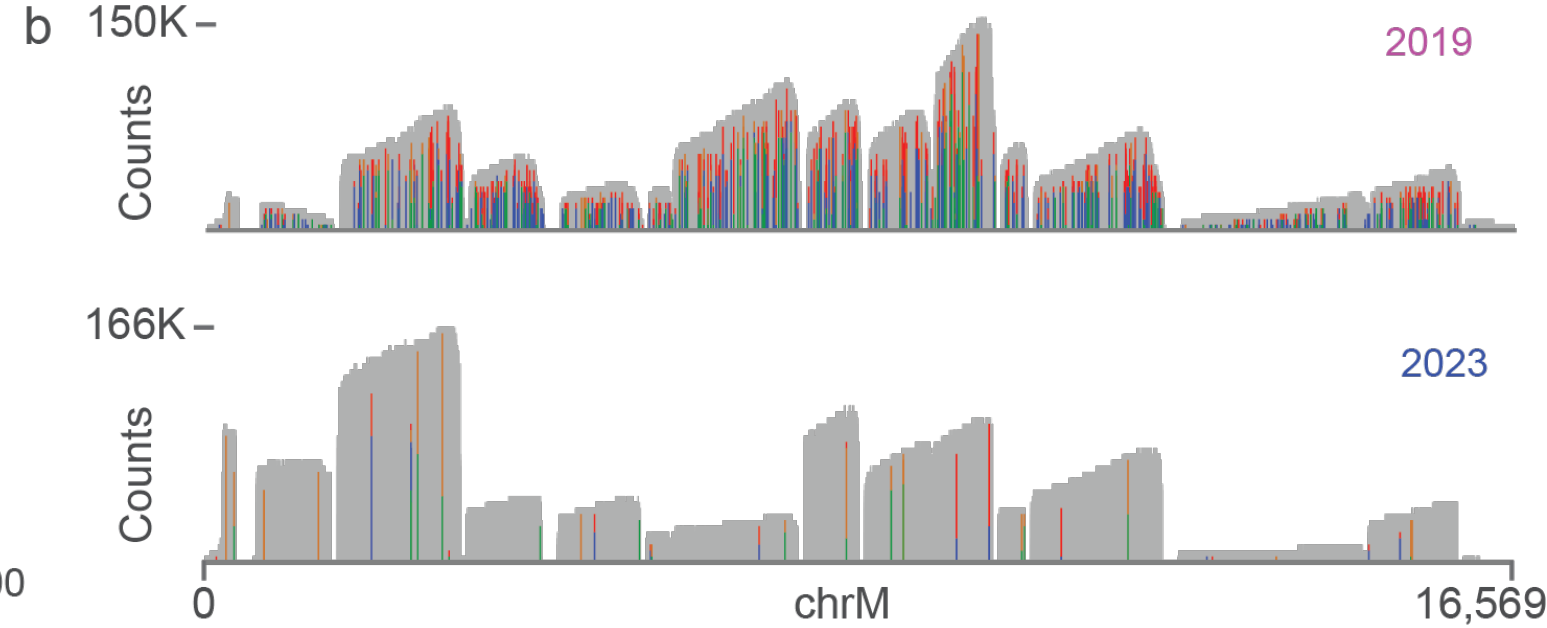
https://www.nature.com/articles/s41588-021-00903-1

Long reads! ONT PromethION



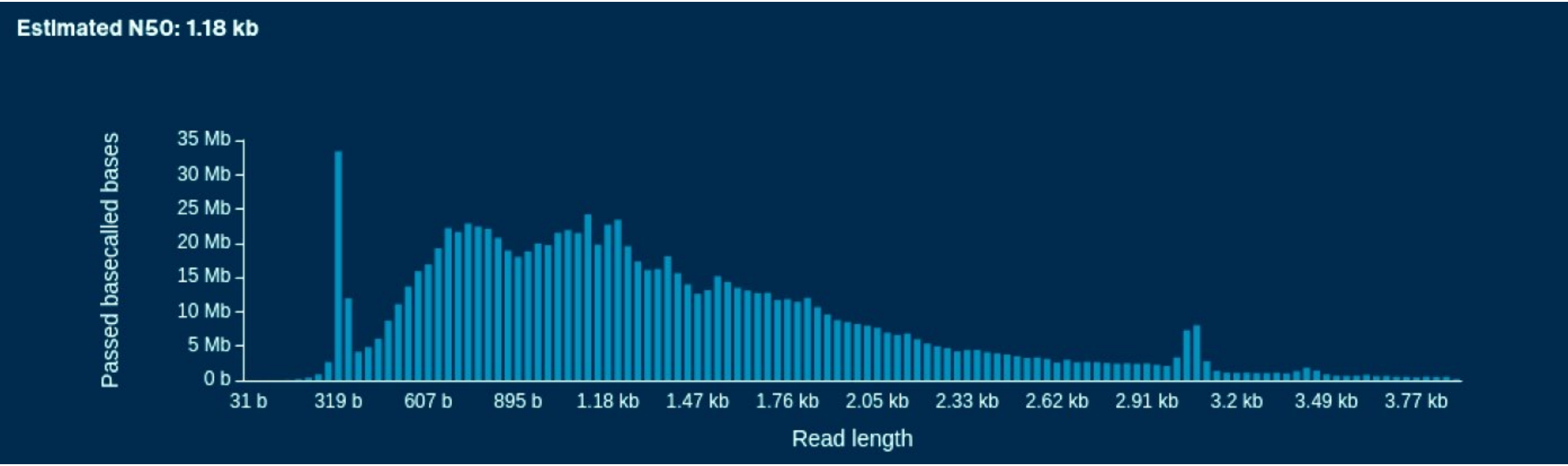
Direct RNA-sequencing from ONT finally reached maturity in 2023





Miten Jain

Direct RNA runs can show a large range of transcript sizes



New Oxford Nanopore Technologies (ONT) pricing enables efficiency at scale, if you have the cash for it

Flow Cell Pack Size	Number of Flow Cells	Cost per Flow Cell (USD)	Cost for Flow Cells (USD)	Cost of Libray Prep (USD)	Total Cost (USD)	Cost/Sample
1	4	$855.00	$3,420.00	$399.33	$3,819.33	$954.83
3	12	$855.00	$10,260.00	$1,198.00	$11,458.00	$954.83
8	32	$820.00	$26,240.00	$3,194.67	$29,434.67	$919.83
24	96	$785.00	$75,360.00	$9,584.00	$84,944.00	$884.83
48	192	$745.00	$143,040.00	$19,168.00	$162,208.00	$844.83
128	512	$680.00	$348,160.00	$51,114.67	$399,274.67	$779.83
256	1024	$630.00	$645,120.00	$102,229.33	$747,349.33	$729.83
720	2880	$600.00	$1,728,000.00	$287,520.00	$2,015,520.00	$699.83

SQK-RNA004, Direct RNA Sequencing Kit & Flow Cell
April 29, 2024 pricing from ONT web site

Please cite this article in press as: Zheng et al., ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2012.10.015

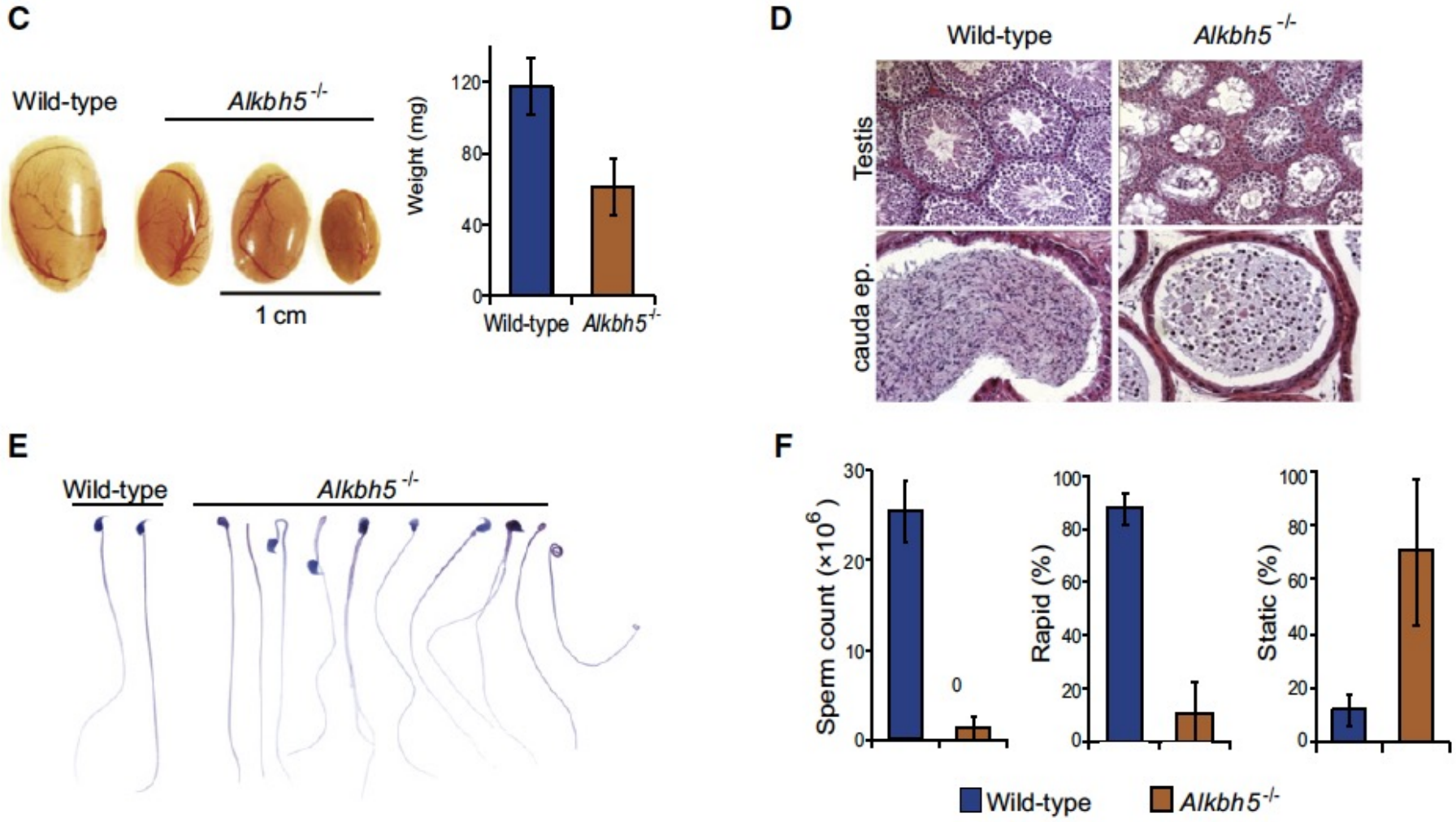
Molecular Cell

Article

ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility

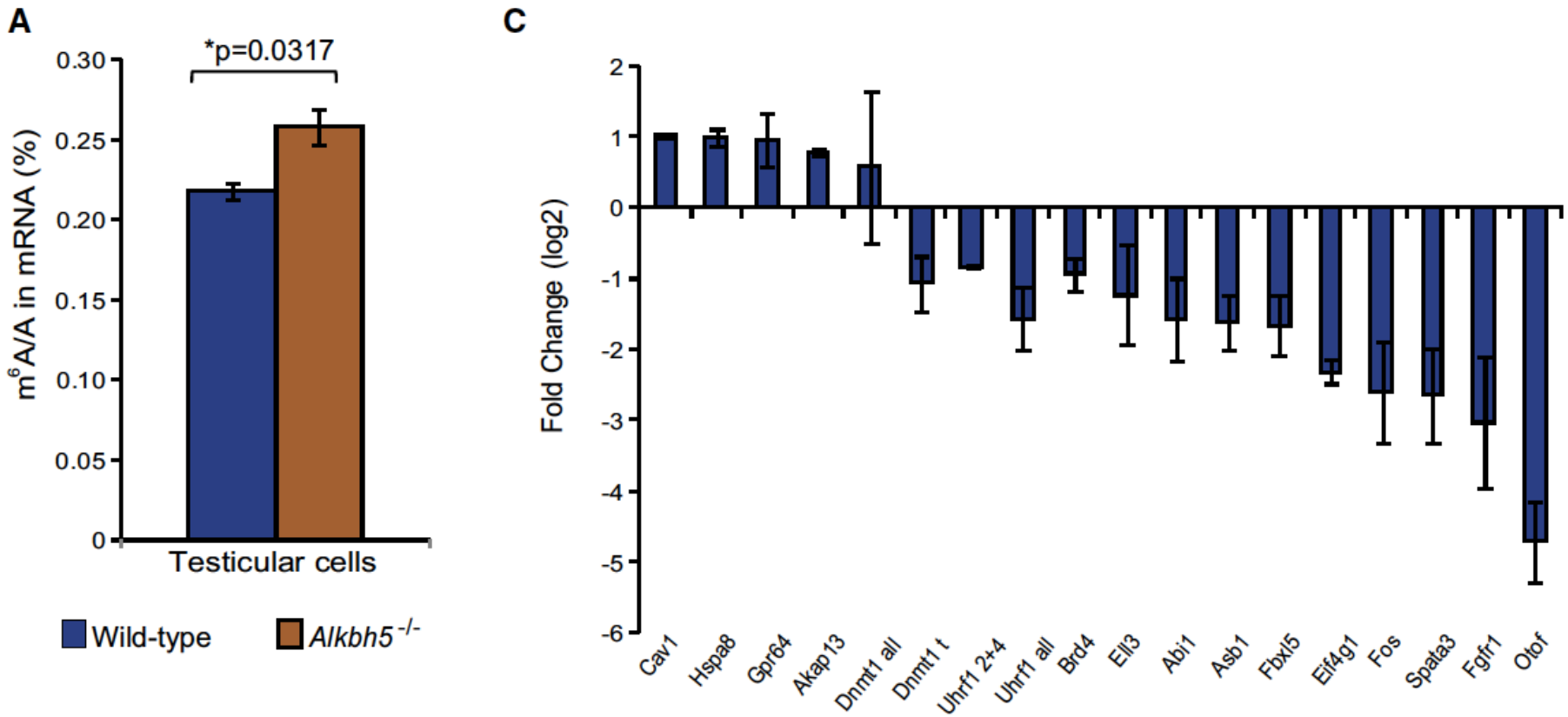
Guanqun Zheng,[1,11] John Arne Dahl,[3,11] Yamei Niu,[2,11] Peter Fedorcsak,[4] Chun-Min Huang,[2] Charles J. Li,[1] Cathrine B. Vågbø,[6] Yue Shi,[2,7] Wen-Ling Wang,[2,7] Shu-Hui Song,[5] Zhike Lu,[1] Ralph P.G. Bosmans,[1] Qing Dai,[1] Ya-Juan Hao,[2,7] Xin Yang,[2,7] Wen-Ming Zhao,[5] Wei-Min Tong,[8] Xiu-Jie Wang,[9] Florian Bogdan,[3] Kari Furu,[3] Ye Fu,[1] Guifang Jia,[1] Xu Zhao,[2,7] Jun Liu,[10] Hans E. Krokan,[6] Arne Klungland,[3,*] Yun-Gui Yang,[2,7,*] and Chuan He[1,*]

RNA m^{6}A defects perturb germline development

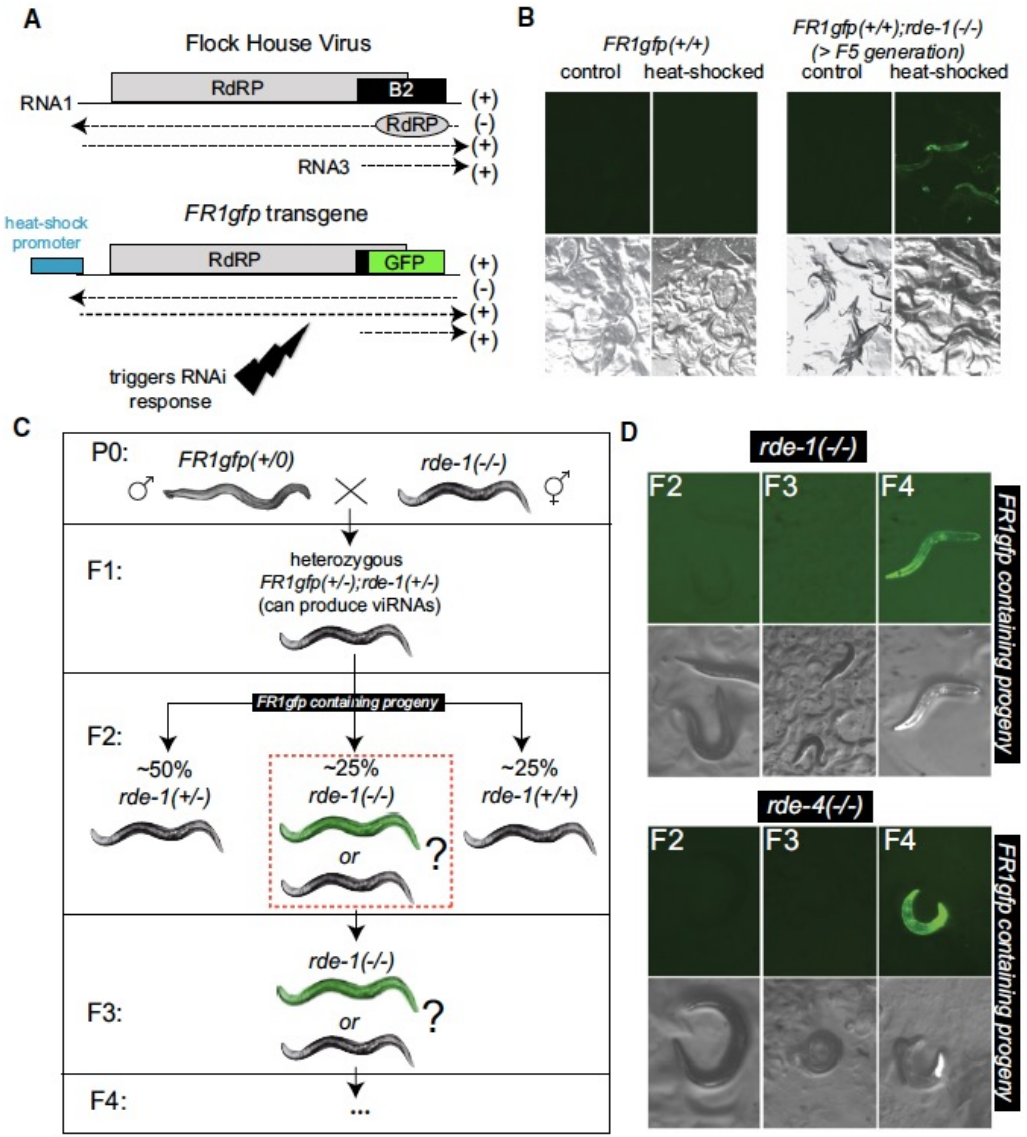


Zheng et al., Mol. Cell, 2013

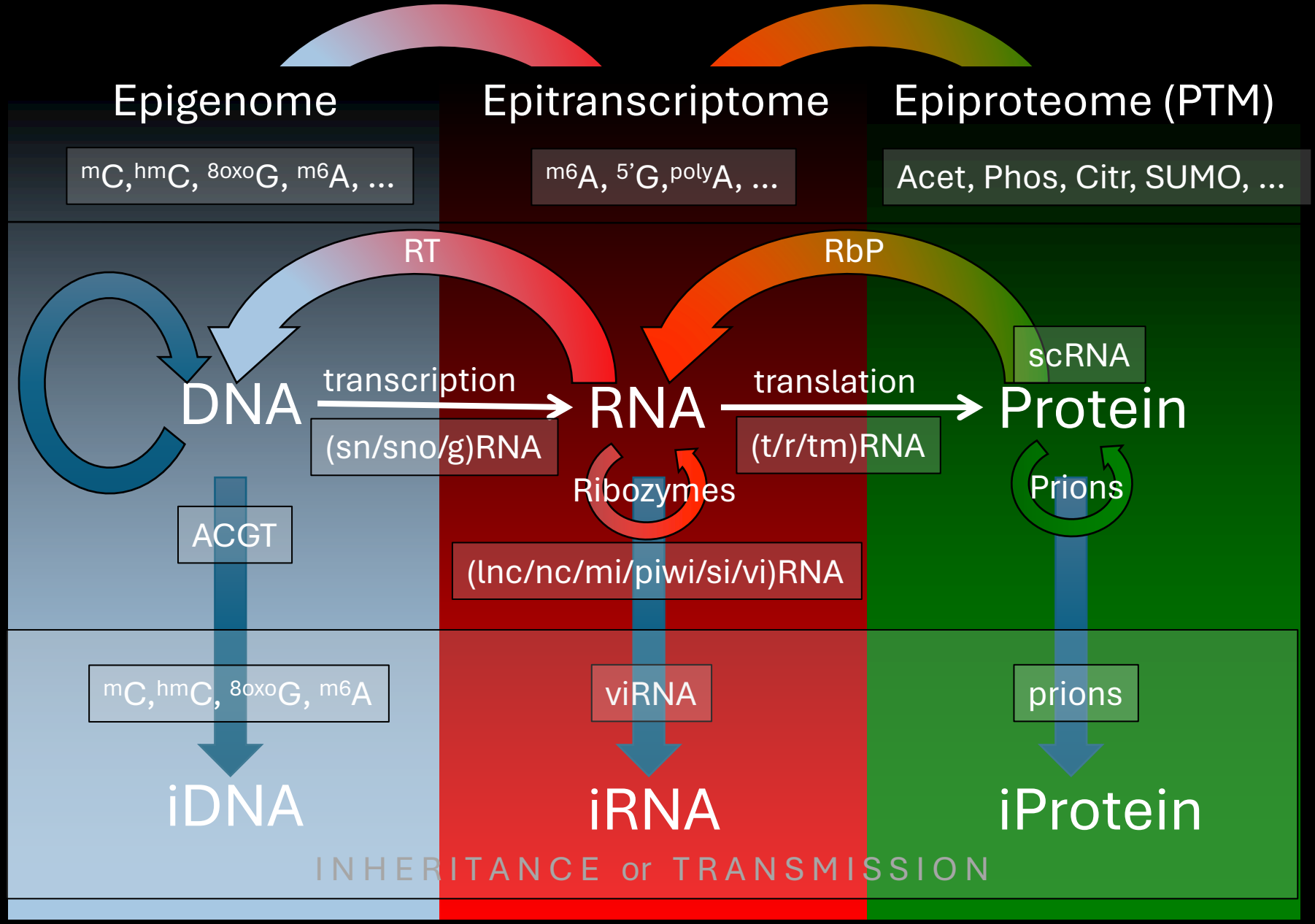
Dysregulated m^{6}A affects many epigenetic modifiers



Information also pass between generations in RNA
Evidence of a Trans-generational Anti-viral RNAi response



Rechavi et al, 2011



from Saletore *et al.*, Genome Biology, 2012

Questions?