Clinical and Research Genomics

Lecture 04 & 06
RNA-Sequencing, Epitranscriptomes, and Single Cells

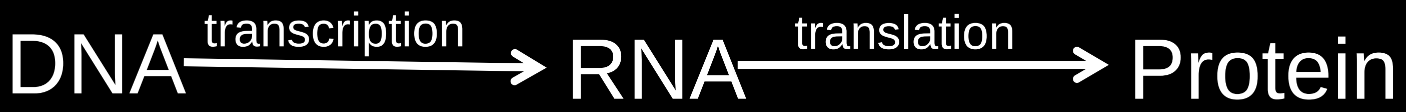
Dr. Christopher E. Mason

-

April 5, 2022

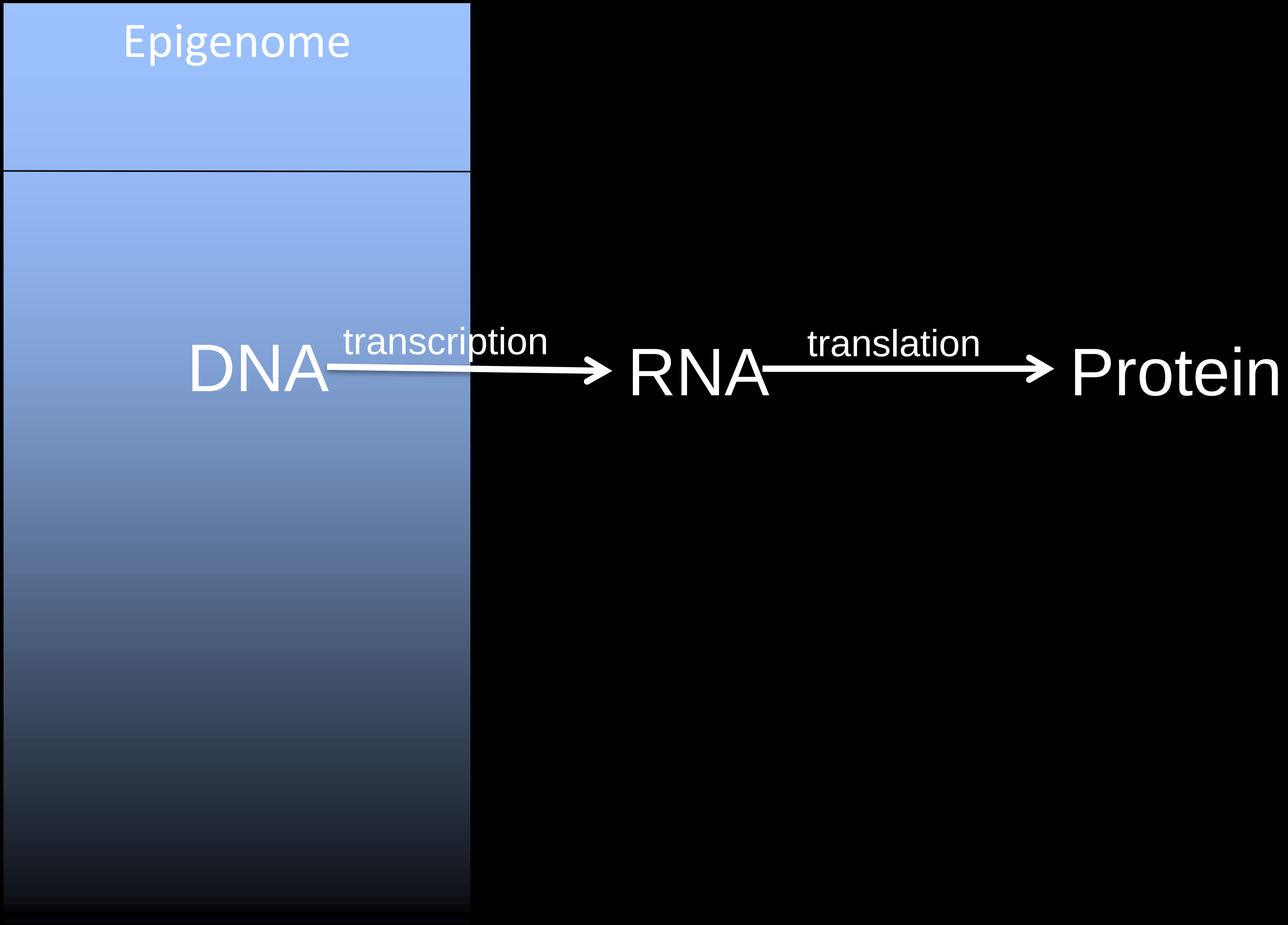
(1)

Background



DNA transcription RNA translation Protein

James Watson, 1958



Epigenome

DNA

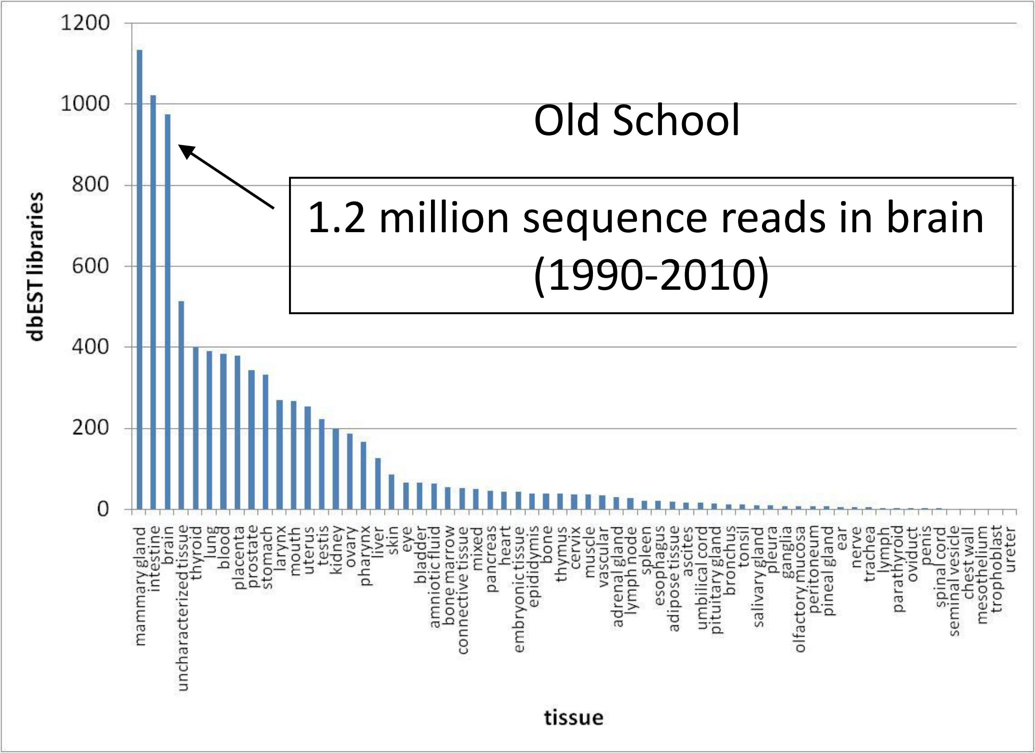
transcription

RNA

translation

Protein

We can observe
many, many more molecules than before

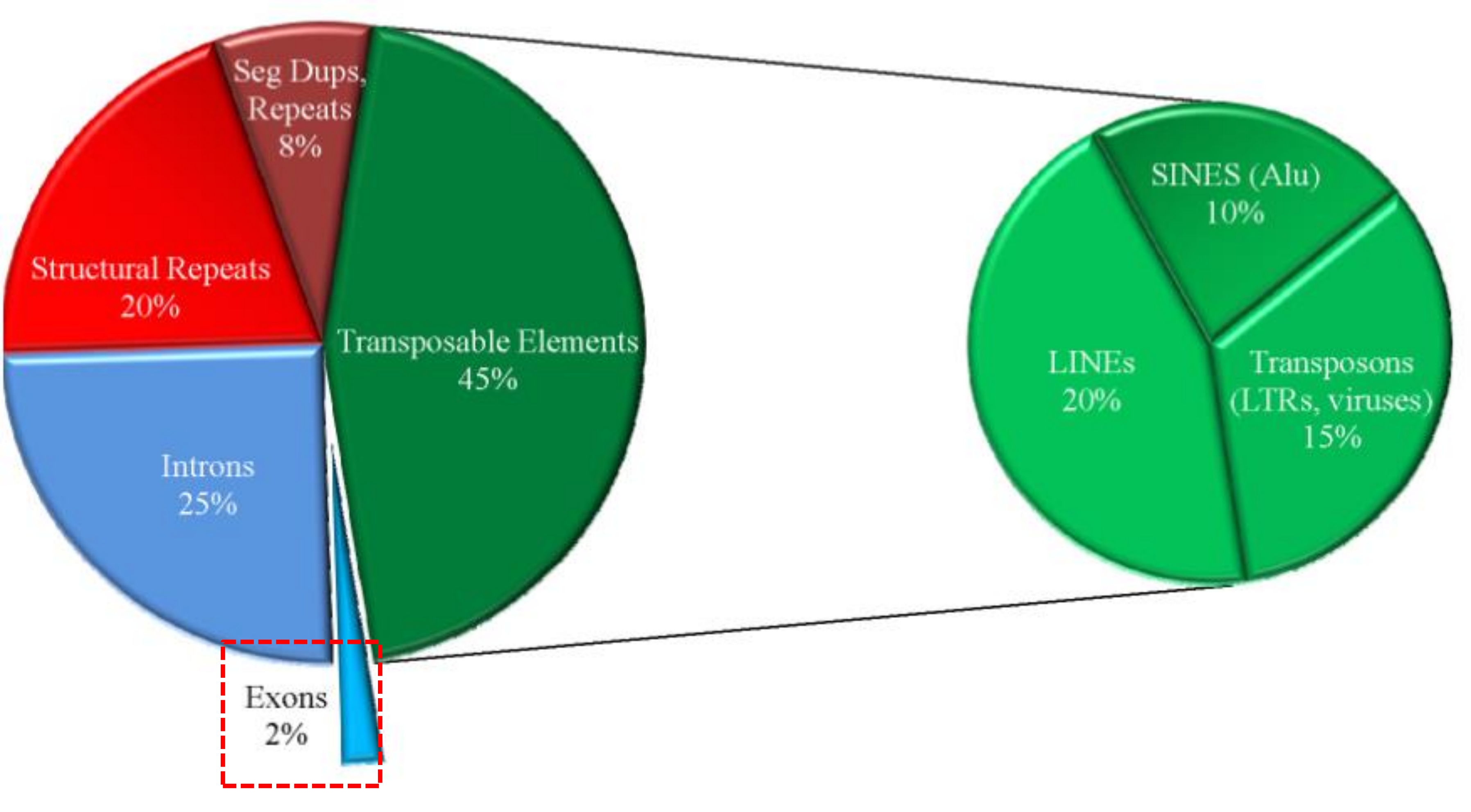


Old School

1.2 million sequence reads in brain
(1990-2010)

New School:
One run of a NGS machine = billions of sequence reads in days

The Annotation/Composition of the Human Genome



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

Validation of known Gene Boundaries

Adult

Amygdala

Hippocampus

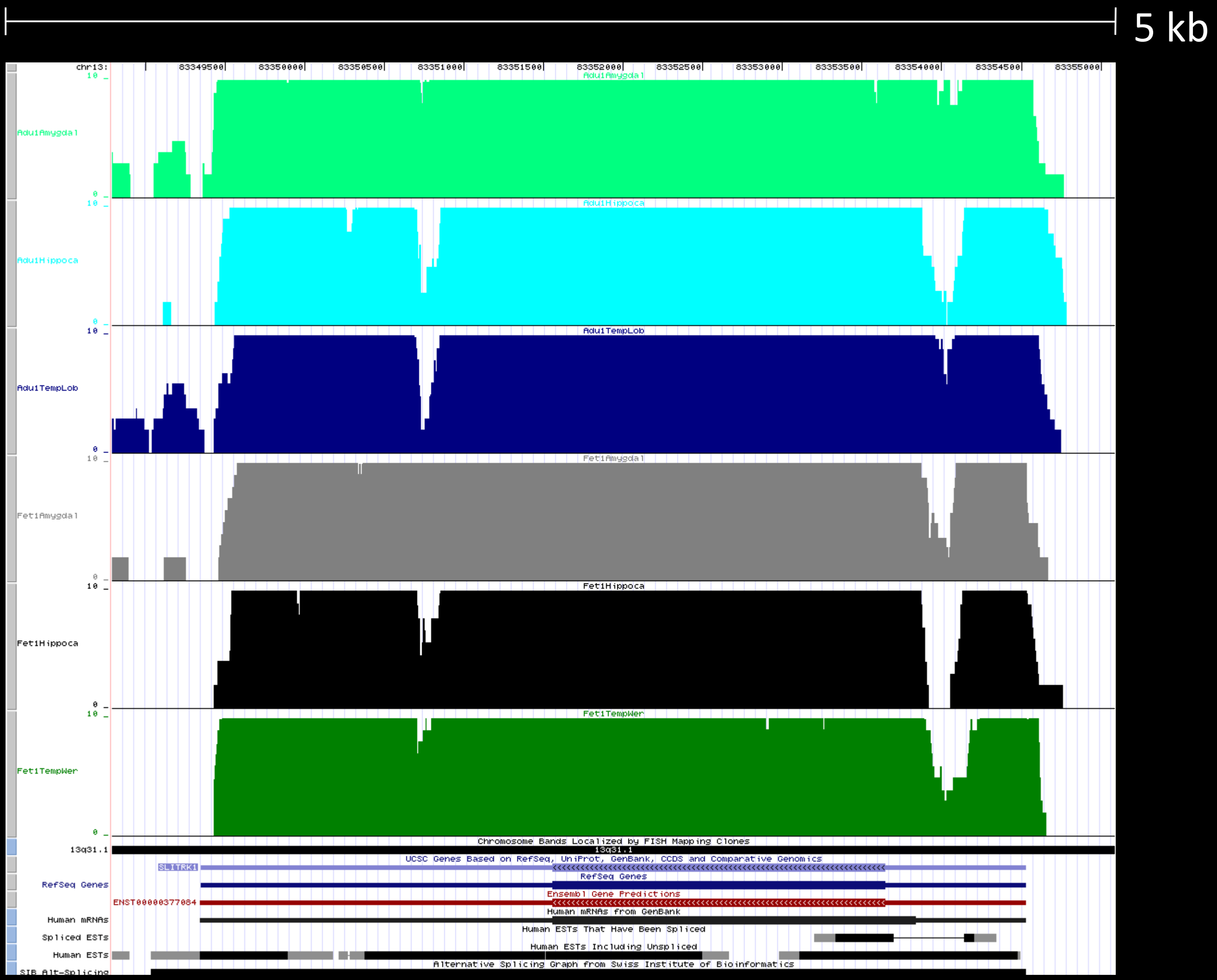
Temporal Lobe

Fetal

Amygdala

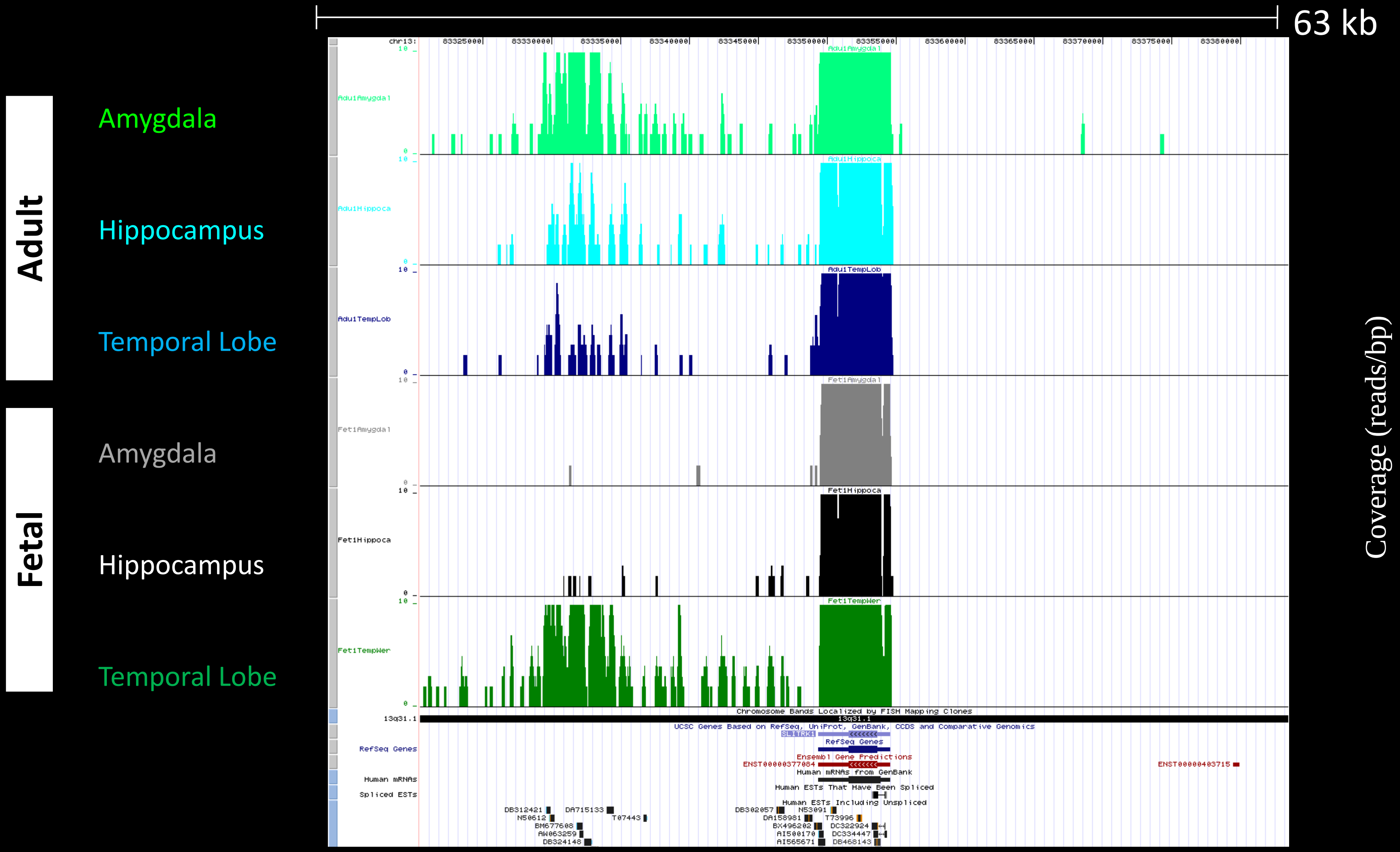
Hippocampus

Temporal Lobe



Coverage (reads/bp)

Find Longer Isoforms



Find New Genes

Adult

Amygdala

Hippocampus

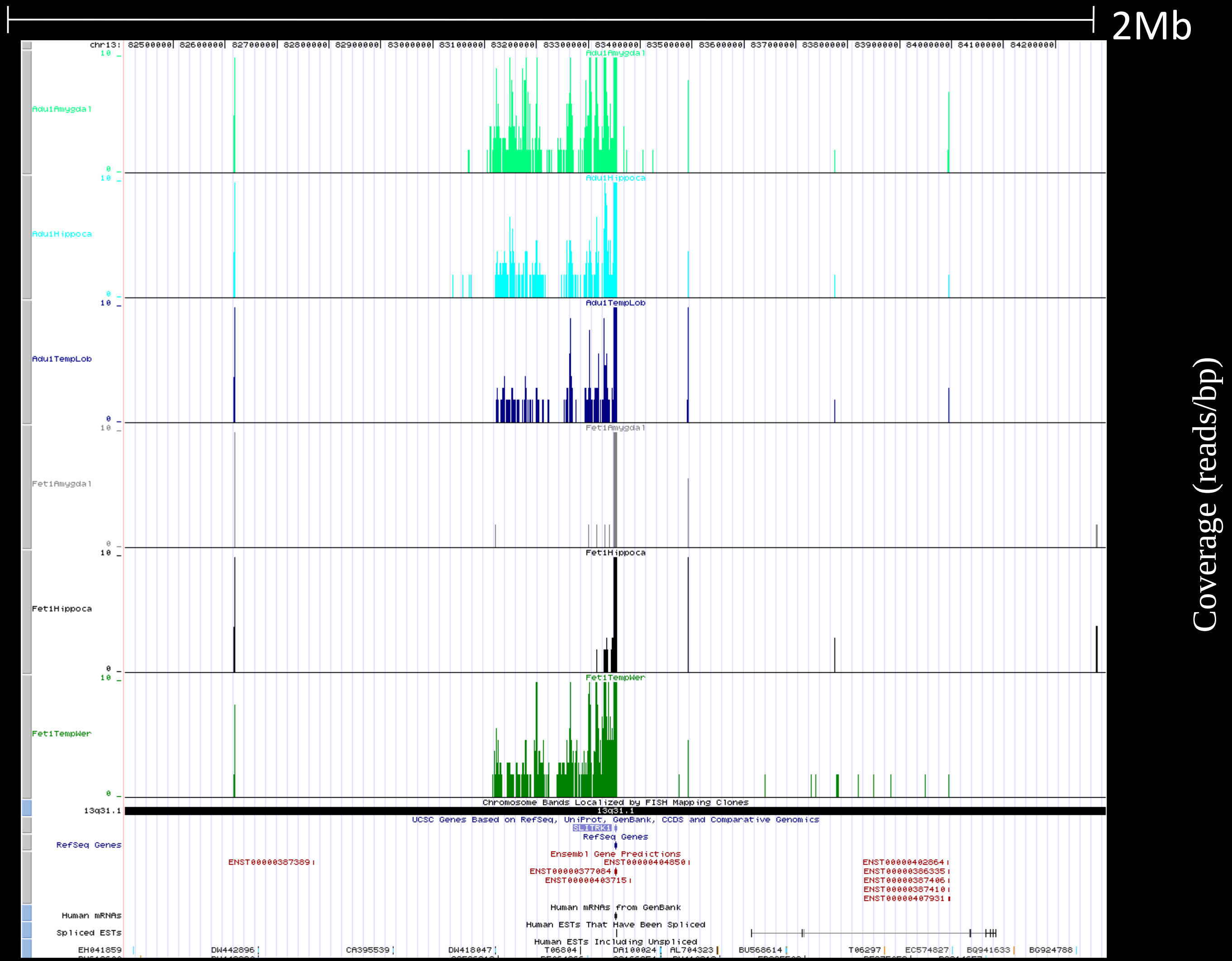
Temporal Lobe

Fetal

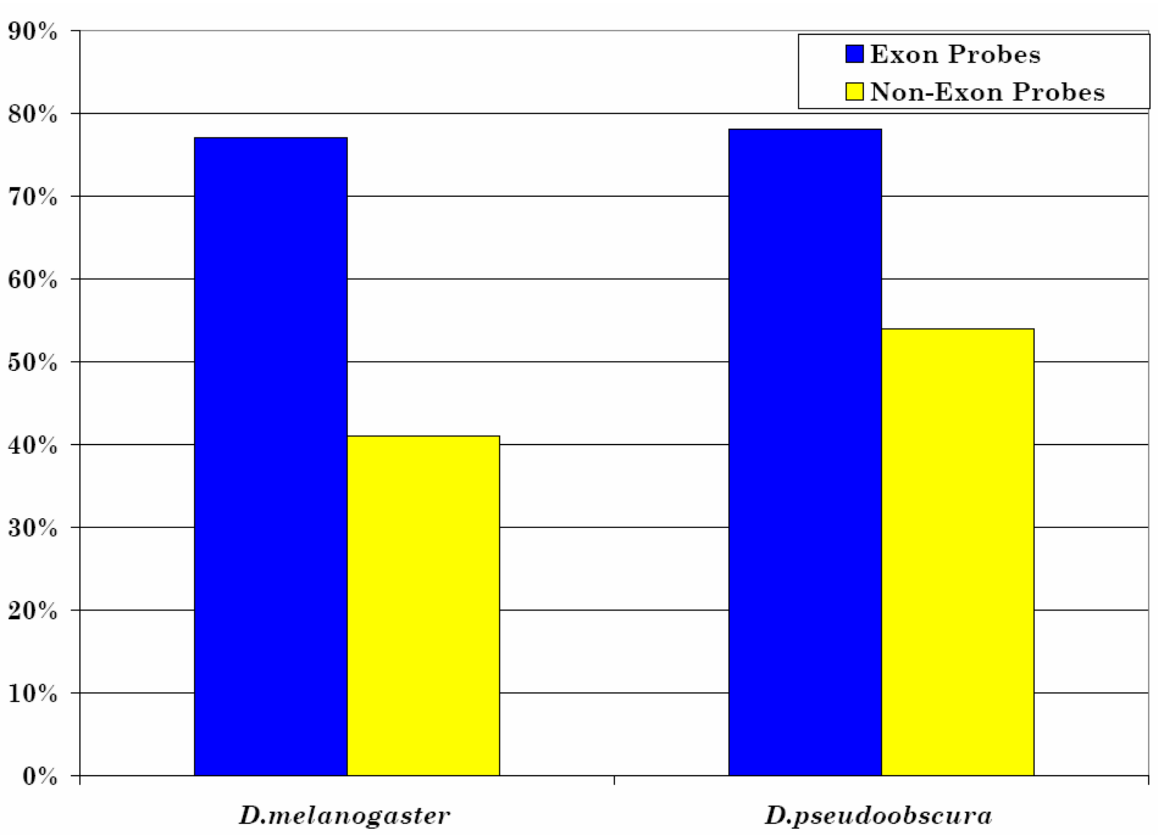
Amygdala

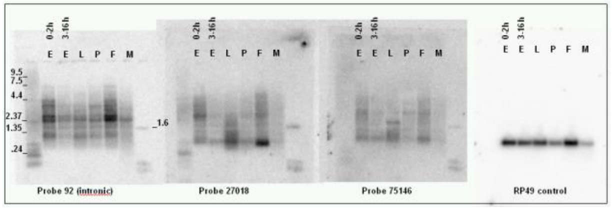
Hippocampus

Temporal Lobe



About Half of the Noncoding Genome is Transcriptionally Active





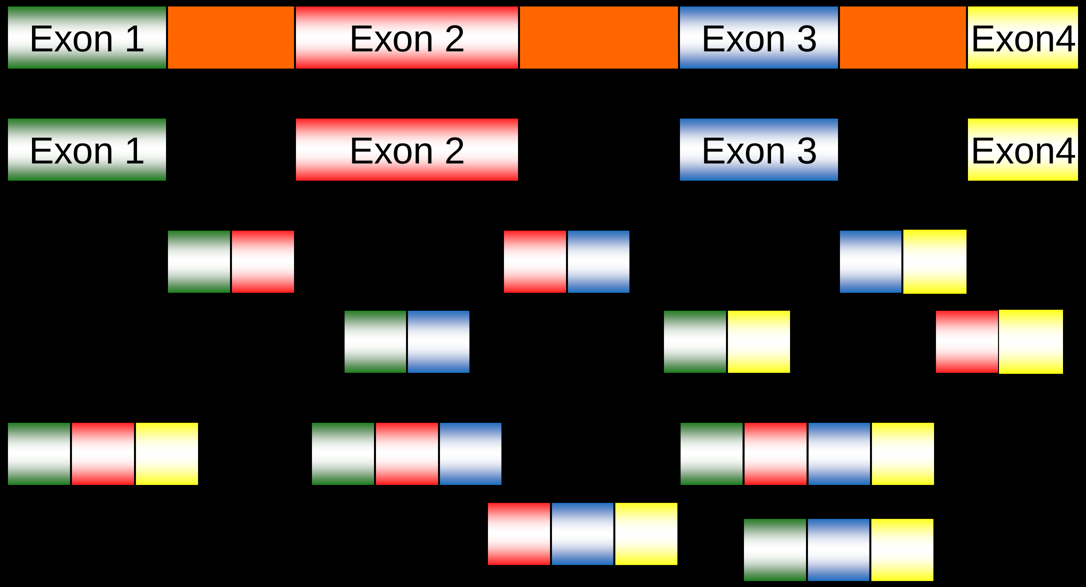
Stolc, Gauhar, Mason et al, *Science*, 2004

Mason, 2006

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

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)

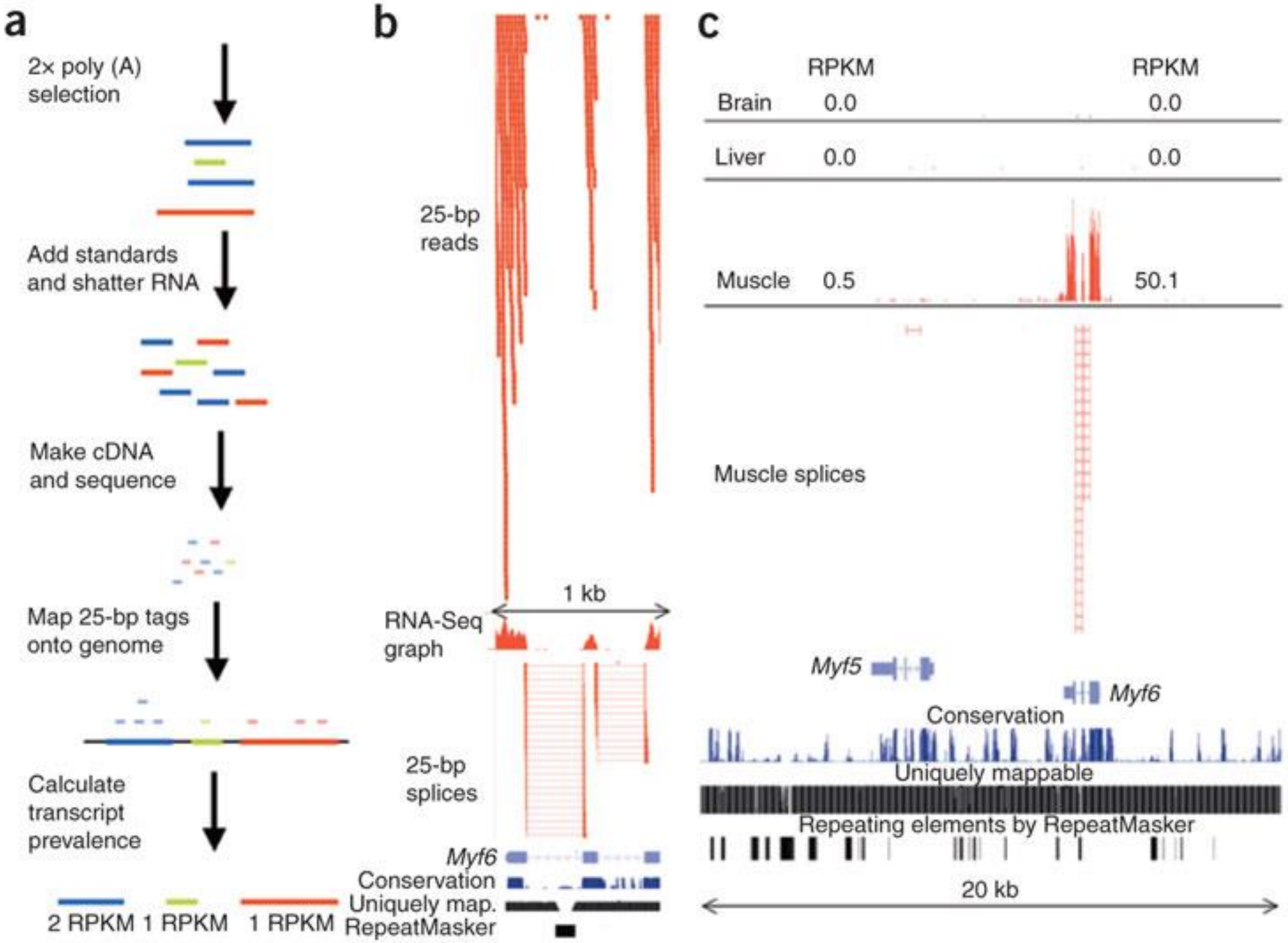
Li and Mason, ARGHG, 2014

(2)

Early Development

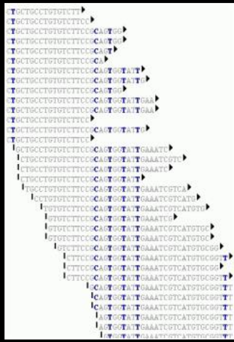
Mapping and quantifying mammalian transcriptomes by RNA-Seq

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



2008

Can RNA-Seq replace microarrays?

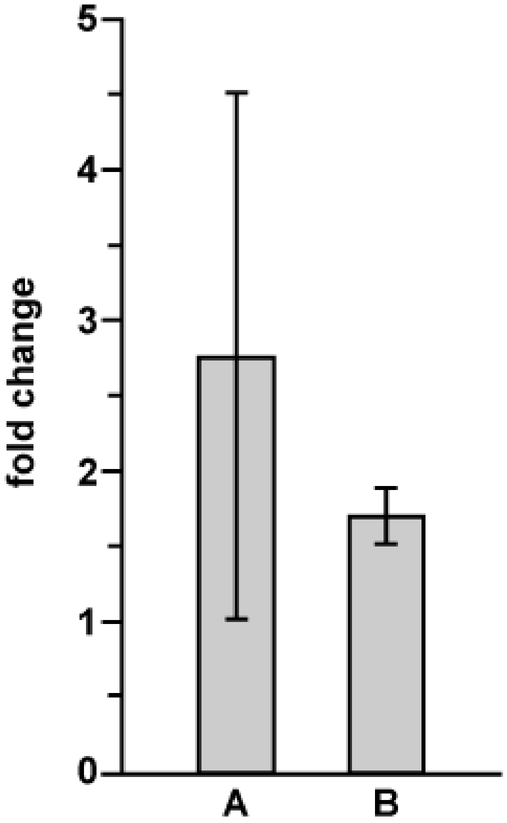


RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays

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

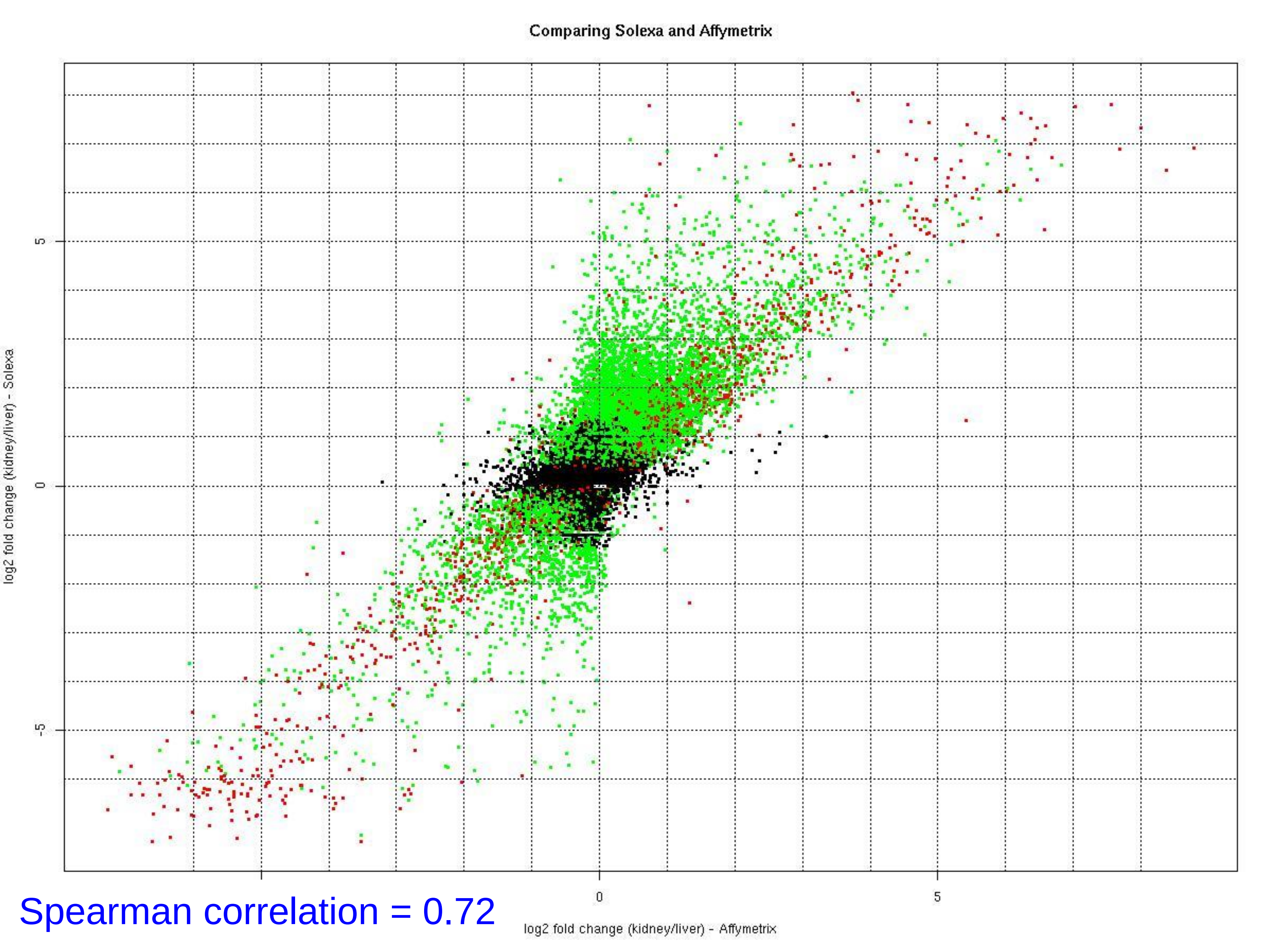
Data Analysis: What genes are differentially expressed?

- 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



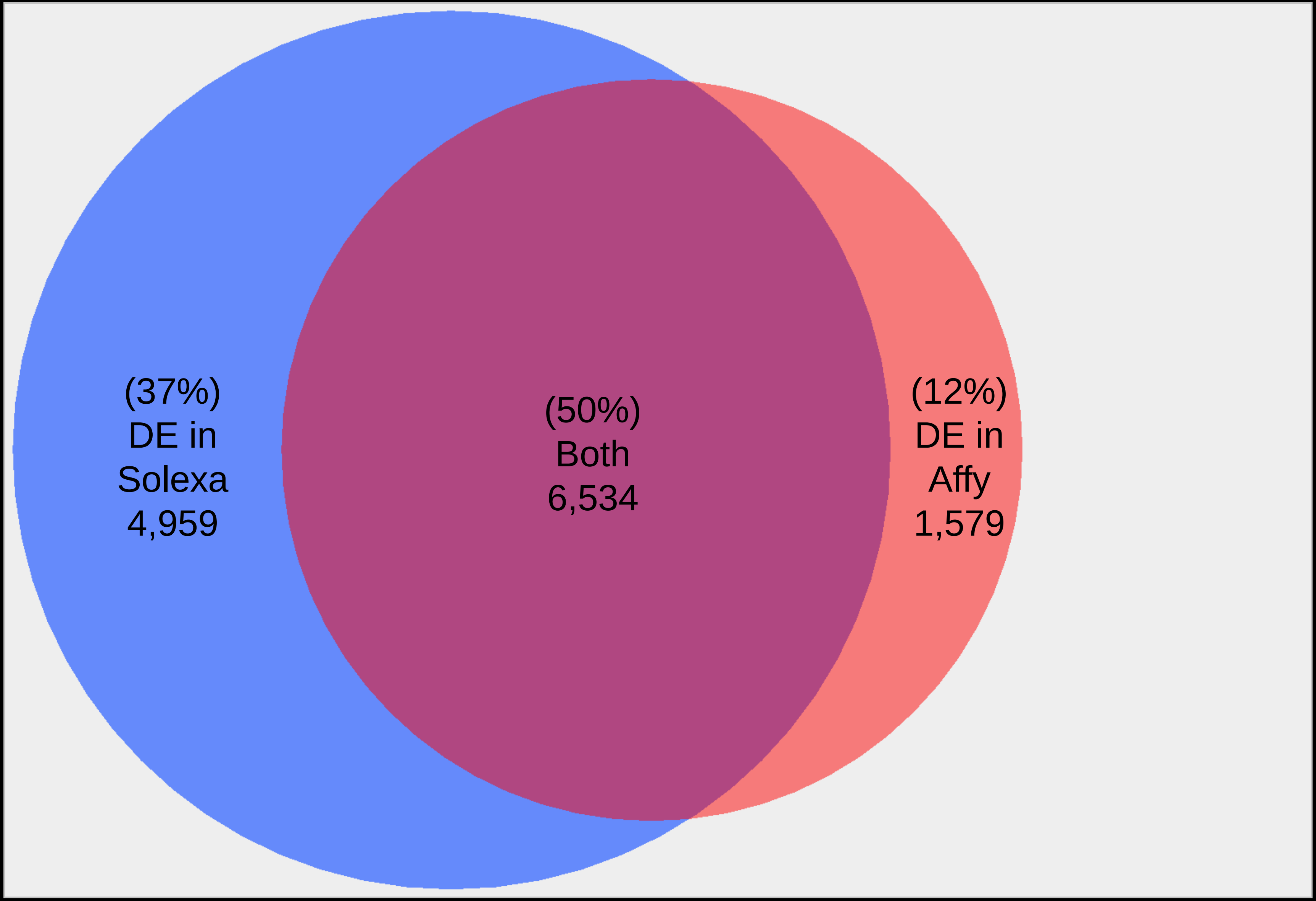
Here, “A” has a fold change >2.5, but varies greatly between replicate experiments. “B” has a fold change of only 1.75, but changes reliably each time the experiment is performed.

Comparing GA and Affy arrays



Spearman correlation = 0.72

13,072 Differentially Expressed (DE) Genes

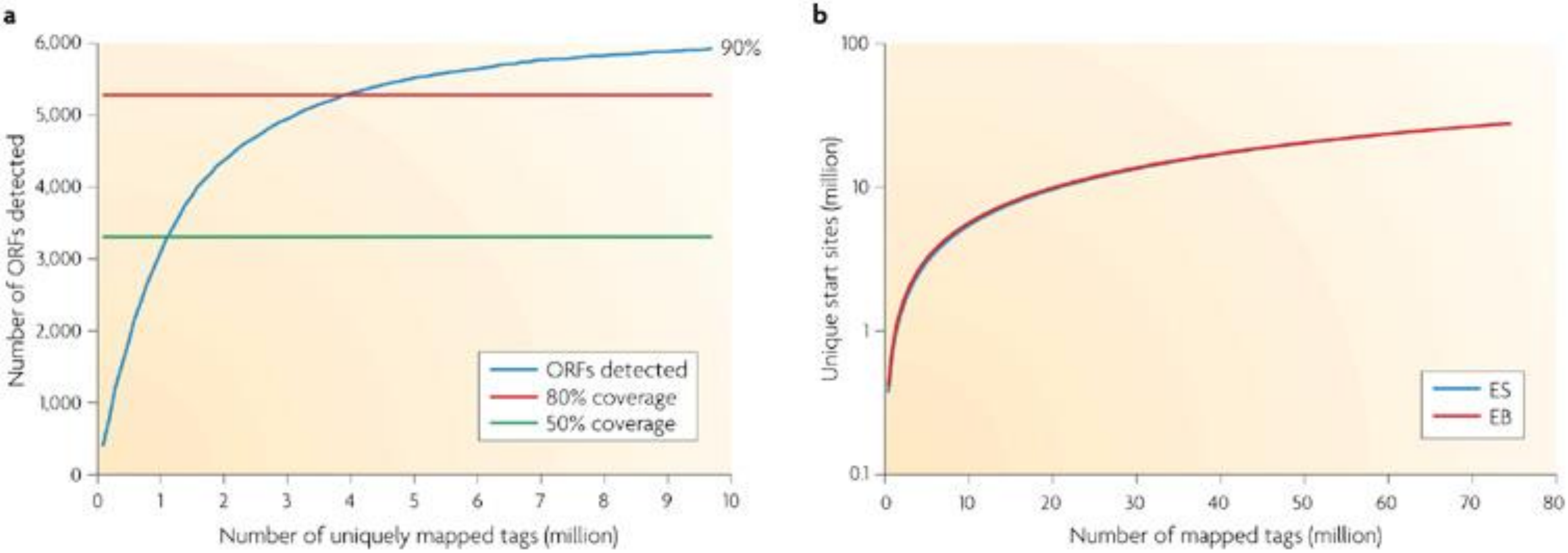


Coverage Requirements: How many lanes/plates/wells?

Depends on:

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

Nature Reviews | Genetics

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)

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

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:

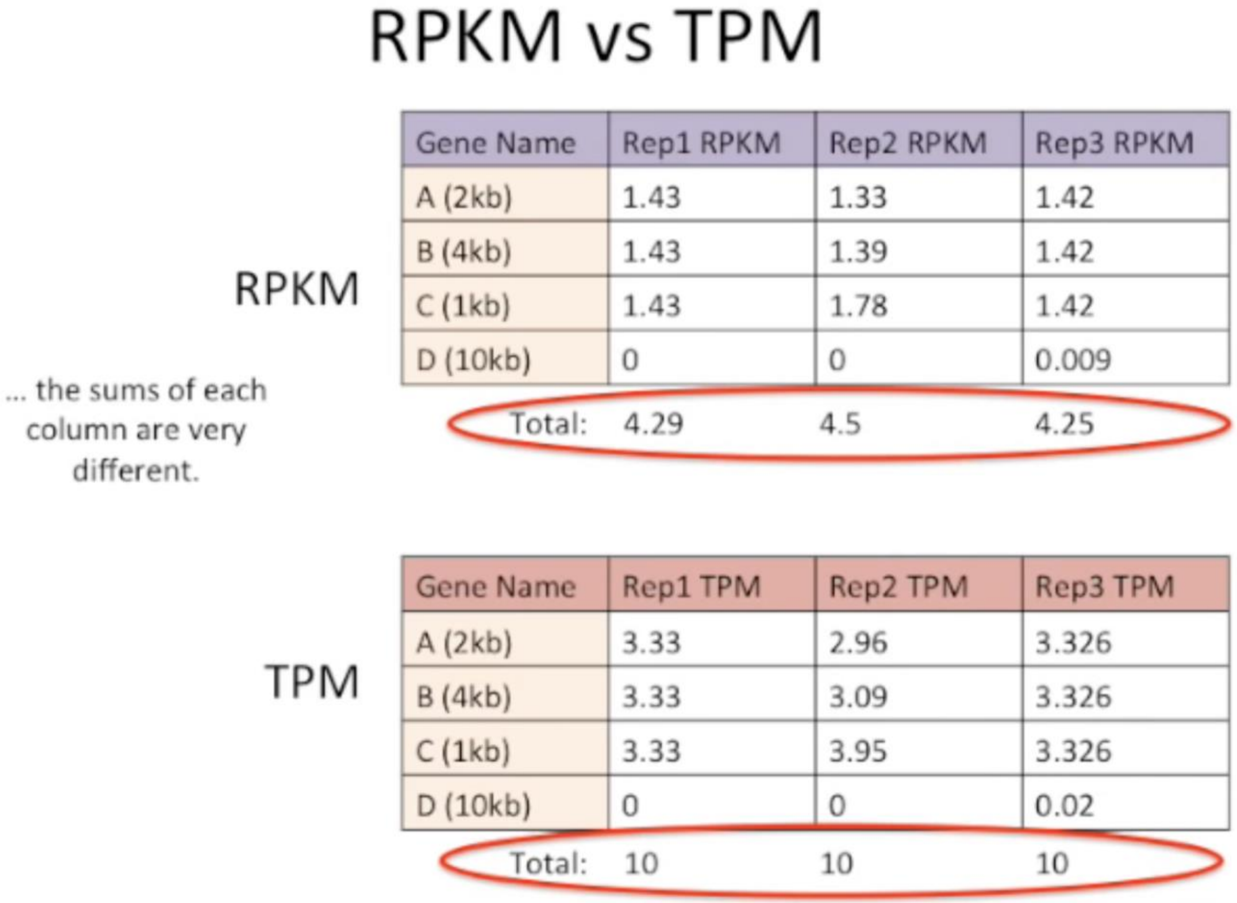
1.Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).

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

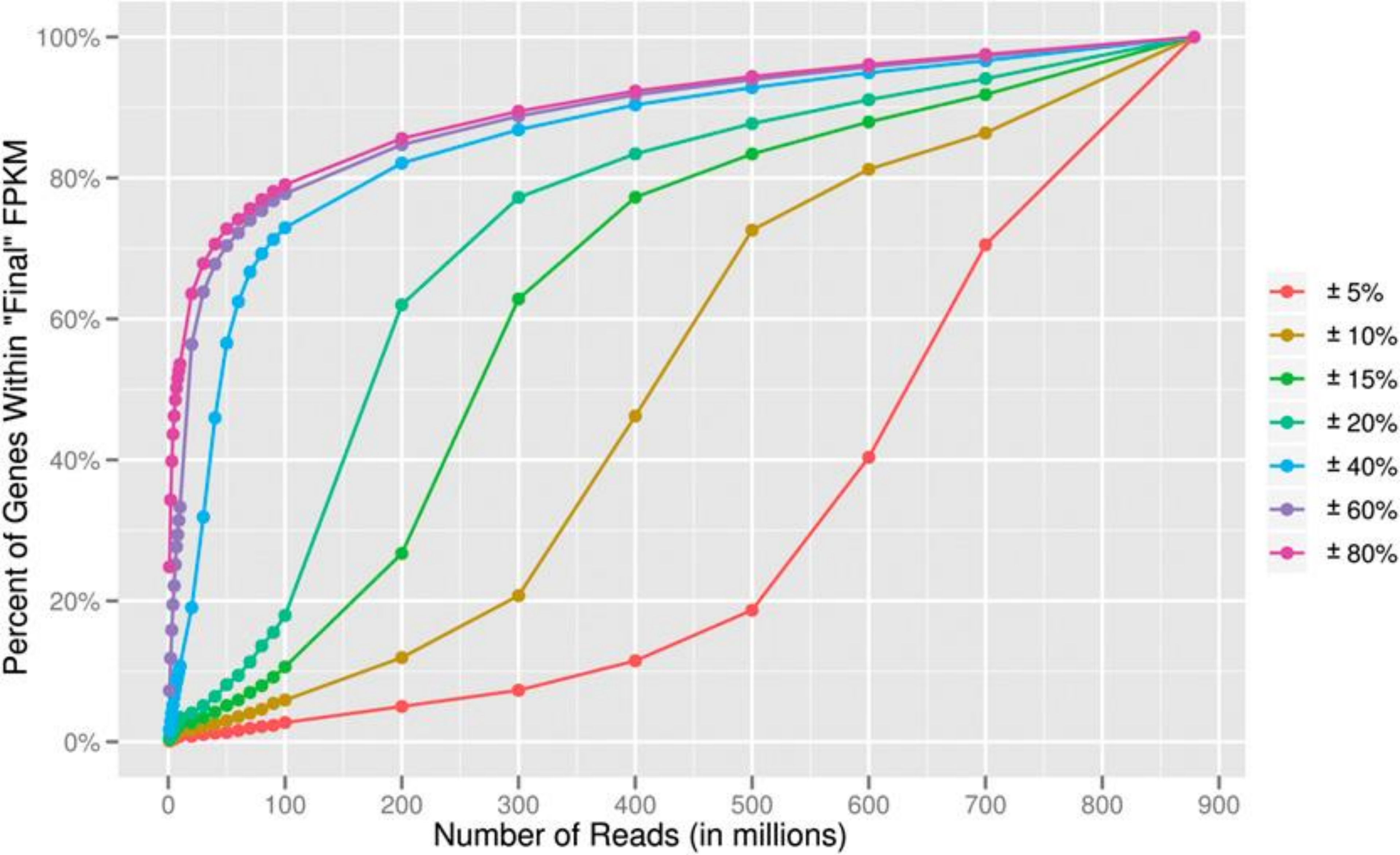
Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

... 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

Accurate gene quantification requires greater depth than gene discovery

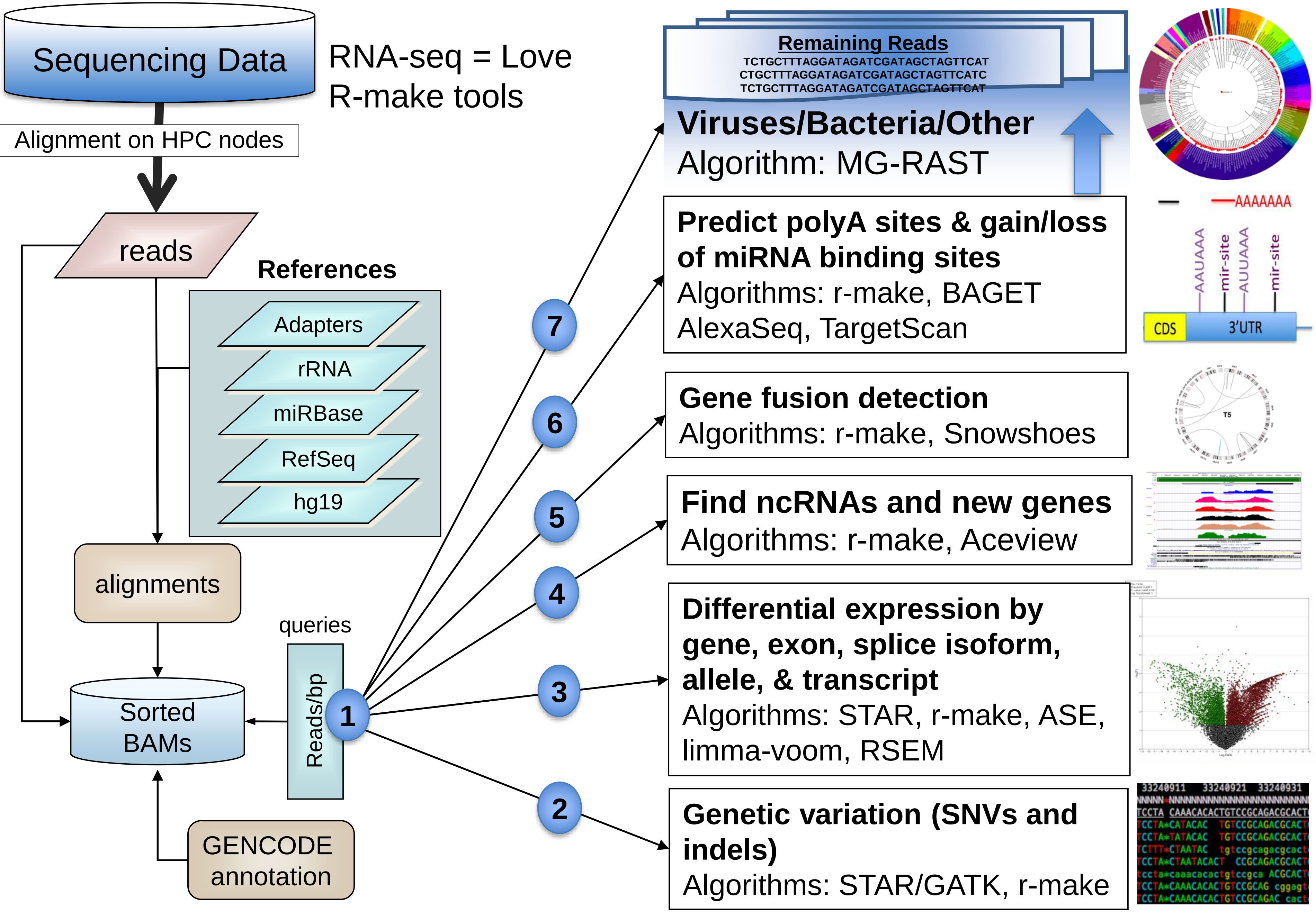


Toung et al, 2011

(3)

Tools & Standards

RNA-Seq and all its flavors create excitement



RNA-seq = Love
R-make tools

Sequencing Data

Alignment on HPC nodes

reads

References

Adapters

rRNA

miRBase

RefSeq

hg19

alignments

Sorted BAMs

GENCODE annotation

queries

Reads/bp

1

2

3

4

5

6

7

Remaining Reads
TCTGCTTTAGGATAGATCGATAGCTAGTTCAT
CTGCTTTAGGATAGATCGATAGCTAGTTCATC
TCTGCTTTAGGATAGATCGATAGCTAGTTCAT

Viruses/Bacteria/Other
Algorithm: MG-RAST

Predict polyA sites & gain/loss of miRNA binding sites
Algorithms: r-make, BAGET AlexaSeq, TargetScan

Gene fusion detection
Algorithms: r-make, Snowshoes

Find ncRNAs and new genes
Algorithms: r-make, Aceview

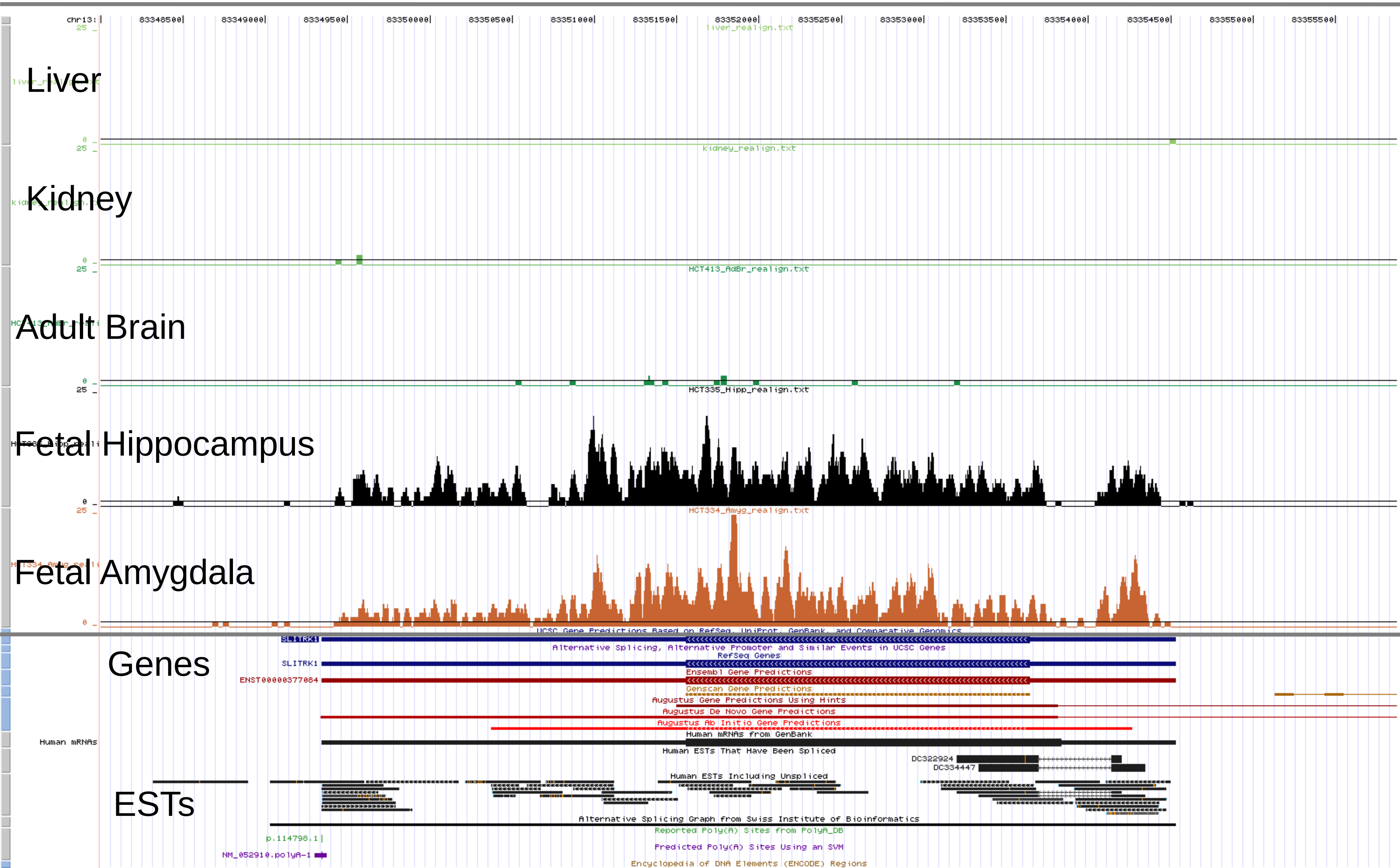
Differential expression by gene, exon, splice isoform, allele, & transcript
Algorithms: STAR, r-make, ASE, limma-voom, RSEM

Genetic variation (SNVs and indels)
Algorithms: STAR/GATK, r-make

But!

There is some noise

What is the source of the wiggles?


chr13:
83348500
83349000
83349500
83350000
83350500
83351000
83351500
83352000
83352500
83353000
83353500
83354000
83354500
83355000
83355500
liver_realign.txt
25
0
kidney_realign.txt
HCT413_AdBr_realign.txt
HCT335_Hipp_realign.txt
HCT334_Amyg_realign.txt
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
SLITRK1
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
RefSeq Genes
SLITRK1
Ensembl Gene Predictions
ENST00000377084
Genscan Gene Predictions
Augustus Gene Predictions Using Hints
Augustus De Novo Gene Predictions
Augustus Ab Initio Gene Predictions
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
DC322924
DC334447
Human ESTs Including Unspliced
Alternative Splicing Graph from Swiss Institute of Bioinformatics
Reported Poly(A) Sites from PolyA_DB
p.114798.1
Predicted Poly(A) Sites Using an SVM
NM_052910.polyA-1
Encyclopedia of DNA Elements (ENCODE) Regions

Liver

Kidney

Adult Brain

Fetal Hippocampus

Fetal Amygdala

Genes

ESTs

The Dirty Dozen: >= 12 Sources of Technical Noise in RNA-Seq

(1) **RNA integrity:** Sample purity or degradation

(2) **Sample RNA complexity:** polyA RNA, total RNA, miRNA

(3) **cDNA synthesis:** random hexamer vs. polyA-primed

(4) **Library isolation:** Gel excision vs. column

(5) **Technical Errors:** Machine, Site, Lane, Technician, Library Size

(6) **Amplification Cycles or Methods:** NuGen, Tn5, Phi29

(7) **Input amount:** (1, 10, 100, 1000 cells)

(8) **Algorithms:** for alignment and assembly

(9) **Fragment size distribution:** Paired-End, Single-End (adaptors)

(10) **Ligation Efficiency:** Multiplexing/Barcoding and RNA ligases

(11) **Depth of Sequencing:** cost/benefit point

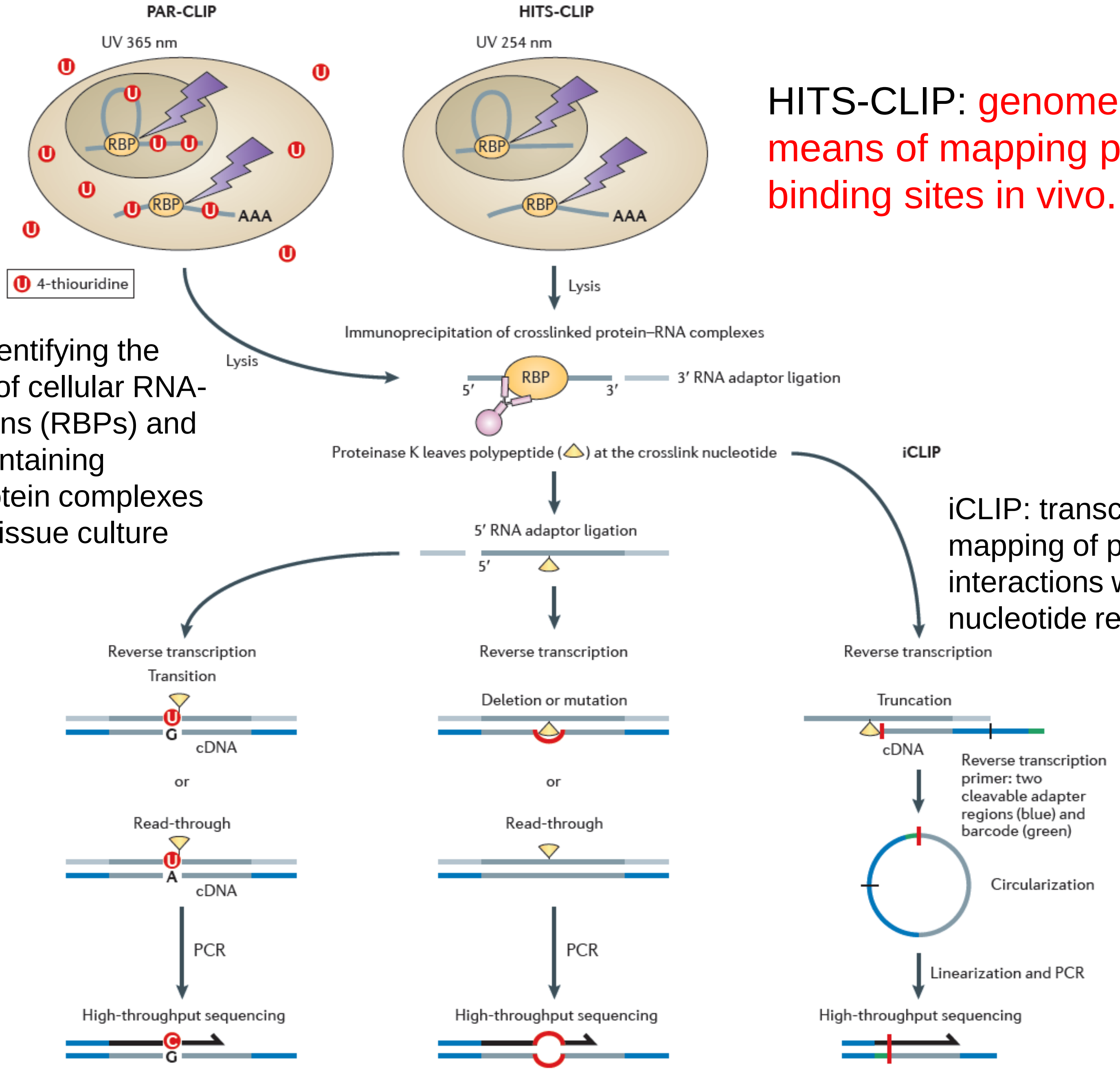
(12) **RNA fragmentation:** cation, enzymatic

Comparison of HITS-CLIP and its latest variants, PAR-CLIP and iCLIP

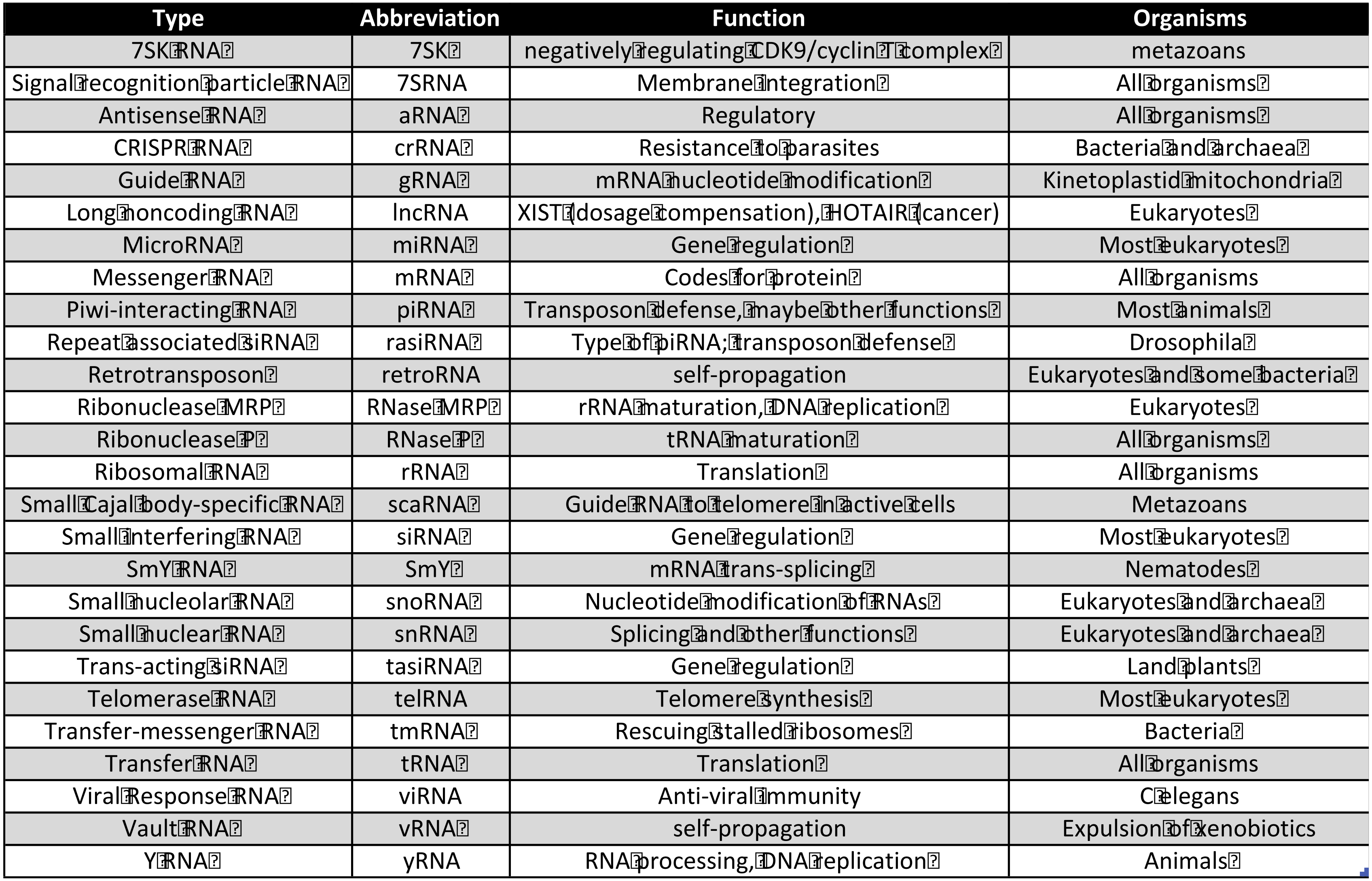
HITS-CLIP: genome-wide means of mapping protein–RNA binding sites in vivo.

PAR-CLIP: identifying the binding sites of cellular RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) in tissue culture cells.

iCLIP: transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution.



Which type of RNA?



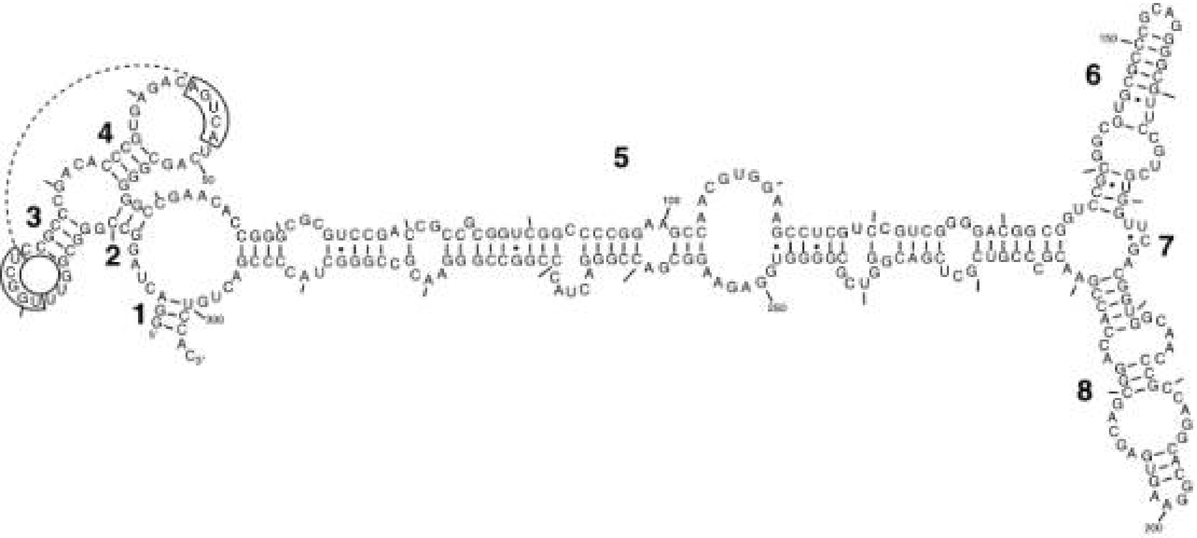
Type	Abbreviation	Function	Organisms
7SK RNA	7SK	negatively regulating CDK9/cyclin T complex	metazoans
Signal recognition particle RNA	7SRNA	Membrane integration	All organisms
Antisense RNA	aRNA	Regulatory	All organisms
CRISPR RNA	crRNA	Resistance to parasites	Bacteria and archaea
Guide RNA	gRNA	mRNA nucleotide modification	Kinetoplastid mitochondria
Long noncoding RNA	lncRNA	XIST (dosage compensation), HOTAIR (cancer)	Eukaryotes
MicroRNA	miRNA	Gene regulation	Most eukaryotes
Messenger RNA	mRNA	Codes for protein	All organisms
Piwi-interacting RNA	piRNA	Transposon defense, maybe other functions	Most animals
Repeat associated siRNA	rasiRNA	Type of piRNA; transposon defense	Drosophila
Retrotransposon	retroRNA	self-propagation	Eukaryotes and some bacteria
Ribonuclease MRP	RNase MRP	rRNA maturation, DNA replication	Eukaryotes
Ribonuclease P	RNase P	tRNA maturation	All organisms
Ribosomal RNA	rRNA	Translation	All organisms
Small Cajal body-specific RNA	scaRNA	Guide RNA to telomere in active cells	Metazoans
Small interfering RNA	siRNA	Gene regulation	Most eukaryotes
SmY RNA	SmY	mRNA trans-splicing	Nematodes
Small nucleolar RNA	snoRNA	Nucleotide modification of RNAs	Eukaryotes and archaea
Small nuclear RNA	snRNA	Splicing and other functions	Eukaryotes and archaea
Trans-acting siRNA	tasiRNA	Gene regulation	Land plants
Telomerase RNA	telRNA	Telomere synthesis	Most eukaryotes
Transfer-messenger RNA	tmRNA	Rescuing stalled ribosomes	Bacteria
Transfer RNA	tRNA	Translation	All organisms
Viral Response RNA	viRNA	Anti-viral immunity	C elegans
Vault RNA	vRNA	self-propagation	Expulsion of xenobiotics
Y RNA	yRNA	RNA processing, DNA replication	Animals

Zumbo and Mason

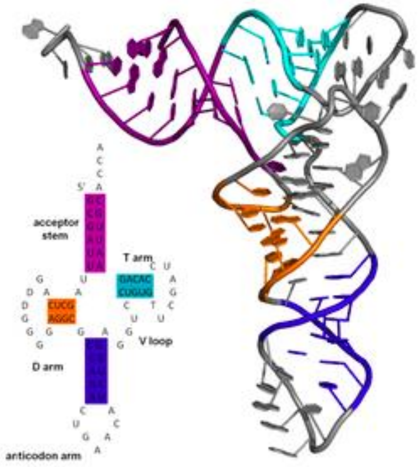
Genome Analysis: Current Procedures and Applications, 2014.

RNAs can have structure/function all their own

- mFOLD/sFOLD
- RNAMotifScan
- RNAfold



Halobacterium halobium SRP RNA
(SRPDB, March 10, 2000)



And - which one do we use?
Technologies Bifurcate into two main realms:

Optical Sequencing

Platform	**Instrument**	**Template Preparation**	**Chemistry**	**Avearge Length**	**Longest Read**
Illumina	HiSeq2500	BridgePCR/cluster	Rev. Term., SBS	100	150
Illumina	HiSeq2000	BridgePCR/cluster	Rev. Term., SBS	100	150
Illumina	MiSeq	BridgePCR/cluster	Rev. Term., SBS	250	300
GnuBio	GnuBio	emPCR	Hyb-Assist Sequencing	1000*	64,000*
Life Technologies	SOLiD 5500	emPCR	Seq. by Lig.	75	100
LaserGen	LaserGen	emPCR	Rev. Term., SBS	25*	100*
Pacific Biosciences	RS	Polymerase Binding	Real-time	1800	15,000
454	Titanium	emPCR	PyroSequencing	650	1100
454	Junior	emPCR	PyroSequencing	400	650
Helicos	Heliscope	none	Rev. Term., SBS	35	57
ZS Genetics	N/A	Atomic Lableing	Electron Microscope	N/A	N/A
Halcyon Molecular	N/A	N/A	Direct Observation of DNA	N/A	N/A

Electrical Sequencing

Platform	**Instrument**	**Template Preparation**	**Chemistry**	**Avearge Length**	**Longest Read**
IBM DNA Transistor	N/A	none	Microchip Nanopore	N/A	N/A
Nabsys	N/A	none	Hyb-Assisted Nanopore (HANS)	N/A	N/A
Life Technologies	PGM	emPCR	Semi-conductor	150	300
Life Technologies	Proton	emPCR	Semi-conductor	300*	500*
Life Technologies	Proton 2	emPCR	Semi-conductor	400*	800*
Oxford Nanopore	MinION	none	Protein Nanopore	1000*	10,000*
Oxford Nanopore	GridION 2K	none	Protein Nanopore	1000*	500,000*
Oxford Nanopore	GridION 8K	none	Protein Nanopore	1000*	500,000*

Nature Biotechnology's Call for Action

Editorial, *Nature Biotechnology*, October 2008

> "… a related endeavor that would help better benchmark the different next-generation sequencing technologies would be to carry out an initiative similar to the Microarray Quality Control [MAQC] consortium where different platforms would be compared against one another as well as against DNA microarrays or quantitative PCR."

There is some hope from at least five places:

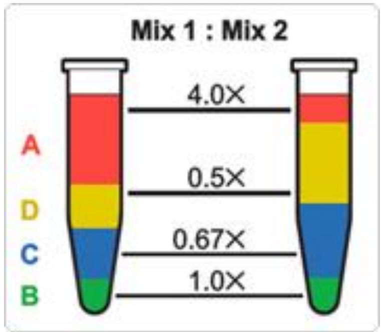
1. **ABRF-NGS Study Consortium**
2. **FDA’s SEQC (MAQC-III) Group**
3. **ENCODE’s RGASP**
4. **RIKEN’s FANTOM**
5. **NIST’s ERCCs**
6. **GEUVADIS Consortium**

But only the first two have data to address technical questions of RNA-Seq

What are ERCCs?

ERCC Spike-In Mixes with synthetic RNAs From Ambion (ERCC=External RNA Control Consortium)

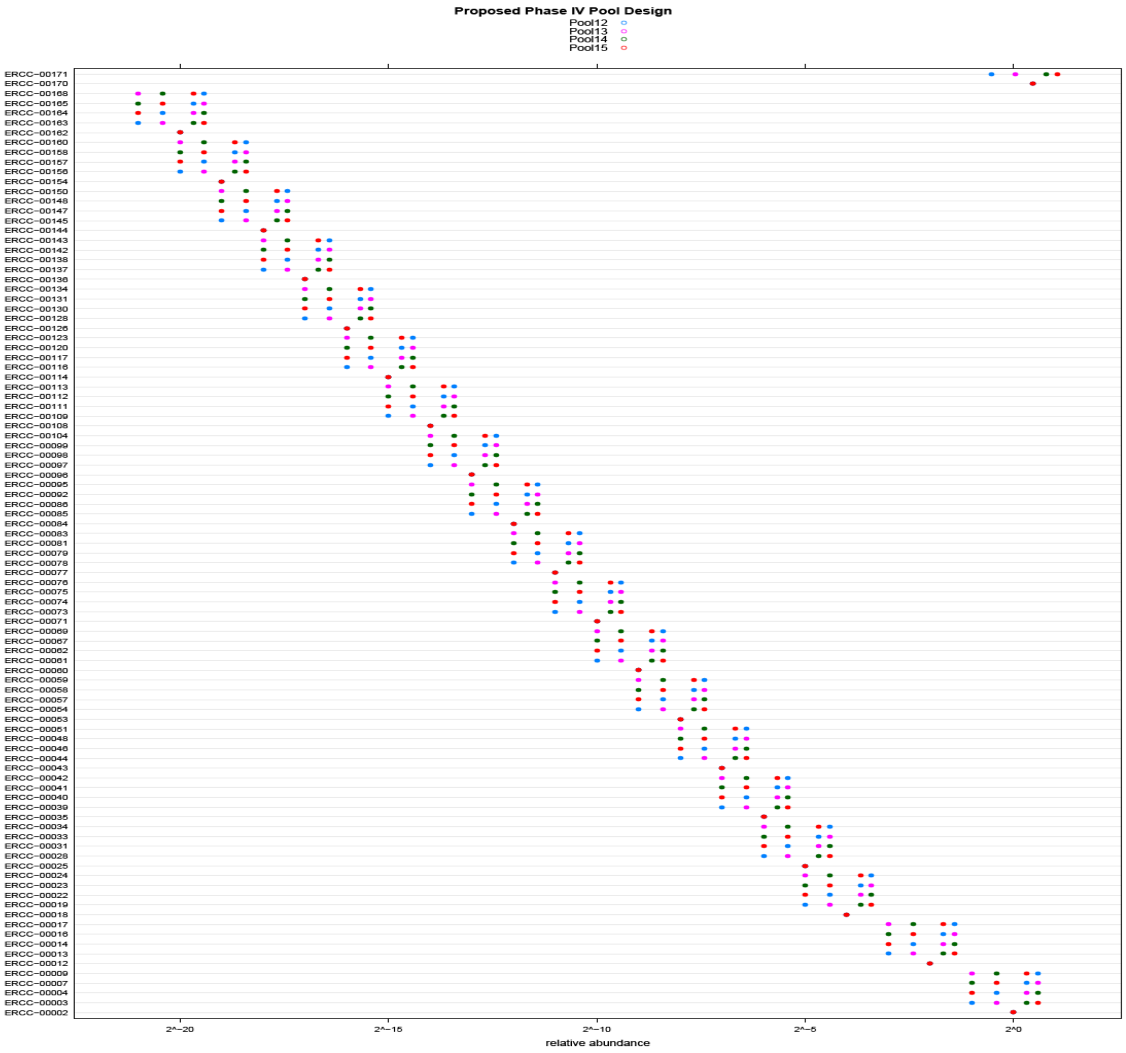
NIST National Institute of Standards and Technology U.S. Department of Commerce



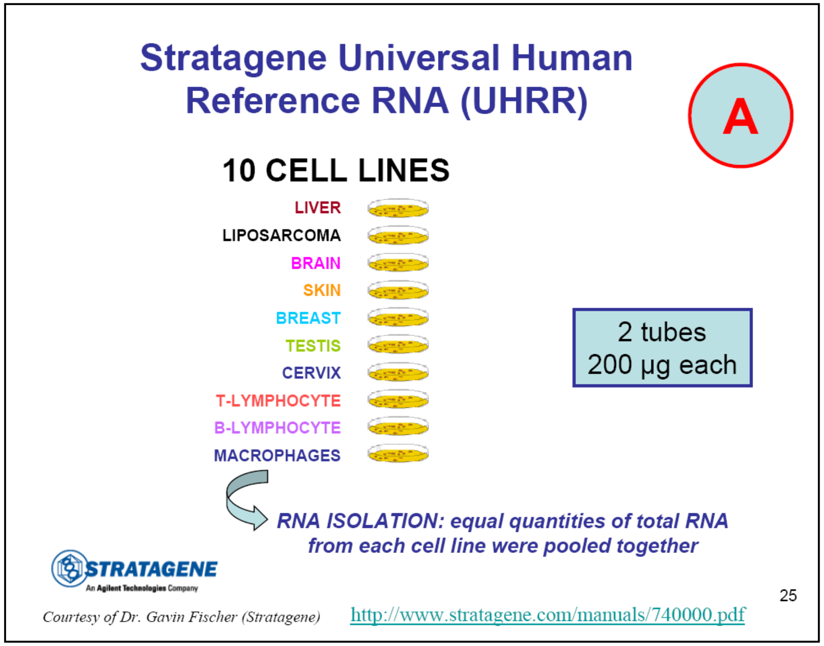
"Ambion® ERCC Spike-In Control Mixes are commercially available, pre-formulated blends of ***92 transcripts,*** *derived and traceable from NIST-certified DNA plasmids. The transcripts are designed to be* ***250 to 2,000 nt*** *in length, which mimic natural eukaryotic mRNAs.*

With two spike-in mix formulations (Spike-In Mix 1 and Spike-In Mix 2), various measurements can be examined to assess different parameters in an experiment or across experiments. Measurements are determined via known molar concentrations for each transcript within a spike-in mix and through association of the two mixes (using ***a combination of ratios across 4 different subgroups of the 92 transcripts****). Furthermore, expression fold-change ratios between two samples can be calculated with a high degree of confidence using the highly concordant relationship between ExFold RNA Spike-In 1 and ExFold RNA Spike-In 2."*

From any species of RNA (left), you can examine it relative to another RNA molecular at a different concentration (x-axis), covering a 2^{20} dynamic range



Samples of the MAQC, SEQC and ABRF-NGS Study



Stratagene Universal Human Reference RNA (UHRR)

A

10 CELL LINES

LIVER
LIPOSARCOMA
BRAIN
SKIN
BREAST
TESTIS
CERVIX
T-LYMPHOCYTE
B-LYMPHOCYTE
MACROPHAGES

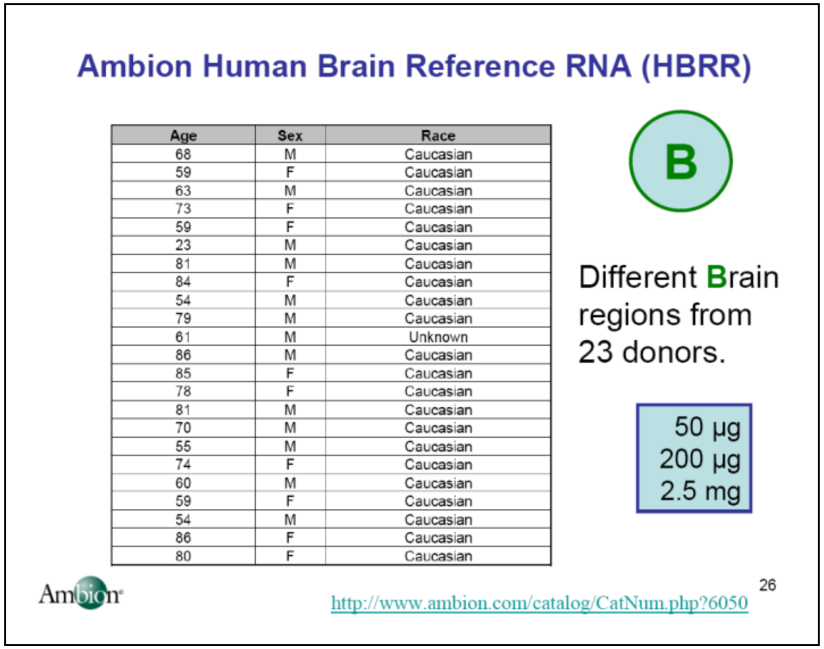
2 tubes
200 µg each

RNA ISOLATION: equal quantities of total RNA from each cell line were pooled together

STRATAGENE
An Agilent Technologies Company

Courtesy of Dr. Gavin Fischer (Stratagene)

http://www.stratagene.com/manuals/740000.pdf



Ambion Human Brain Reference RNA (HBRR)

Age	Sex	Race
68	M	Caucasian
59	F	Caucasian
63	M	Caucasian
73	F	Caucasian
59	F	Caucasian
23	M	Caucasian
81	M	Caucasian
84	F	Caucasian
54	M	Caucasian
79	M	Caucasian
61	M	Unknown
86	M	Caucasian
85	F	Caucasian
78	F	Caucasian
81	M	Caucasian
70	M	Caucasian
55	M	Caucasian
74	F	Caucasian
60	M	Caucasian
59	F	Caucasian
54	M	Caucasian
86	F	Caucasian
80	F	Caucasian

B

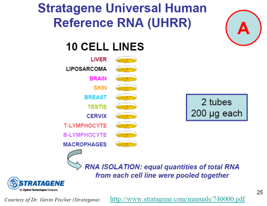
Different **B**rain regions from 23 donors.

50 µg
200 µg
2.5 mg

Ambion

http://www.ambion.com/catalog/CatNum.php?6050

SEQC Samples = MAQC A,B,C,D with ERCC spike-ins



Stratagene Universal Human Reference RNA (UHRR)

A

10 CELL LINES

- LIVER
- LIPOSARCOMA
- BRAIN
- SKIN
- BREAST
- TESTIS
- CERVIX
- T-LYMPHOCYTE
- B-LYMPHOCYTE
- MACROPHAGES

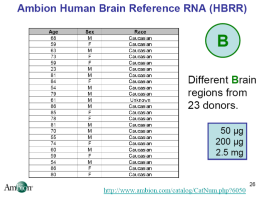
2 tubes
200 µg each

RNA ISOLATION: equal quantities of total RNA from each cell line were pooled together

STRATAGENE
An Agilent Technologies Company

Courtesy of Dr. Gavin Fischer (Stratagene) http://www.stratagene.com/manuals/740000.pdf

25



Ambion Human Brain Reference RNA (HBRR)

Age	Sex	Race
68	M	Caucasian
59	F	Caucasian
63	M	Caucasian
73	F	Caucasian
59	F	Caucasian
23	M	Caucasian
81	M	Caucasian
84	F	Caucasian
54	M	Caucasian
79	M	Caucasian
61	M	Unknown
86	M	Caucasian
85	F	Caucasian
78	F	Caucasian
81	M	Caucasian
70	M	Caucasian
55	M	Caucasian
74	F	Caucasian
60	M	Caucasian
59	F	Caucasian
54	M	Caucasian
86	F	Caucasian
80	F	Caucasian

B

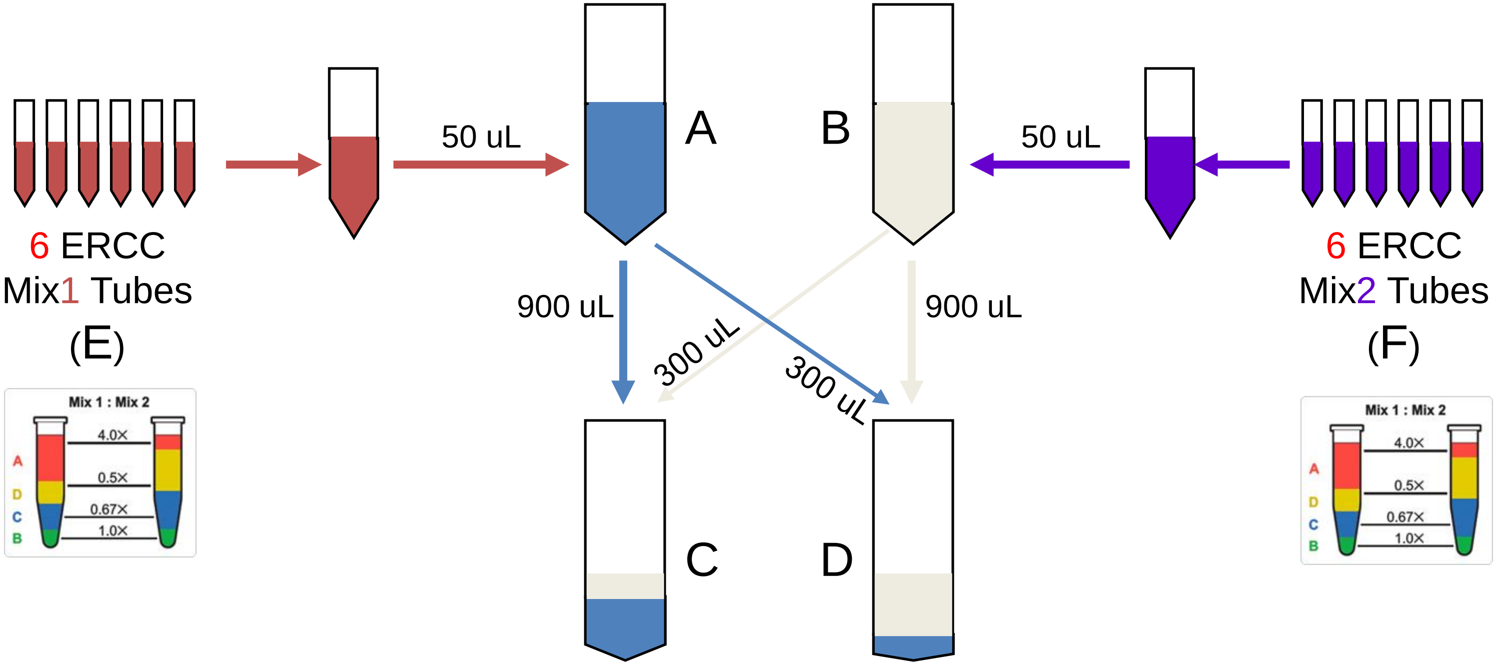
Different Brain regions from 23 donors.

50 µg
200 µg
2.5 mg

Ambion

http://www.ambion.com/catalog/CatNum.php?6050

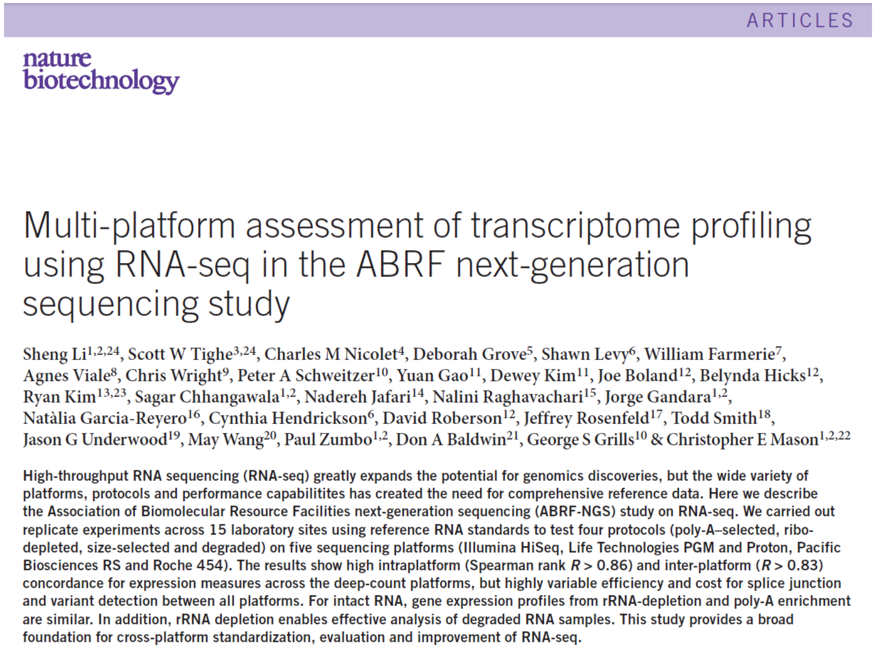
26



ABRF Next Generation Sequencing Study

nature biotechnology
VOLUME 32 NUMBER 9 SEPTEMBER 2014
www.nature.com/naturebiotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY
Focus on RNA sequencing quality control (SEQC)
ABRF evaluation of RNA-seq
Genome editing in hexaploid wheat

Current Results
Phase I: RNA Standards



ARTICLES

nature biotechnology

Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study

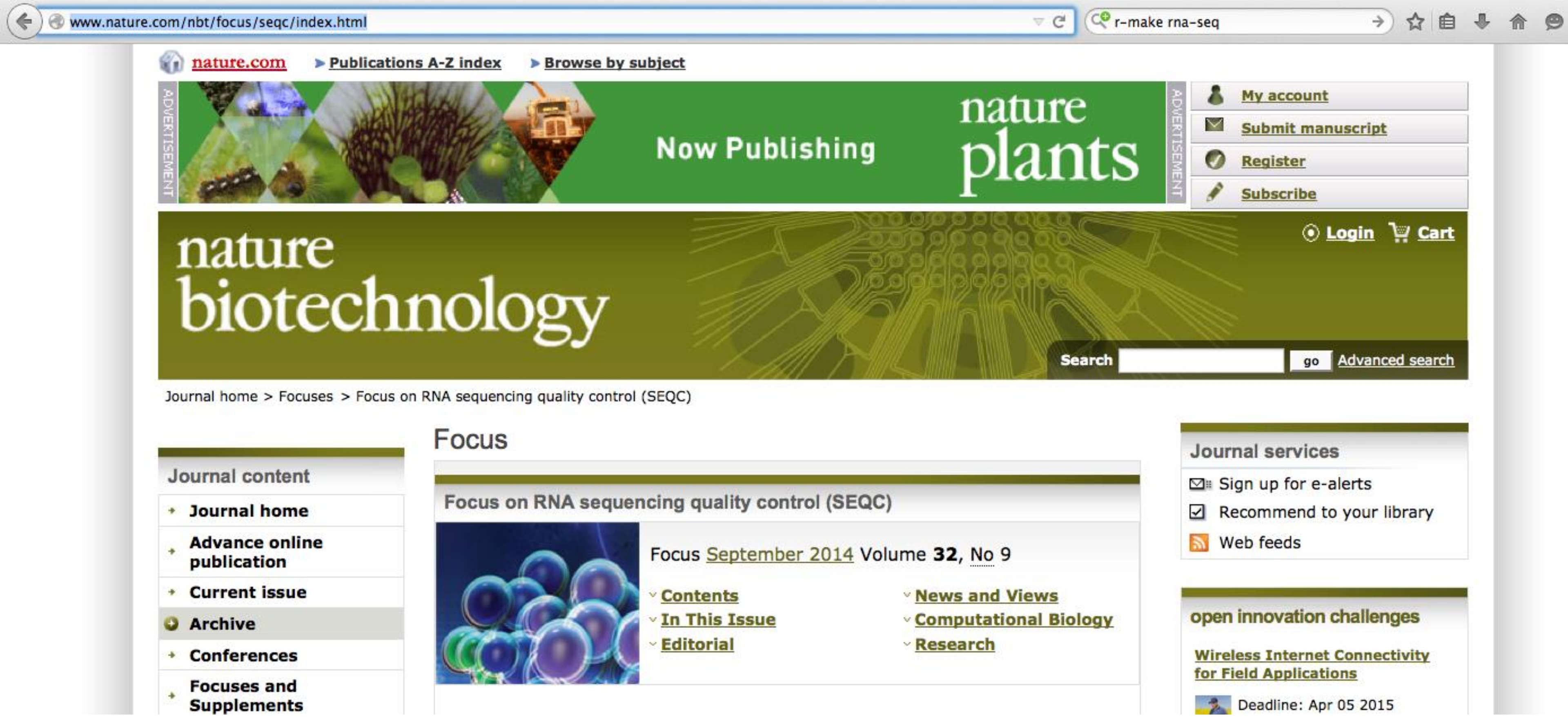
Sheng Li[1,2,24], Scott W Tighe[3,24], Charles M Nicolet[4], Deborah Grove[5], Shawn Levy[6], William Farmerie[7], Agnes Viale[8], Chris Wright[9], Peter A Schweitzer[10], Yuan Gao[11], Dewey Kim[11], Joe Boland[12], Belynda Hicks[12], Ryan Kim[13,23], Sagar Chhangawala[1,2], Nadereh Jafari[14], Nalini Raghavachari[15], Jorge Gandara[1,2], Natàlia Garcia-Reyero[16], Cynthia Hendrickson[6], David Roberson[12], Jeffrey Rosenfeld[17], Todd Smith[18], Jason G Underwood[19], May Wang[20], Paul Zumbo[1,2], Don A Baldwin[21], George S Grills[10] & Christopher E Mason[1,2,22]

High-throughput RNA sequencing (RNA-seq) greatly expands the potential for genomics discoveries, but the wide variety of platforms, protocols and performance capabilitites has created the need for comprehensive reference data. Here we describe the Association of Biomolecular Resource Facilities next-generation sequencing (ABRF-NGS) study on RNA-seq. We carried out replicate experiments across 15 laboratory sites using reference RNA standards to test four protocols (poly-A–selected, ribo-depleted, size-selected and degraded) on five sequencing platforms (Illumina HiSeq, Life Technologies PGM and Proton, Pacific Biosciences RS and Roche 454). The results show high intraplatform (Spearman rank $R > 0.86$) and inter-platform ($R > 0.83$) concordance for expression measures across the deep-count platforms, but highly variable efficiency and cost for splice junction and variant detection between all platforms. For intact RNA, gene expression profiles from rRNA-depletion and poly-A enrichment are similar. In addition, rRNA depletion enables effective analysis of degraded RNA samples. This study provides a broad foundation for cross-platform standardization, evaluation and improvement of RNA-seq.

Li S, et al. *Nat Biotechnol.* 2014 Sep;32(9):915-925.
doi: 10.1038/nbt.2972. Epub 2014 Aug 24. PMID: 25150835

http://www.nature.com/nbt/focus/seqc/index.html

Special issue printed and hosted site



www.nature.com/nbt/focus/seqc/index.html

r-make rna-seq

nature.com ➤ Publications A-Z index ➤ Browse by subject

My account
Submit manuscript
Register
Subscribe

nature biotechnology

Login Cart

Search go Advanced search

Journal home > Focuses > Focus on RNA sequencing quality control (SEQC)

Journal content
- Journal home
- Advance online publication
- Current issue
- Archive
- Conferences
- Focuses and Supplements

Focus

Focus on RNA sequencing quality control (SEQC)

Focus September 2014 Volume **32**, No 9

- Contents
- In This Issue
- Editorial
- News and Views
- Computational Biology
- Research

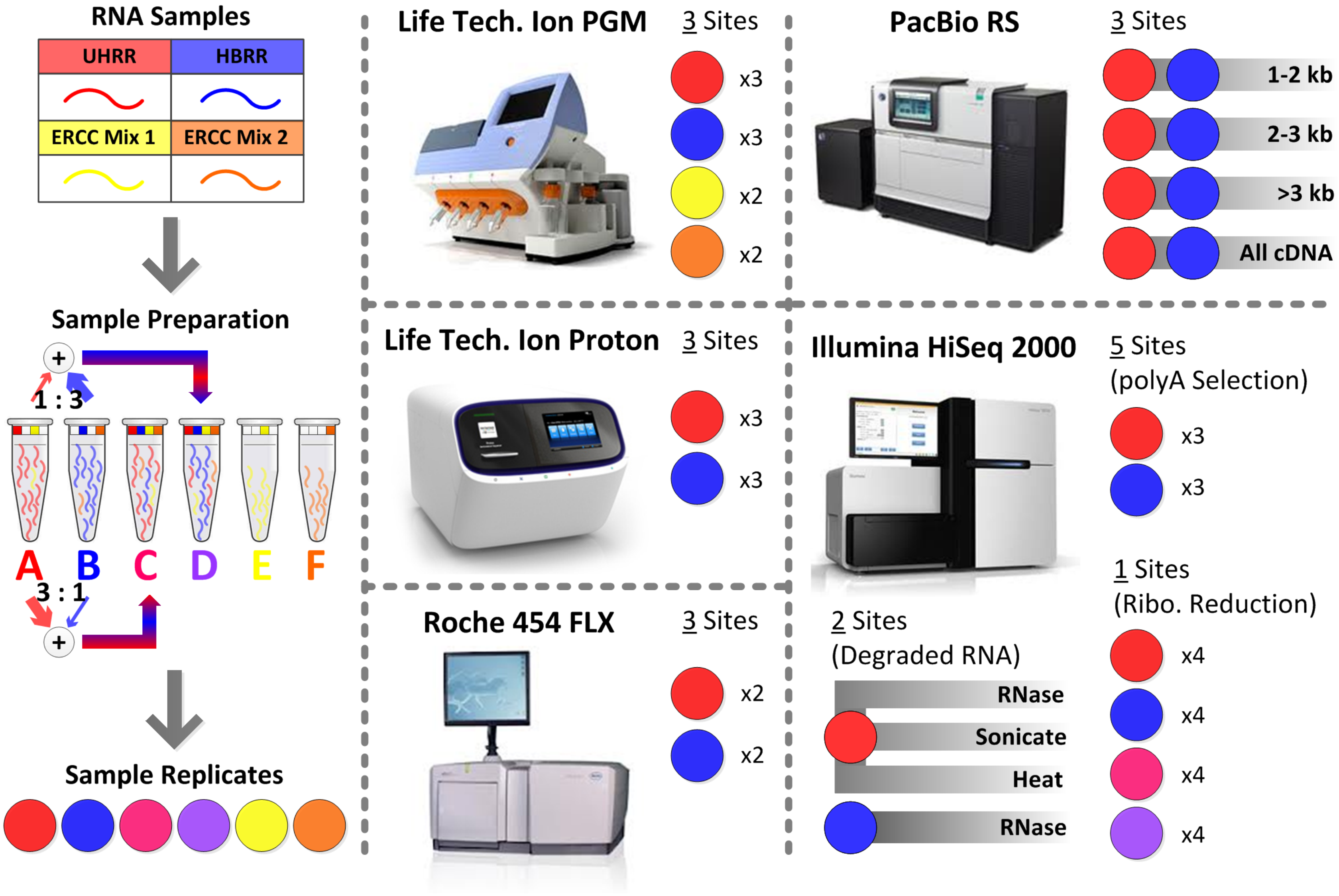
Journal services
- Sign up for e-alerts
- ☑ Recommend to your library
- Web feeds

open innovation challenges

Wireless Internet Connectivity for Field Applications

Deadline: Apr 05 2015

ABRF-NGS study



RNA Samples

UHRR	HBRR
ERCC Mix 1	ERCC Mix 2

Sample Preparation

1 : 3

A B C D E F

3 : 1

Sample Replicates

Life Tech. Ion PGM — 3 Sites

● x3
● x3
● x2
● x2

Life Tech. Ion Proton — 3 Sites

● x3
● x3

Roche 454 FLX — 3 Sites

● x2
● x2

PacBio RS — 3 Sites

1-2 kb
2-3 kb
>3 kb
All cDNA

Illumina HiSeq 2000

5 Sites
(polyA Selection)

● x3
● x3

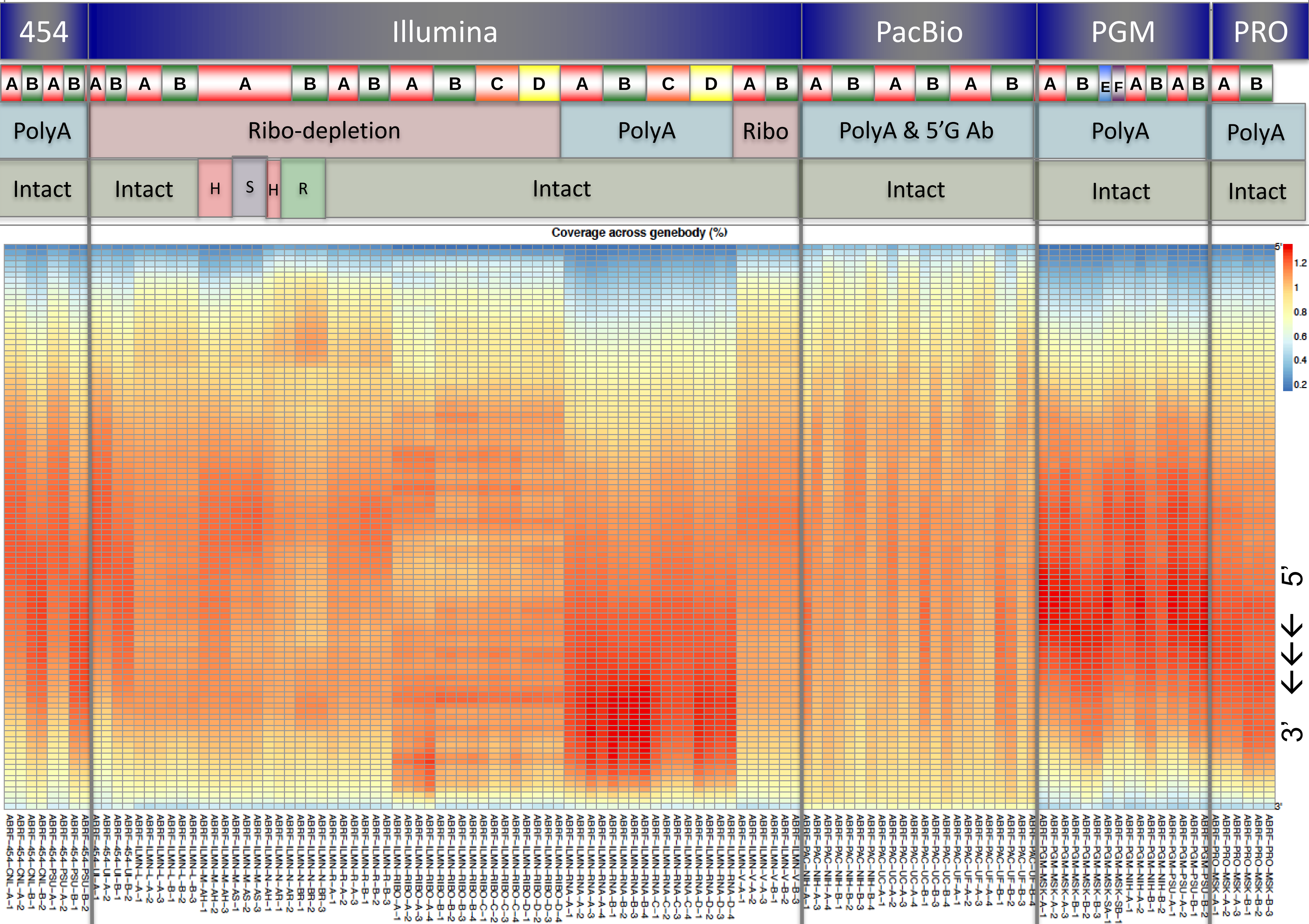
1 Sites
(Ribo. Reduction)

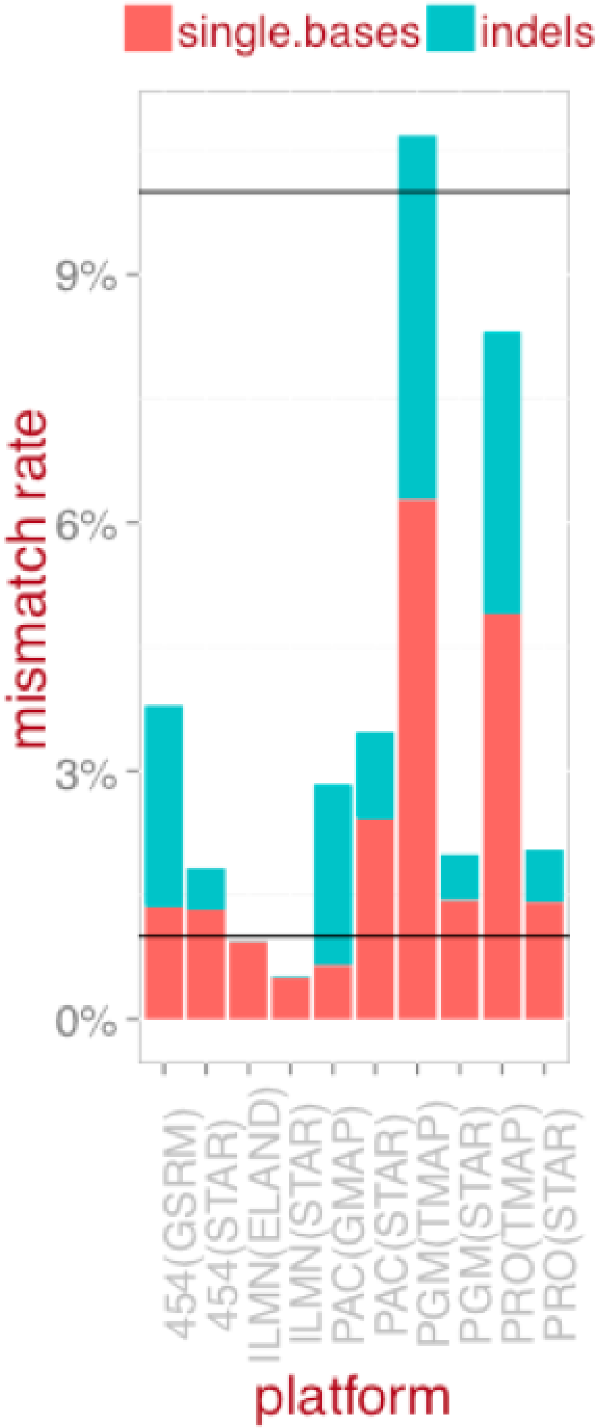
● x4
● x4
● x4
● x4

2 Sites
(Degraded RNA)

RNase
Sonicate
Heat
RNase

Gene coverage distributions reveal platform- and prep-specific effects





Error models highly variable among platforms

454: Roche 454 GS FLX+

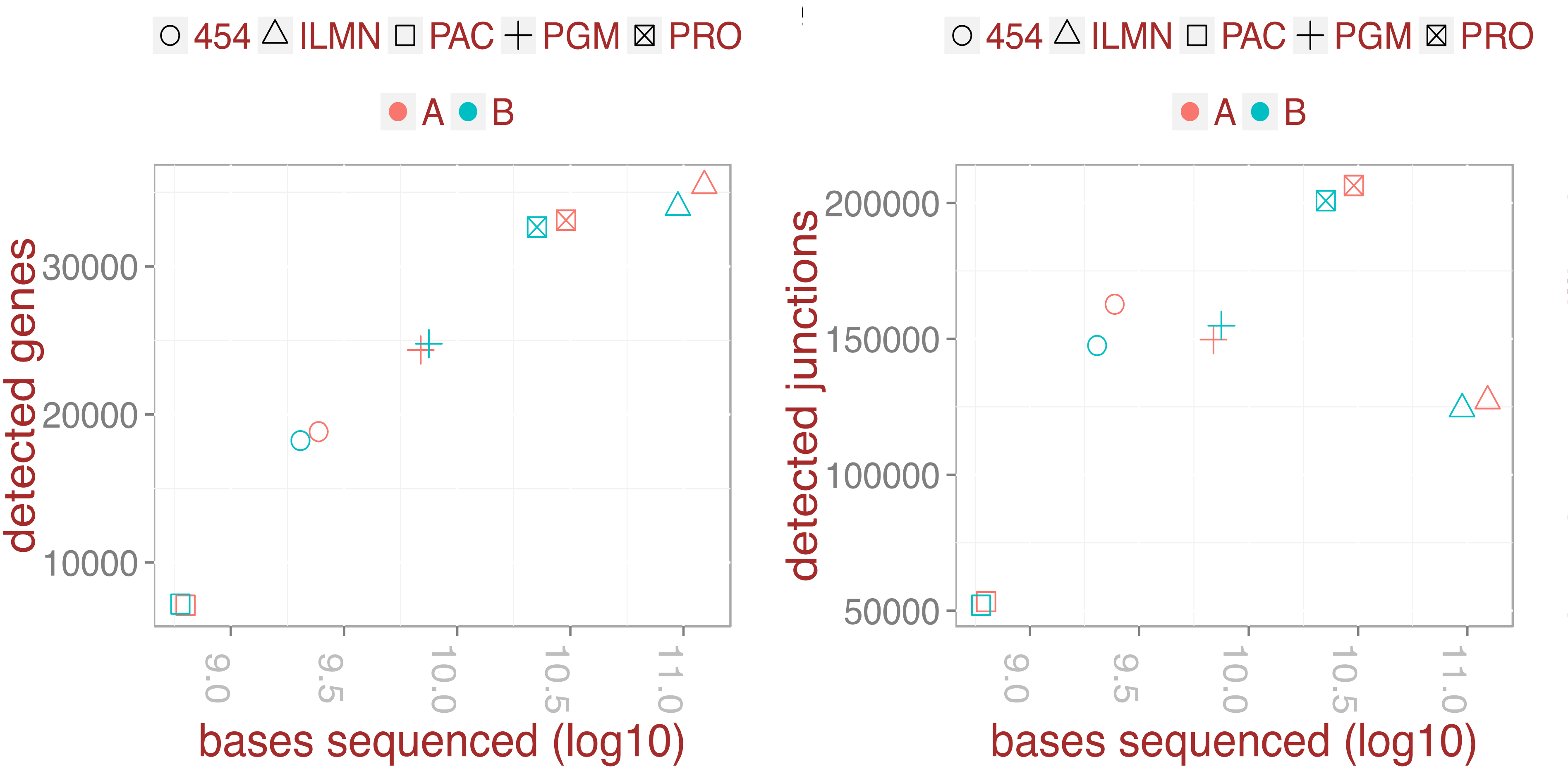
ILMN: Illumina HiSeq 2000/2500

PAC: Pacific Biosciences RS I

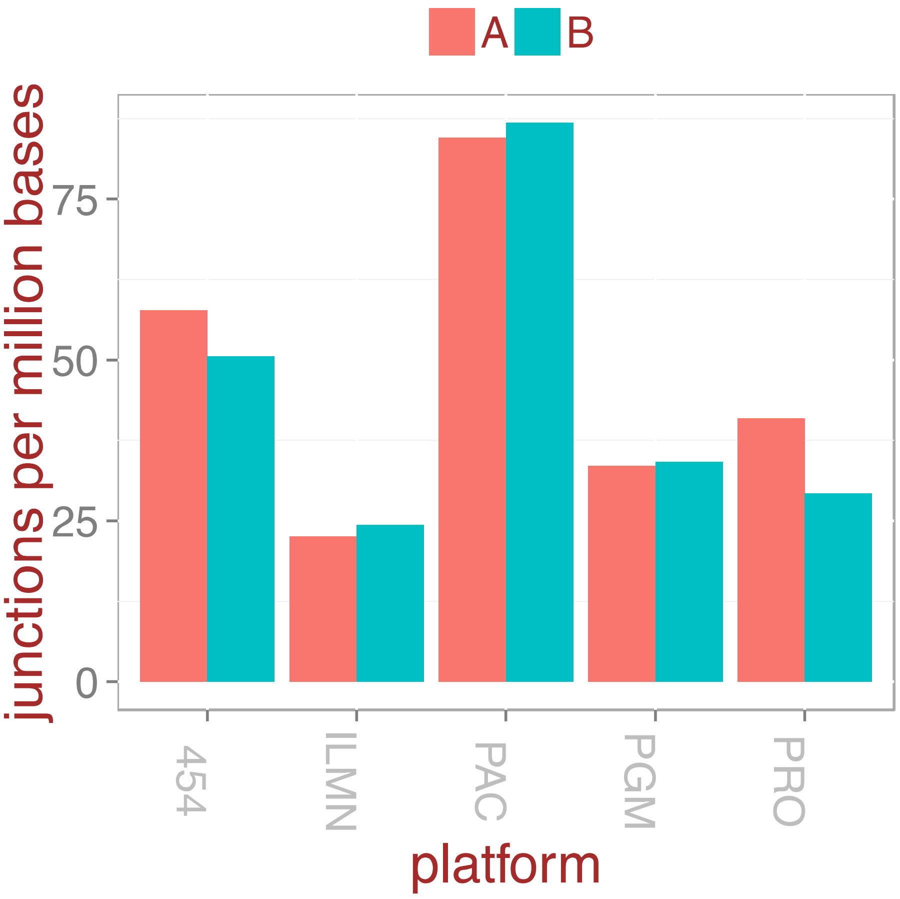
PGM: Ion Personal Genome Machine

PRO: Ion Torrent Proton

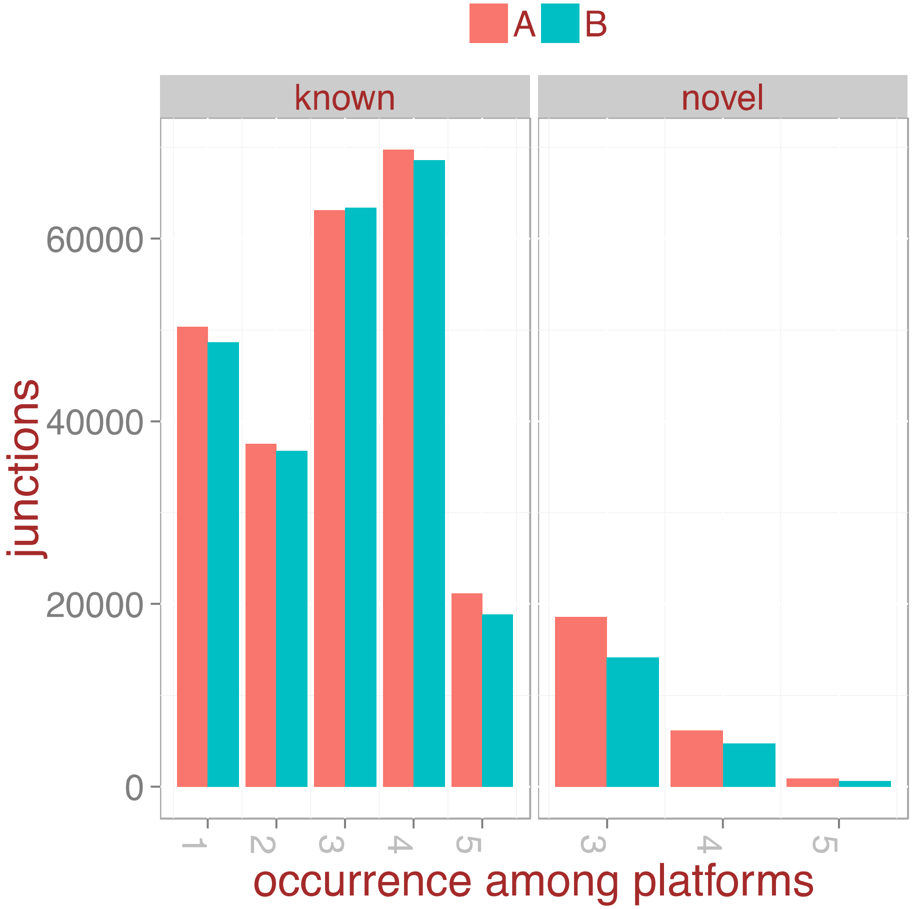
Genes detection is log-linear; Junction detection is length-dependent



Junctions detection efficiency highly variable; agreement common

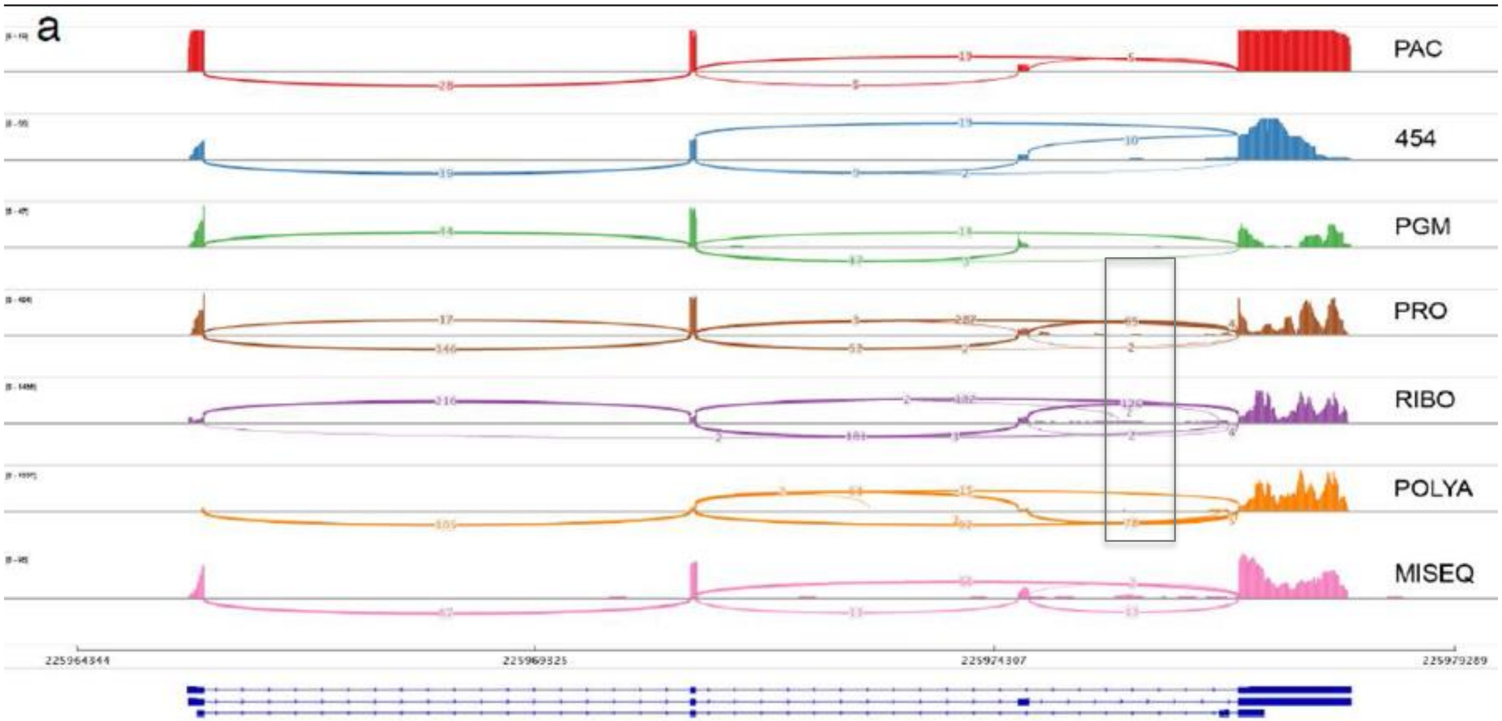


Note: Only use a subset of reads among platforms to normalize the scale

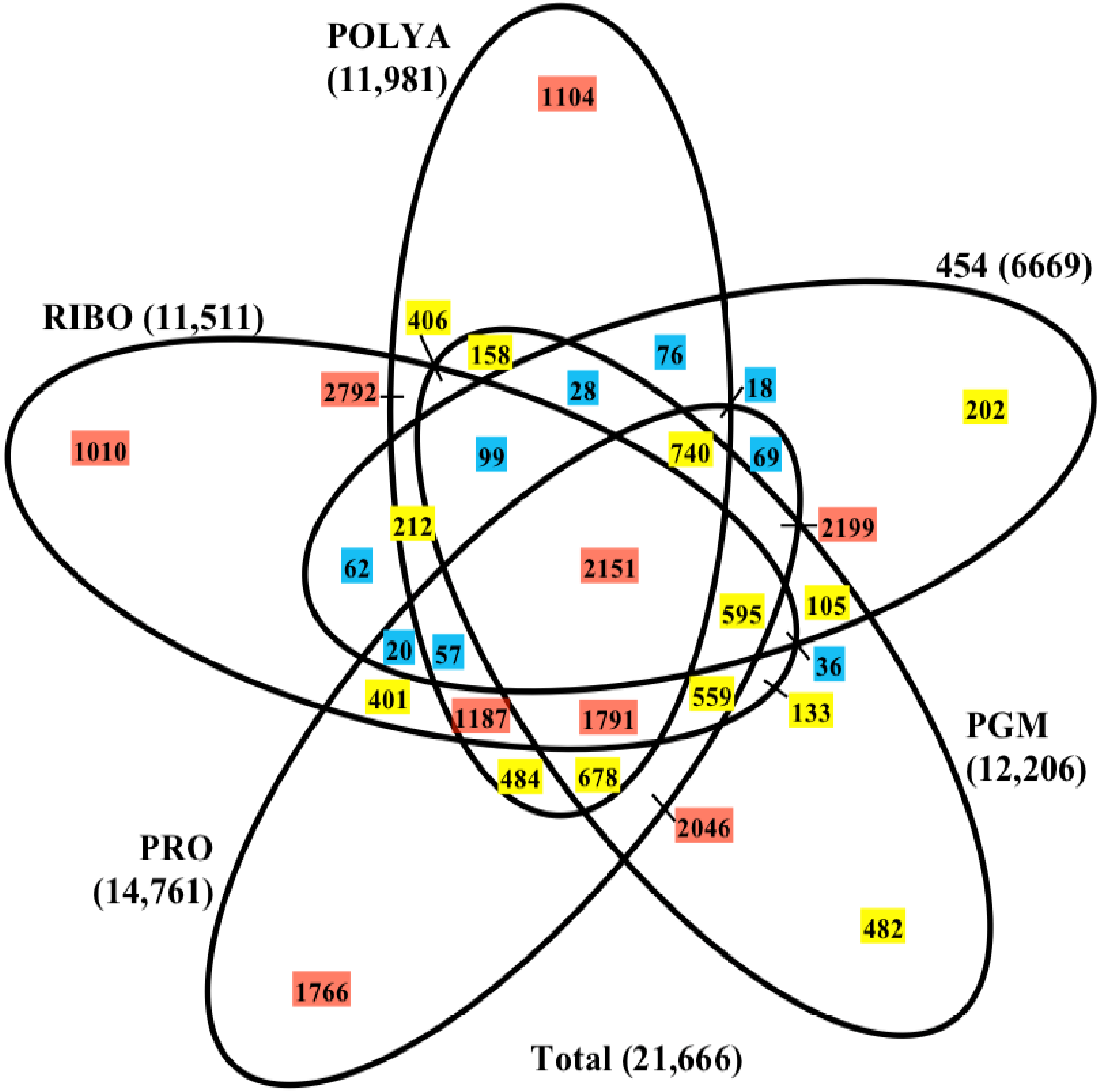


Most of the known junctions are shared by at least 3 platforms

Sequencing depth is important to discover low abundance transcripts



Inter-platform differential gene expression show 88-97% agreement

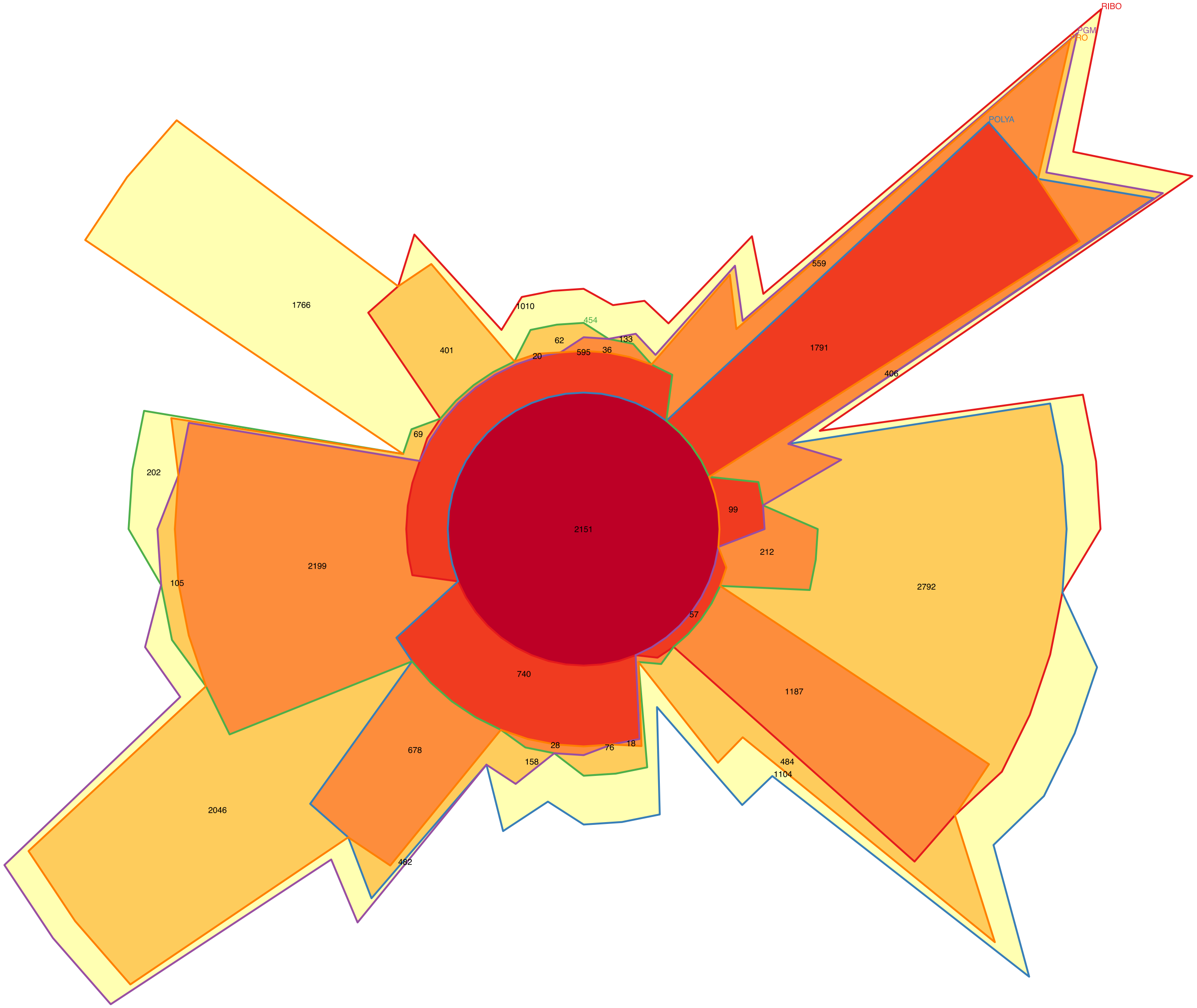


Shared sets of greater than 1000 genes are indicated in red,
100-999 yellow,
<100 blue.

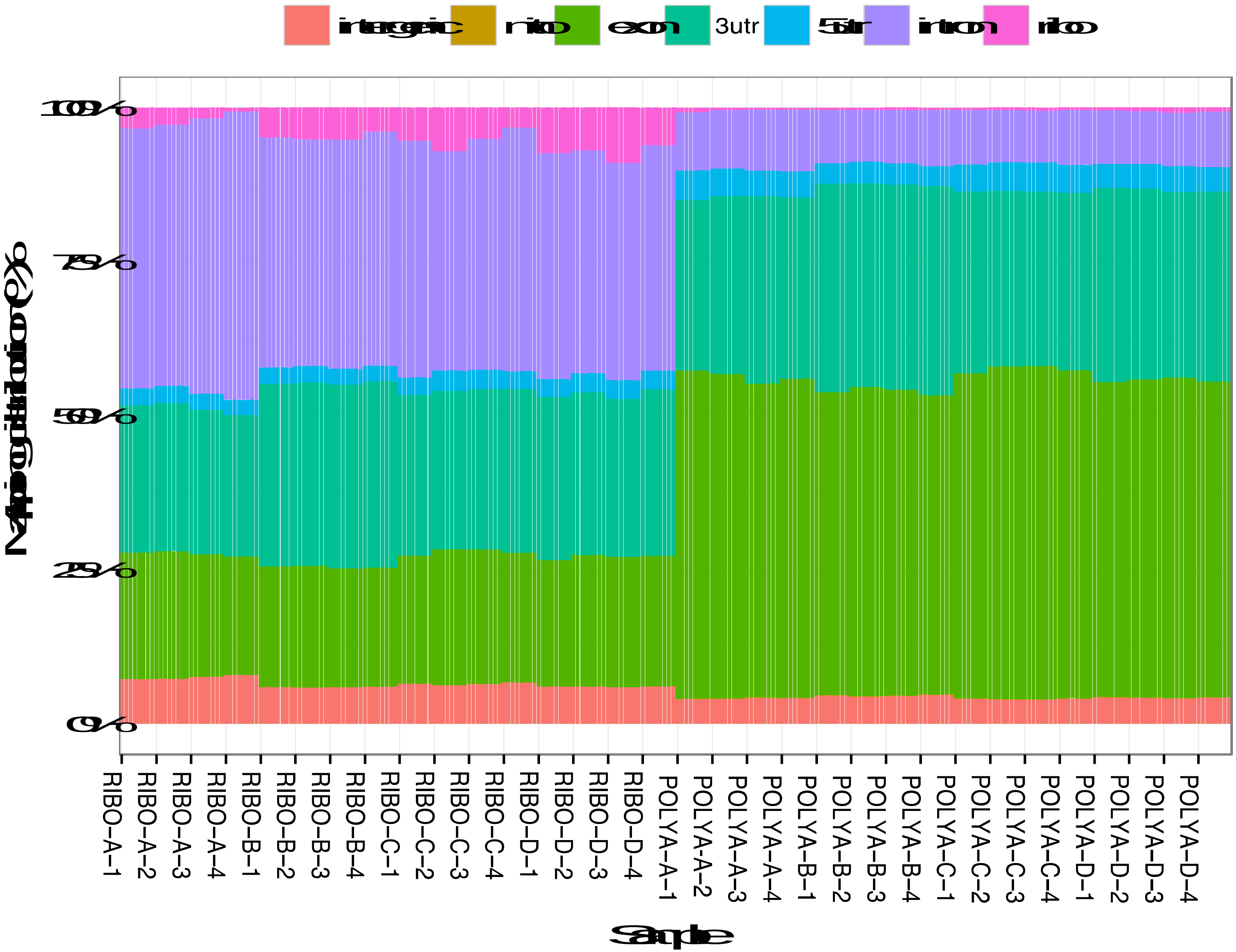
Unique DEGs:

454	- 3.0%
POLYA	- 9.2%
RIBO	- 8.8%
PRO	- 11.9%
PGM	- 3.9%

Proportional venn diagrams don’t always add clarity, but they are pretty

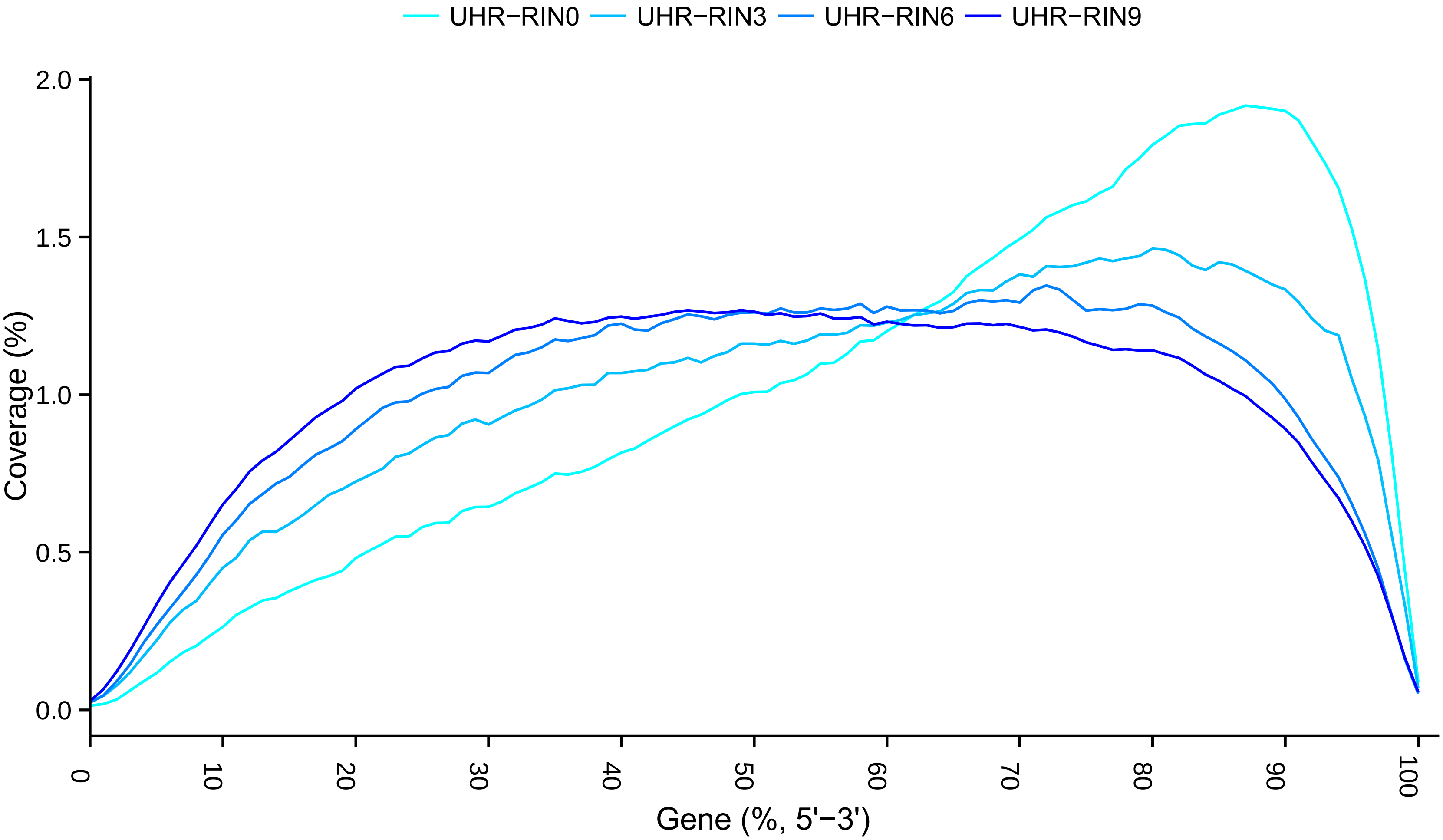


Gene regions distribution varies between protocols

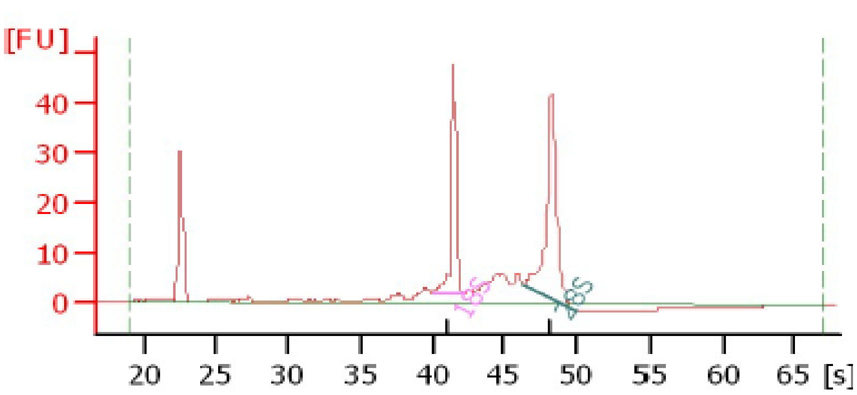


Supplemental Table 3 - Top 25 Genes with Highest Enrichment from Ribo-Depletion Preparation

ENSEMBL Gene ID	Gene Symbol	Description	Length	PolyA Reads	RiboDep Reads	PolyA FPKM	RiboDep FPKM	FPKM Diff
ENSG00000210082	J01415.4	Mt_rRNA	1559	17040155	133679955	18568.31	200404.74	-181836.43
ENSG00000211459	J01415.24	Mt_rRNA	954	2232892	14445488	3976.16	35389.28	-31413.11
ENSG00000202198	RN7SK	misc_RNA	331	3303	2818202	16.95	19899.03	-19882.08
ENSG00000258486	RN7SL1	antisense	300	3061	520624	17.33	4055.93	-4038.60
ENSG00000202364	SNORD3A	snoRNA	216	718	186779	5.65	2020.98	-2015.33
ENSG00000202538	RNU4-2	snRNA	141	168	102560	2.02	1699.99	-1697.97
ENSG00000251562	MALAT1	lincRNA	8708	437102	4931496	85.27	1323.57	-1238.30
ENSG00000199916	RMRP	misc_RNA	264	148	83019	0.95	734.96	-734.00
ENSG00000201098	RNY1	misc_RNA	113	172	19210	2.59	397.32	-394.73
ENSG00000238741	SCARNA7	snoRNA	330	53	50182	0.27	355.40	-355.13
ENSG00000200795	RNU4-1	snRNA	141	35	18724	0.42	310.36	-309.94
ENSG00000200087	SNORA73B	snoRNA	204	336	23778	2.80	272.42	-269.62
ENSG00000252010	SCARNA5	snoRNA	276	29	25379	0.18	214.91	-214.73
ENSG00000199568	RNU5A-1	snRNA	116	38	10224	0.56	205.99	-205.44
ENSG00000252481	SCARNA13	snoRNA	275	145	24034	0.90	204.26	-203.36
ENSG00000212232	SNORD17	snoRNA	237	102	20571	0.73	202.86	-202.13
ENSG00000207008	SNORA54	snoRNA	123	8	9655	0.11	183.46	-183.35
ENSG00000200156	RNU5B-1	snRNA	116	28	8301	0.41	167.25	-166.84
ENSG00000209582	SNORA48	snoRNA	135	38	9272	0.48	160.52	-160.04
ENSG00000239002	SCARNA10	snoRNA	330	19	21801	0.10	154.40	-154.30
ENSG00000254911	SCARNA9	antisense	353	501	19842	2.41	131.37	-128.96
ENSG00000230043	TMSB4XP6	pseudogene	135	0	6272	0.00	108.58	-108.58
ENSG00000239039	SNORD13	snoRNA	104	0	4521	0.00	101.60	-101.60
ENSG00000208892	SNORA49	snoRNA	136	7	5352	0.09	91.97	-91.89
ENSG00000223336	RNU2-6P	snRNA	190	22	6365	0.20	78.29	-78.10

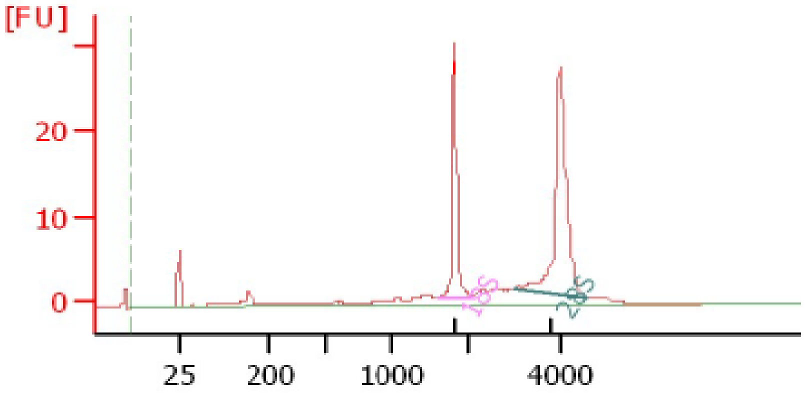

UHR−RIN0
UHR−RIN3
UHR−RIN6
UHR−RIN9
Coverage (%)
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
90
100
Gene (%, 5'−3')

Entropy is usually a source of fear



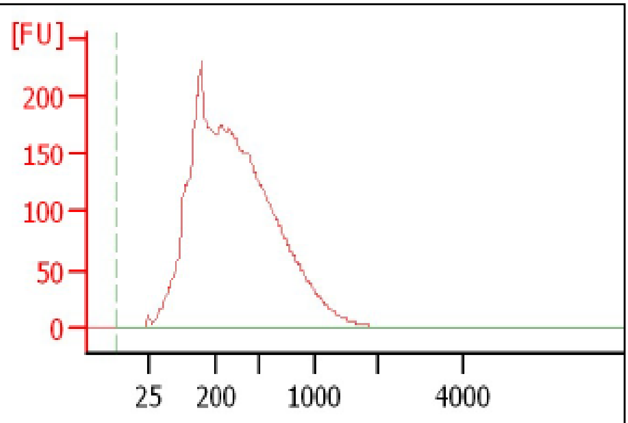
Overall Results for sample 6 : HBR-SV-ad1002sv

RNA Area: 199.0
RNA Concentration: 191 ng/µl
rRNA Ratio [28s / 18s]: 1.5
RNA Integrity Number (RIN): 8.9 (A.01.01)



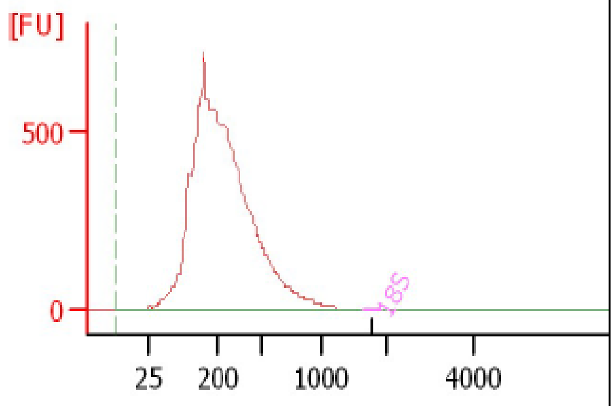
Overall Results for sample 1 : cont

RNA Area: 128.7
RNA Concentration: 119 ng/µl
rRNA Ratio [28s / 18s]: 1.8
RNA Integrity Number (RIN): 9.4 (B.02.08)



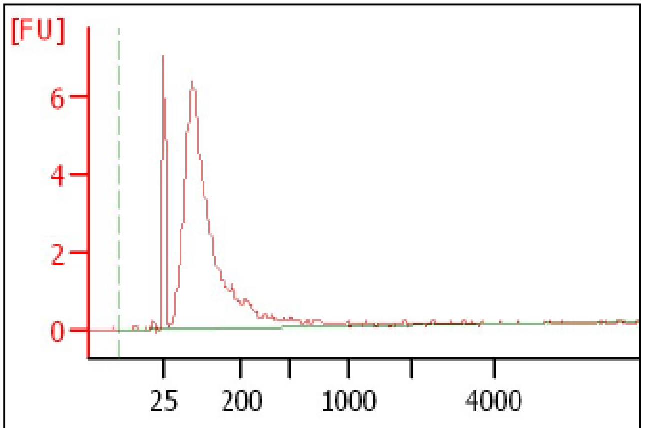
Overall Results for sample 5 : Heat-UR

RNA Area: 5,349.7
RNA Concentration: 38,678 pg/µl
rRNA Ratio [28s / 18s]: 0.0
RNA Integrity Number (RIN): 1.7 (B.02.08)



Overall Results for sample 4 : COV-UR

RNA Area: 11,841.2
RNA Concentration: 85,610 pg/µl
rRNA Ratio [28s / 18s]: 0.0
RNA Integrity Number (RIN): 1.9 (B.02.08)



Overall Results for sample 1 : UR RNA

RNA Area: 62.2
RNA Concentration: 61 ng/µl
rRNA Ratio [28s / 18s]: 0.0
RNA Integrity Number (RIN): 2.6 (B.02.08)

TECHNICIANS
the building blocks of research
International Laboratory Animal Technician Week
January 27–February 2, 2013

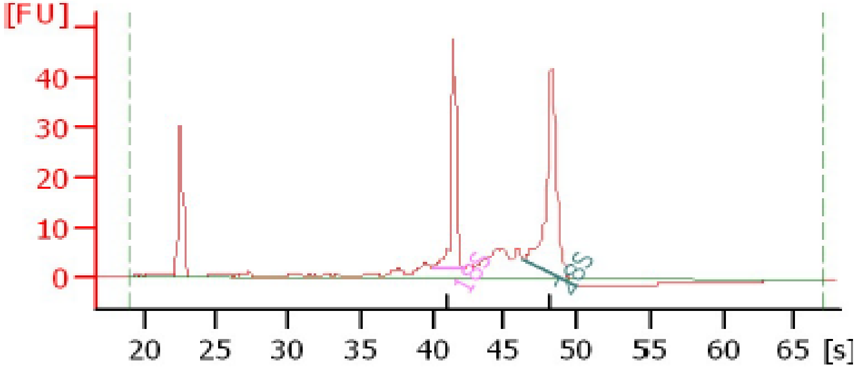
"**FEAR** is the main source of superstition,

and one of the main sources of cruelty

To conquer fear is the beginning of wisdom."

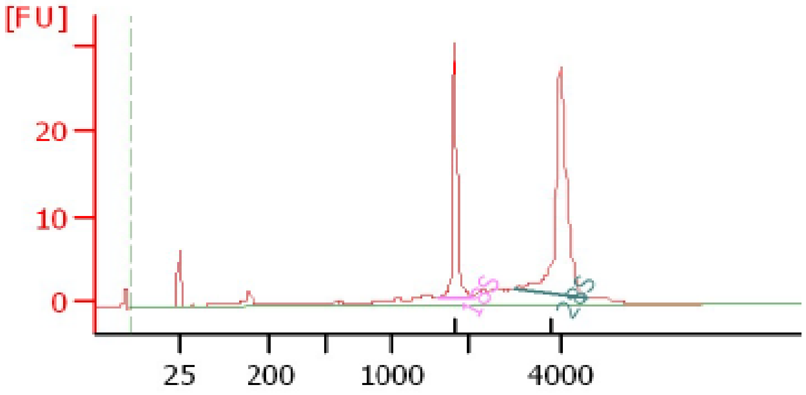
– Bertrand Russell

Can we remove superstition?



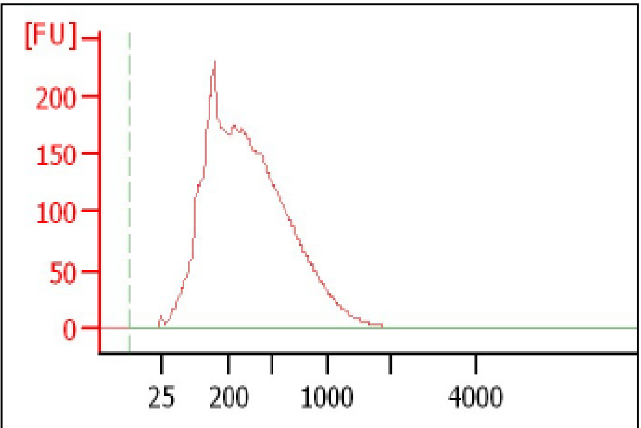
Overall Results for sample 6 : HBR-SV-ad1002sv

RNA Area:	199.0
RNA Concentration:	191 ng/µl
rRNA Ratio [28s / 18s]:	1.5
RNA Integrity Number (RIN):	8.9 (A.01.01)



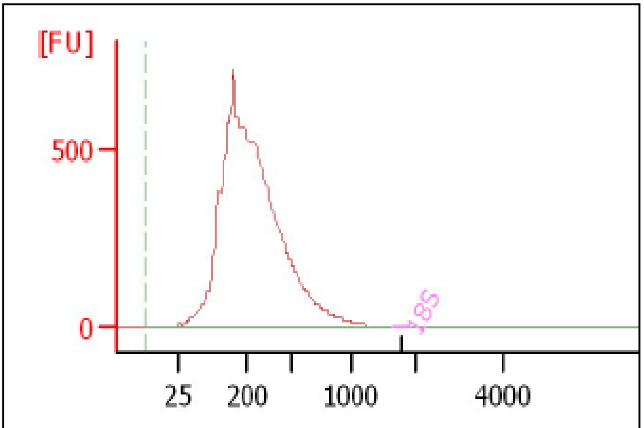
Overall Results for sample 1 : cont

RNA Area:	128.7
RNA Concentration:	119 ng/µl
rRNA Ratio [28s / 18s]:	1.8
RNA Integrity Number (RIN):	9.4 (B.02.08)



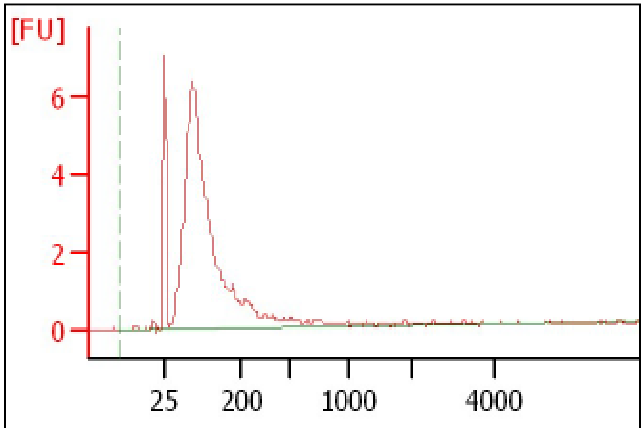
Overall Results for sample 5 : Heat-UR

RNA Area:	5,349.7
RNA Concentration:	38,678 pg/µl
rRNA Ratio [28s / 18s]:	0.0
RNA Integrity Number (RIN):	1.7 (B.02.08)



Overall Results for sample 4 : COV-UR

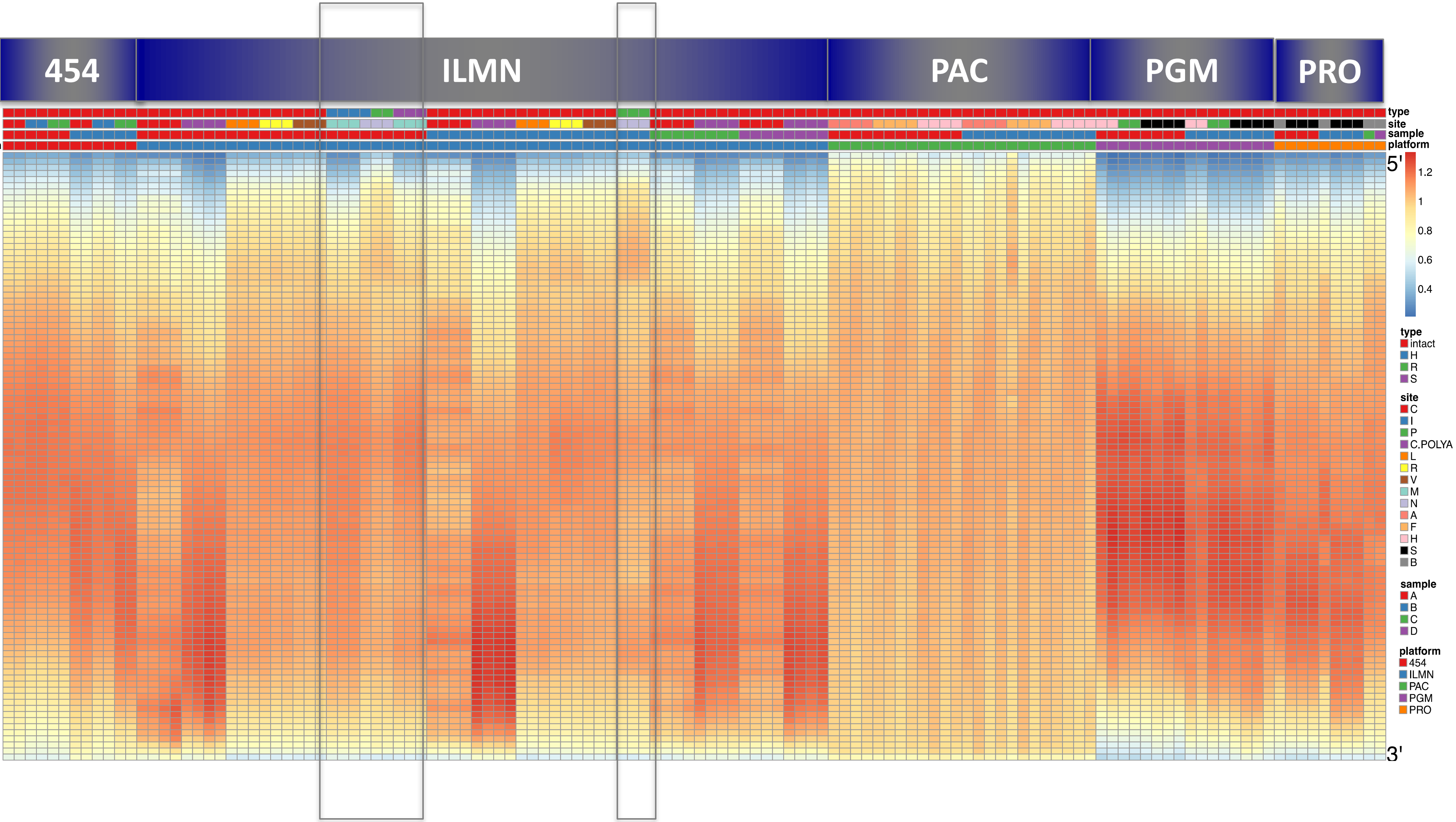
RNA Area:	11,841.2
RNA Concentration:	85,610 pg/µl
rRNA Ratio [28s / 18s]:	0.0
RNA Integrity Number (RIN):	1.9 (B.02.08)



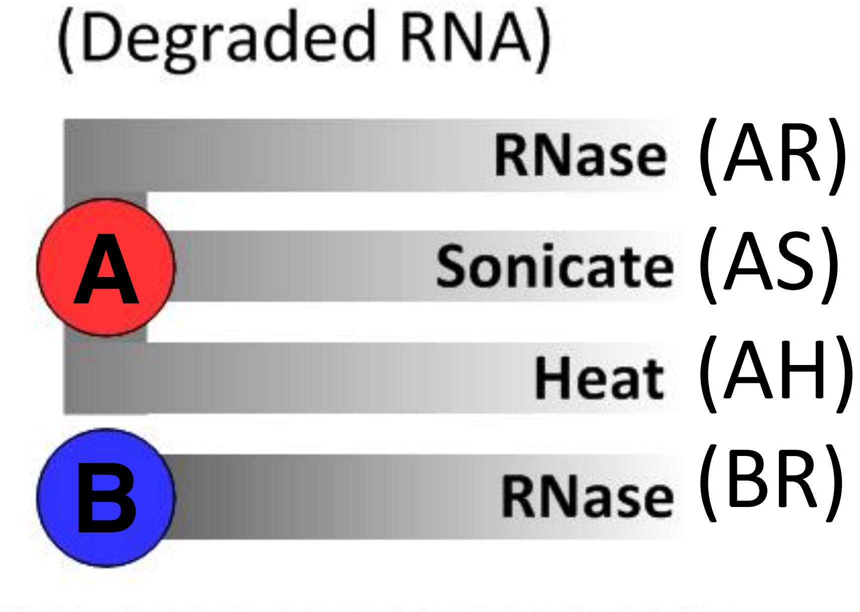
Overall Results for sample 1 : UR RNA

RNA Area:	62.2
RNA Concentration:	61 ng/µl
rRNA Ratio [28s / 18s]:	0.0
RNA Integrity Number (RIN):	2.6 (B.02.08)

Degraded RNA looks great!

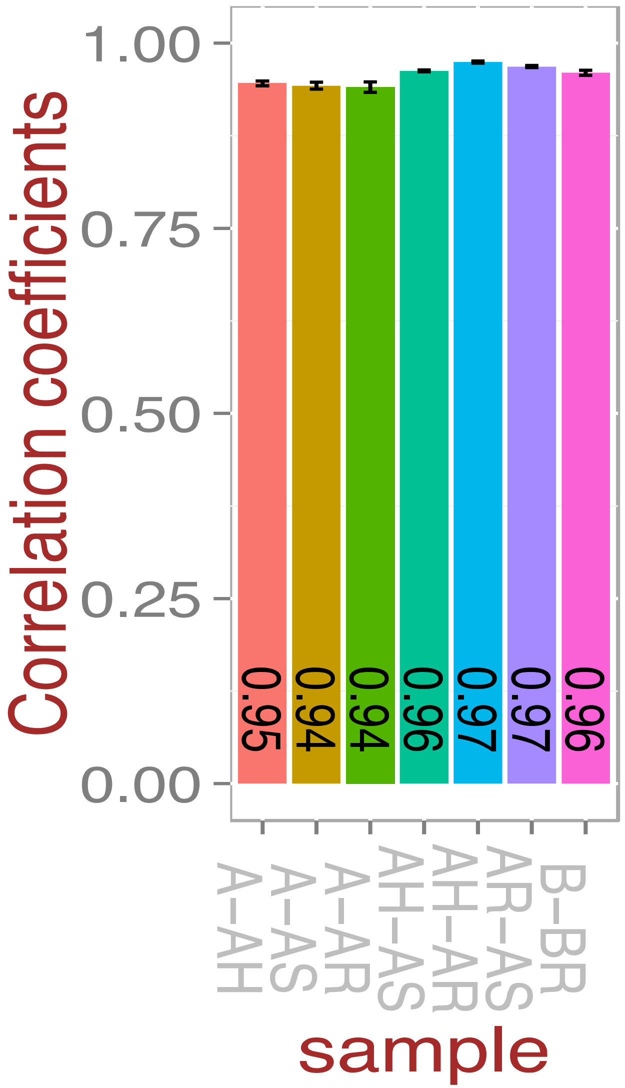


Degraded RNA highly correlates with intact RNA gene expression

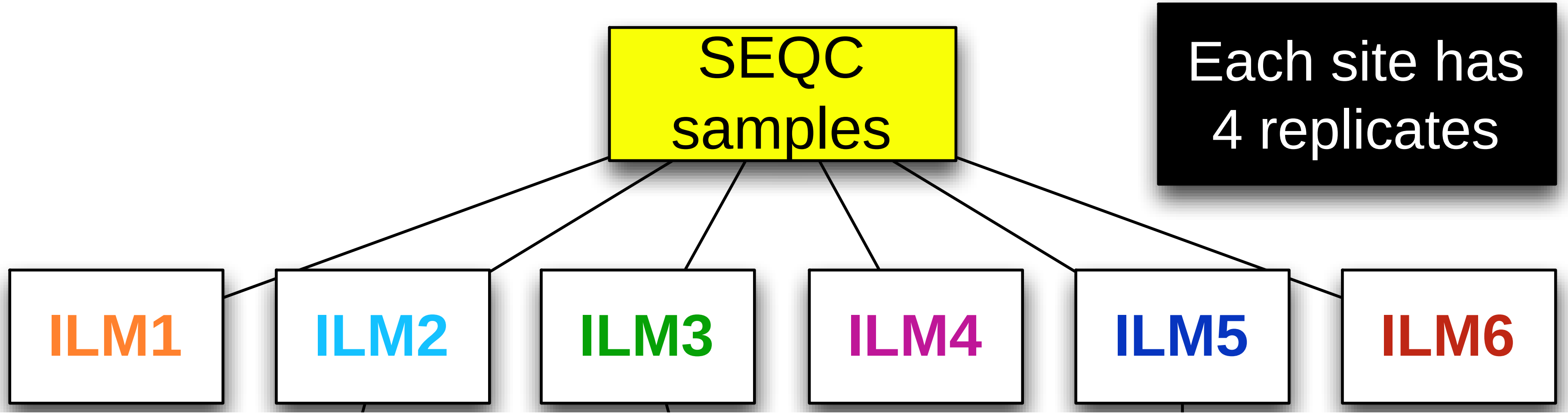


(Degraded RNA)

Illumina Ribo-depletion protocol



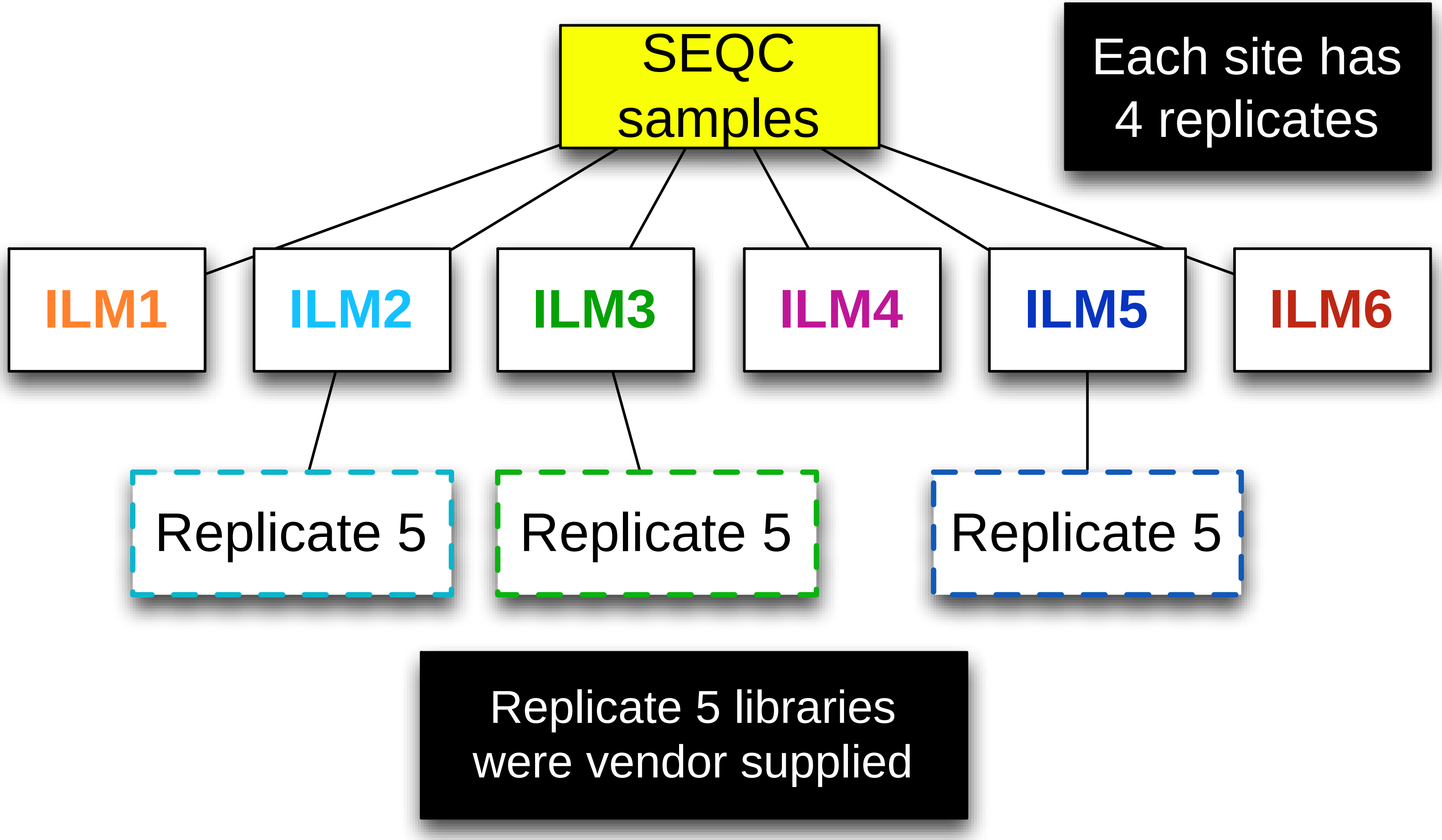
Ameliorating Inter-site Variation



Each site has 4 replicates

HiSeq 2000

Ameliorating Inter-site Variation

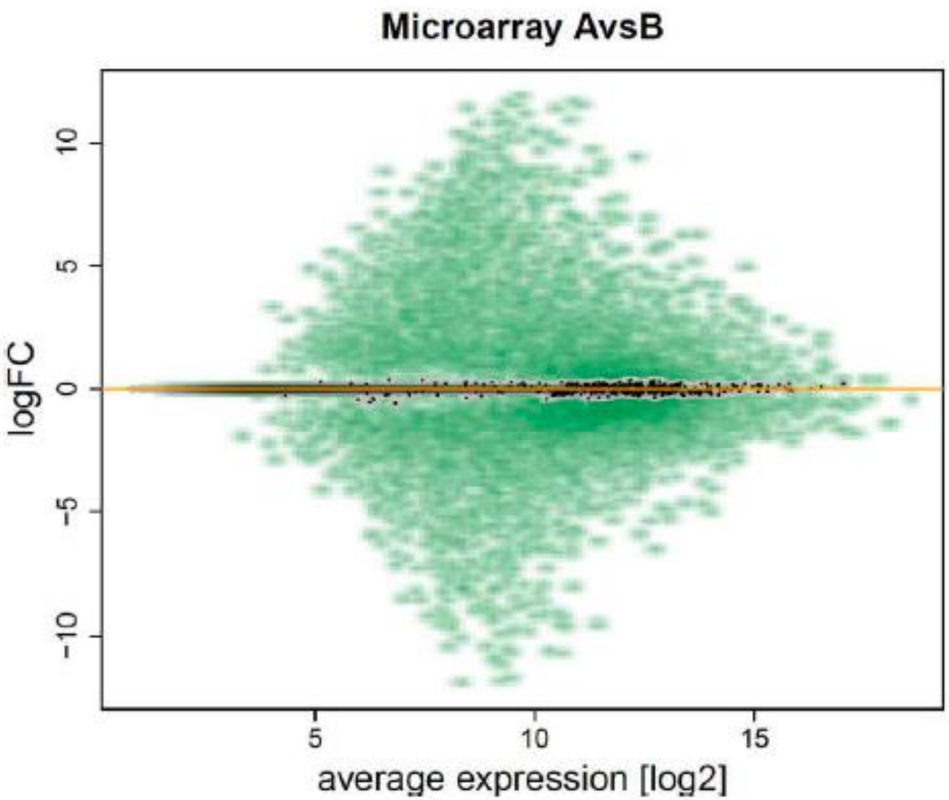


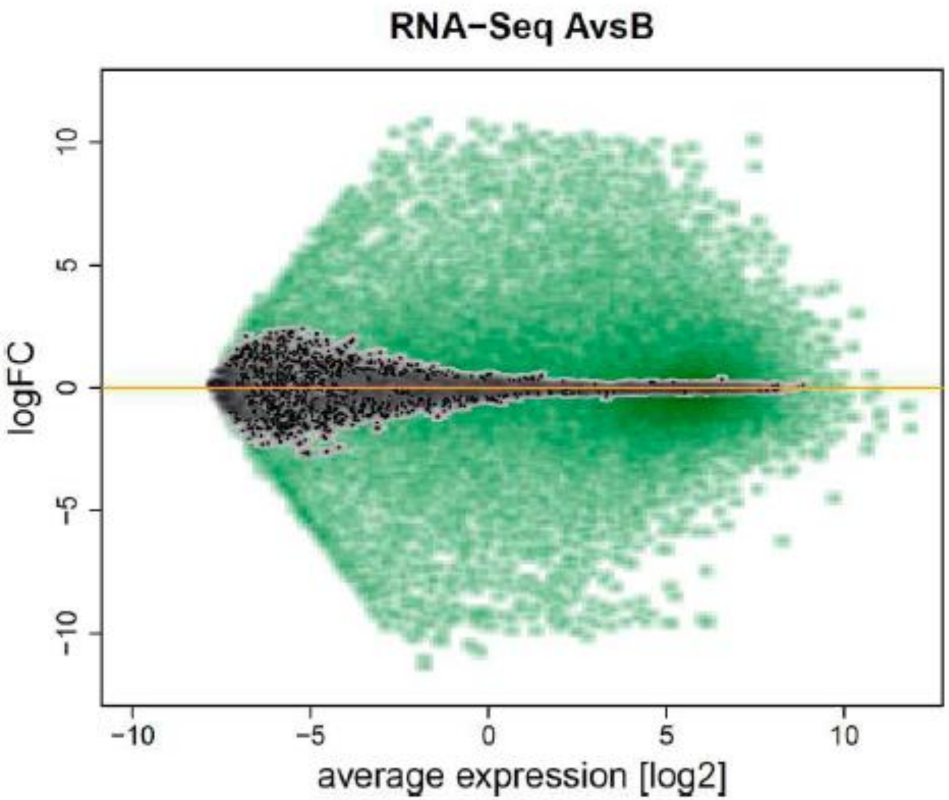
Each site has 4 replicates

Replicate 5 libraries were vendor supplied

Differential expression calls – AvsB

significant @ $p < 1\%$

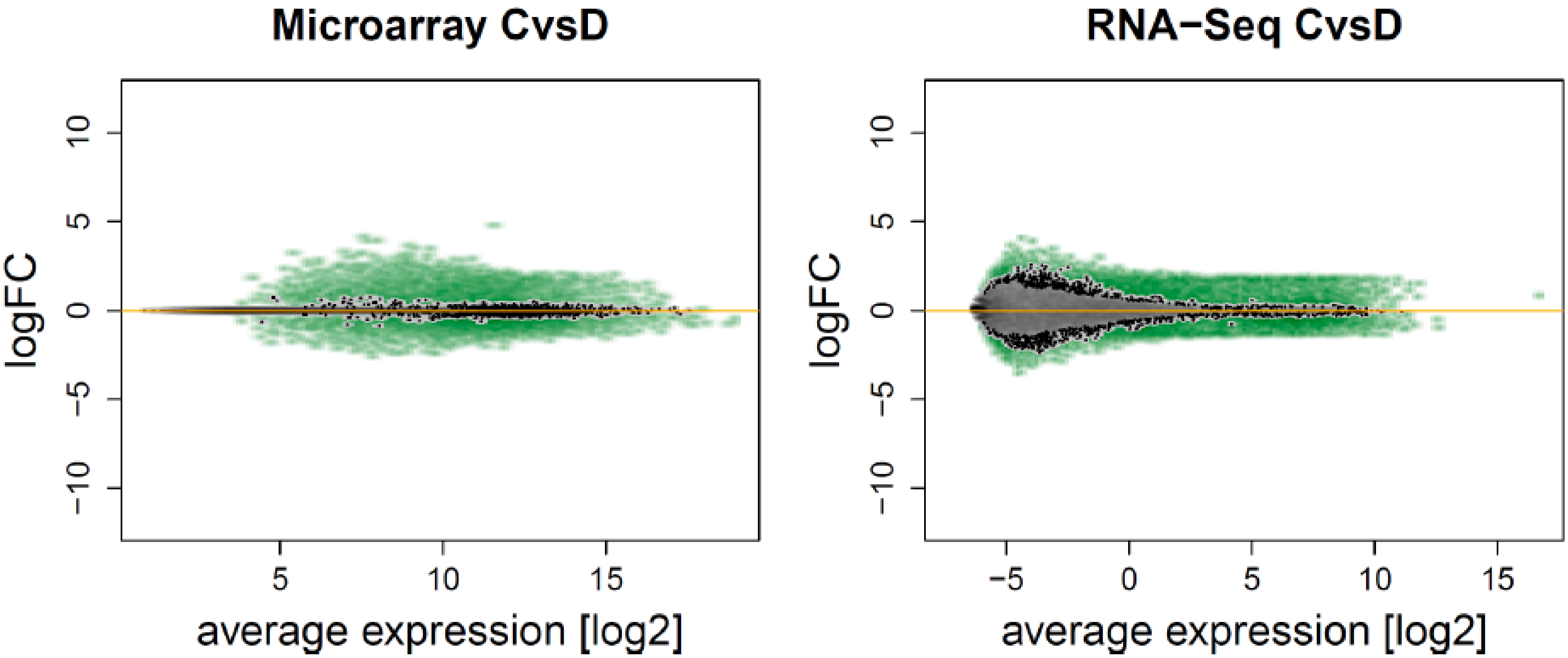

Microarray AvsB


RNA−Seq AvsB

Paweł Łabaj

Differential expression calls – CvsD

significant @ *p* < 1%



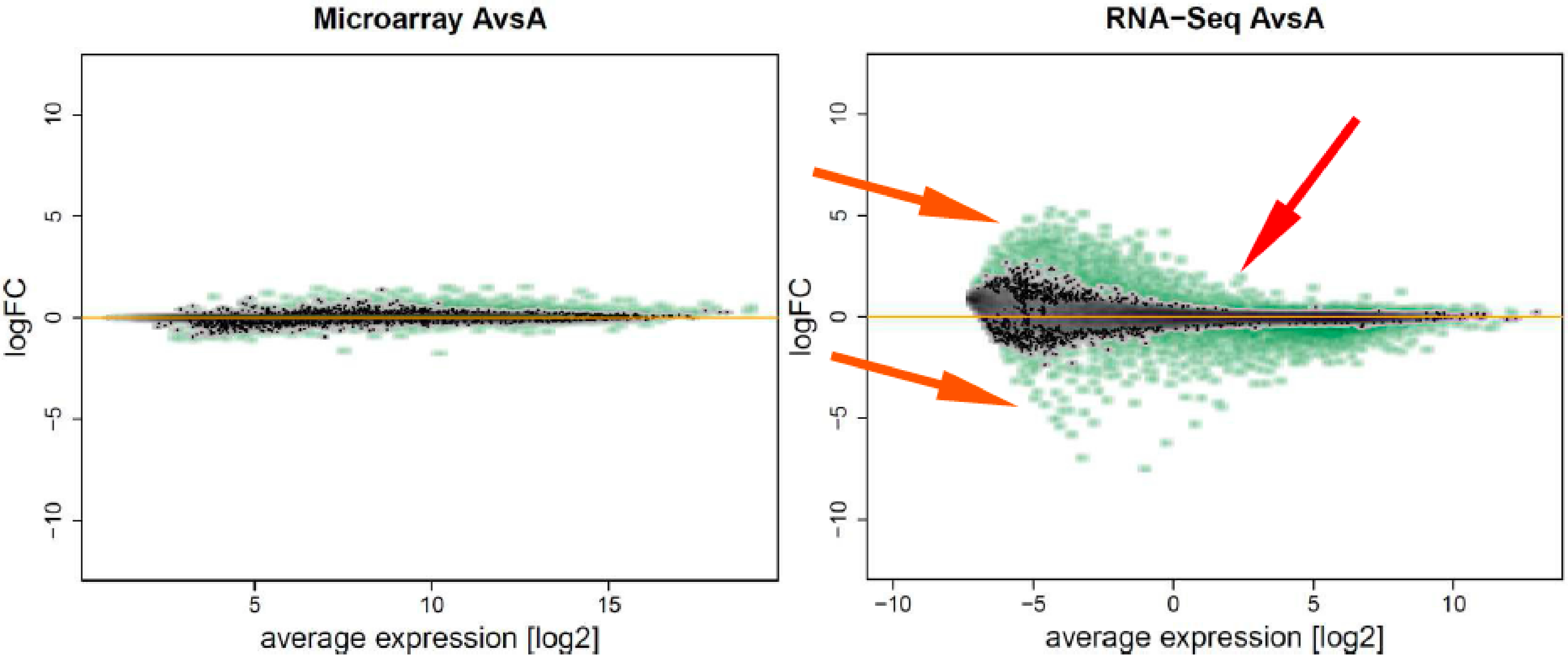
Microarray CvsD

RNA–Seq CvsD

Paweł Łabaj

Differential expression calls – AvsA

significant @ $p < 1\%$



RNA-Seq and microarrays (MAQC-I): (Nature Biotech, 2006)

– site to site variation → ~50% eFDR

Paweł Łaba

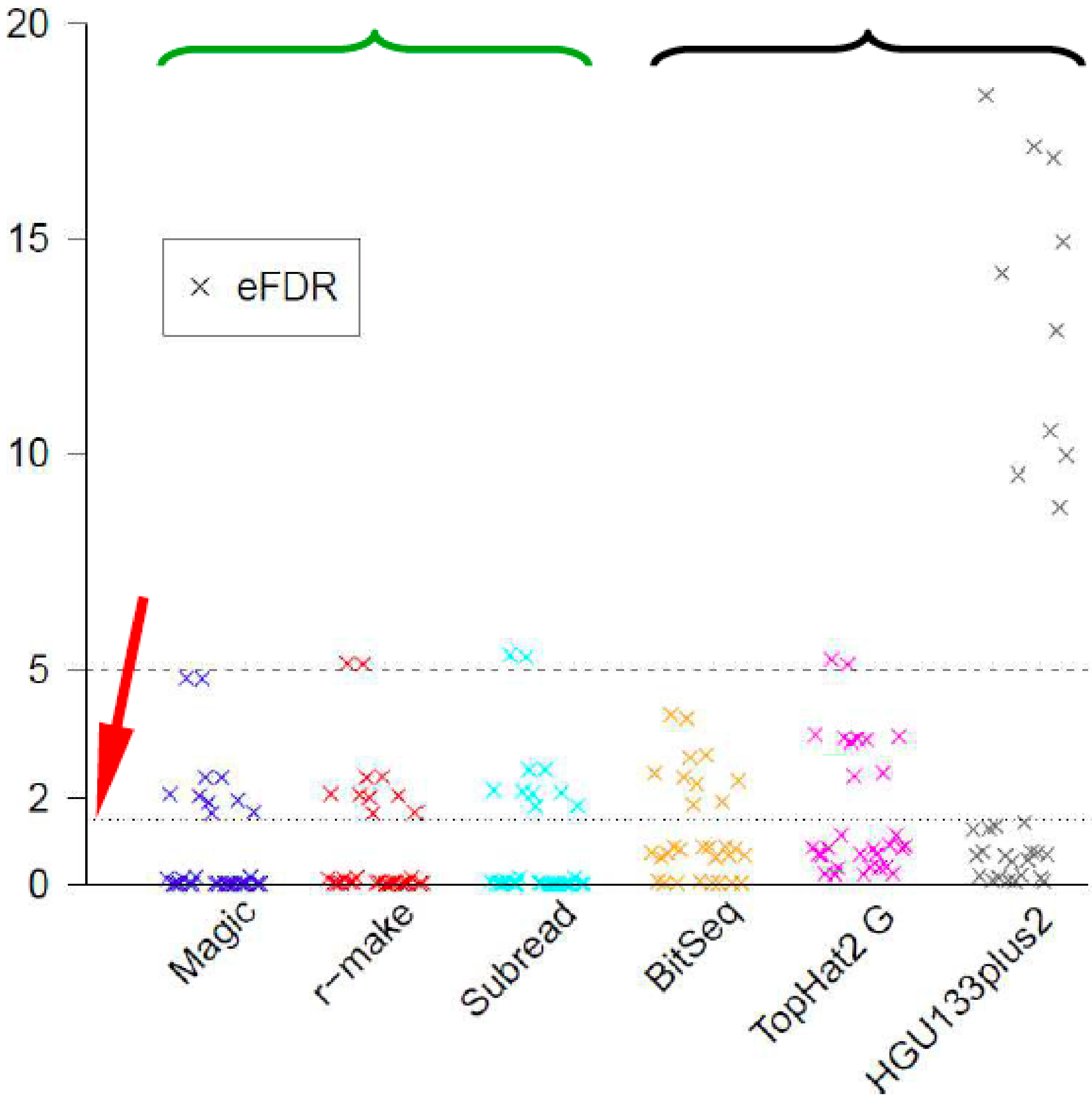
Differential expression calls – reproducibility across sites

All platforms suffer outlier sites

- validation!

With the right filters

- eFDR < 1.5% without outlier sites
- RNA-Seq *can be* more specific than microarrays (pipeline!)



Paweł Łabaj

Systematic variation removal - False Positives

GC-content bias correction:

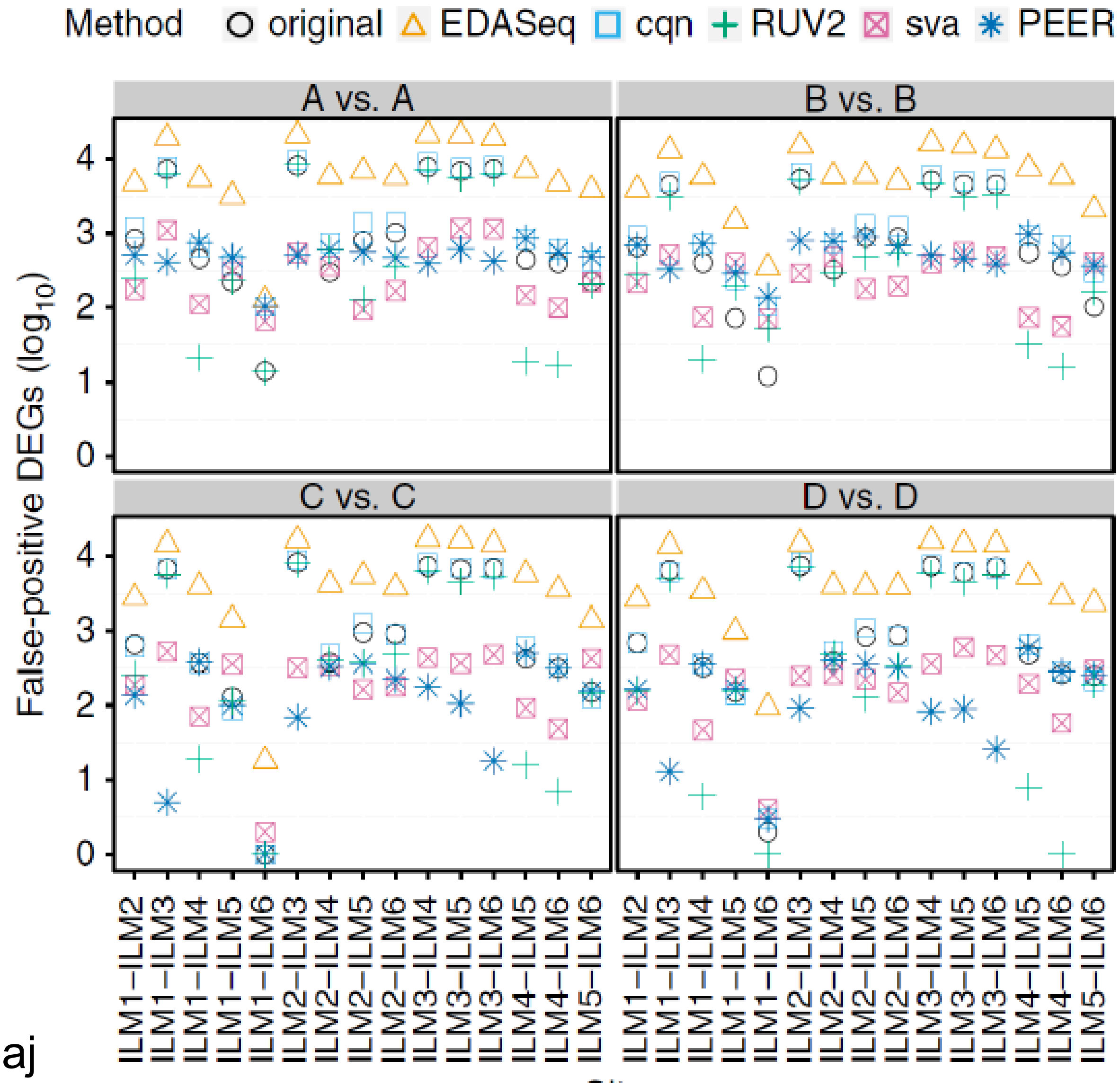
- EDASeq
- cqn

Factor analysis based on ERCCs:

- RUV2 (ERCC)

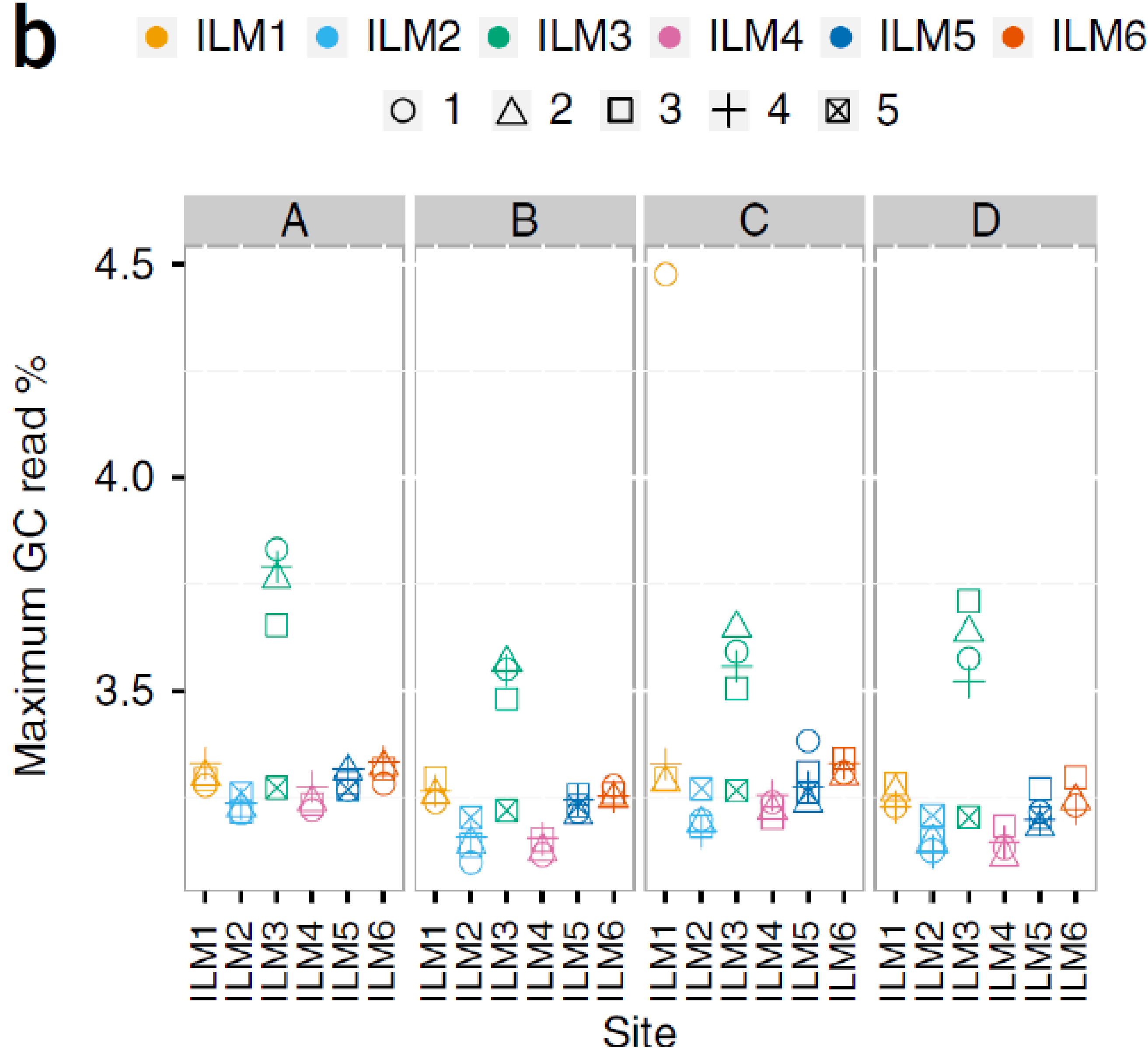
Latent variables:

- sva
- PEER



Paweł Łabaj

Identification of the underlying sources of variation – GC content

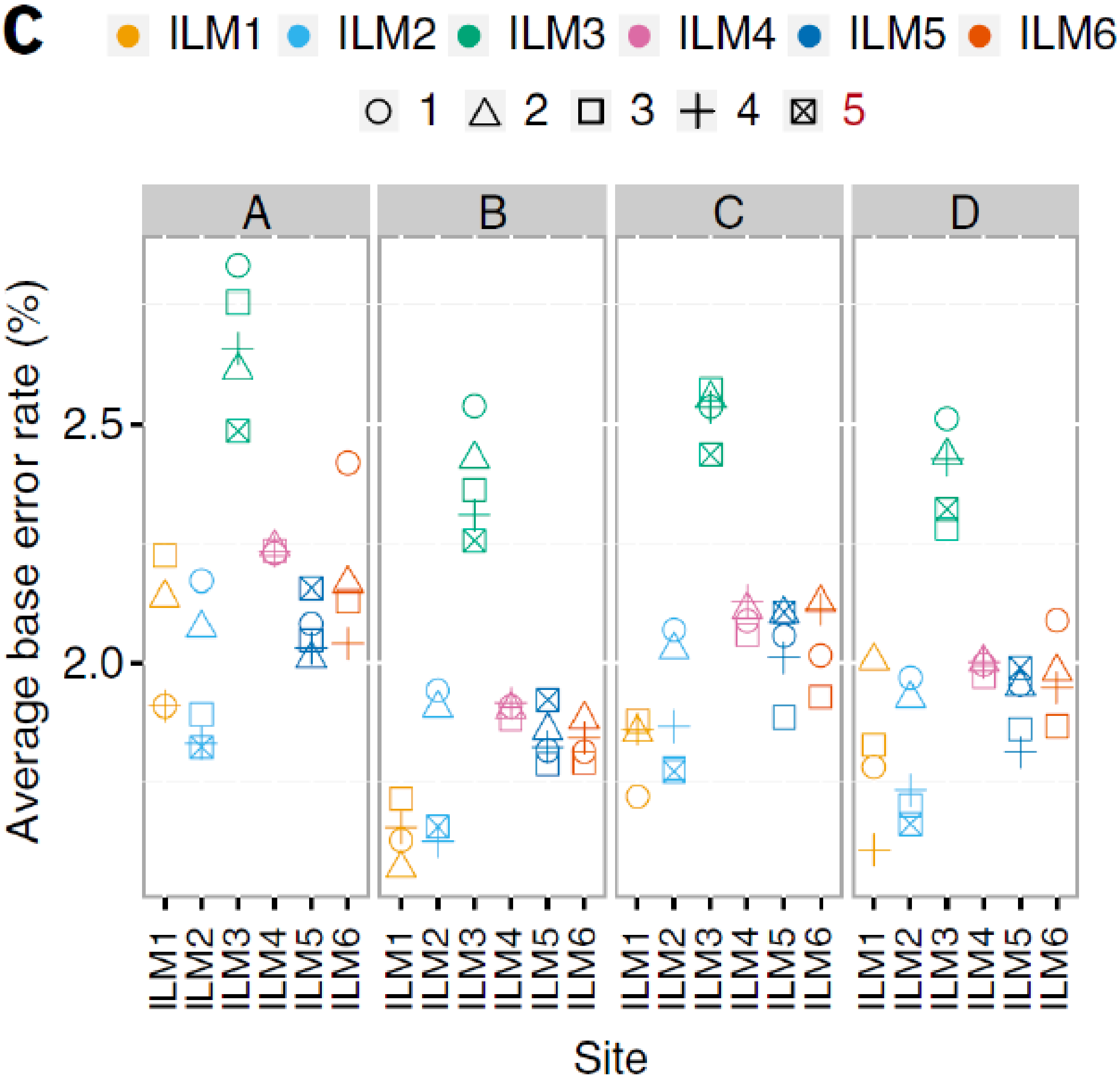


Outlier site:
ILM 3

5th replicate
→ **not** affected

Paweł Łabaj

Identification of the underlying sources of variation – base error rate

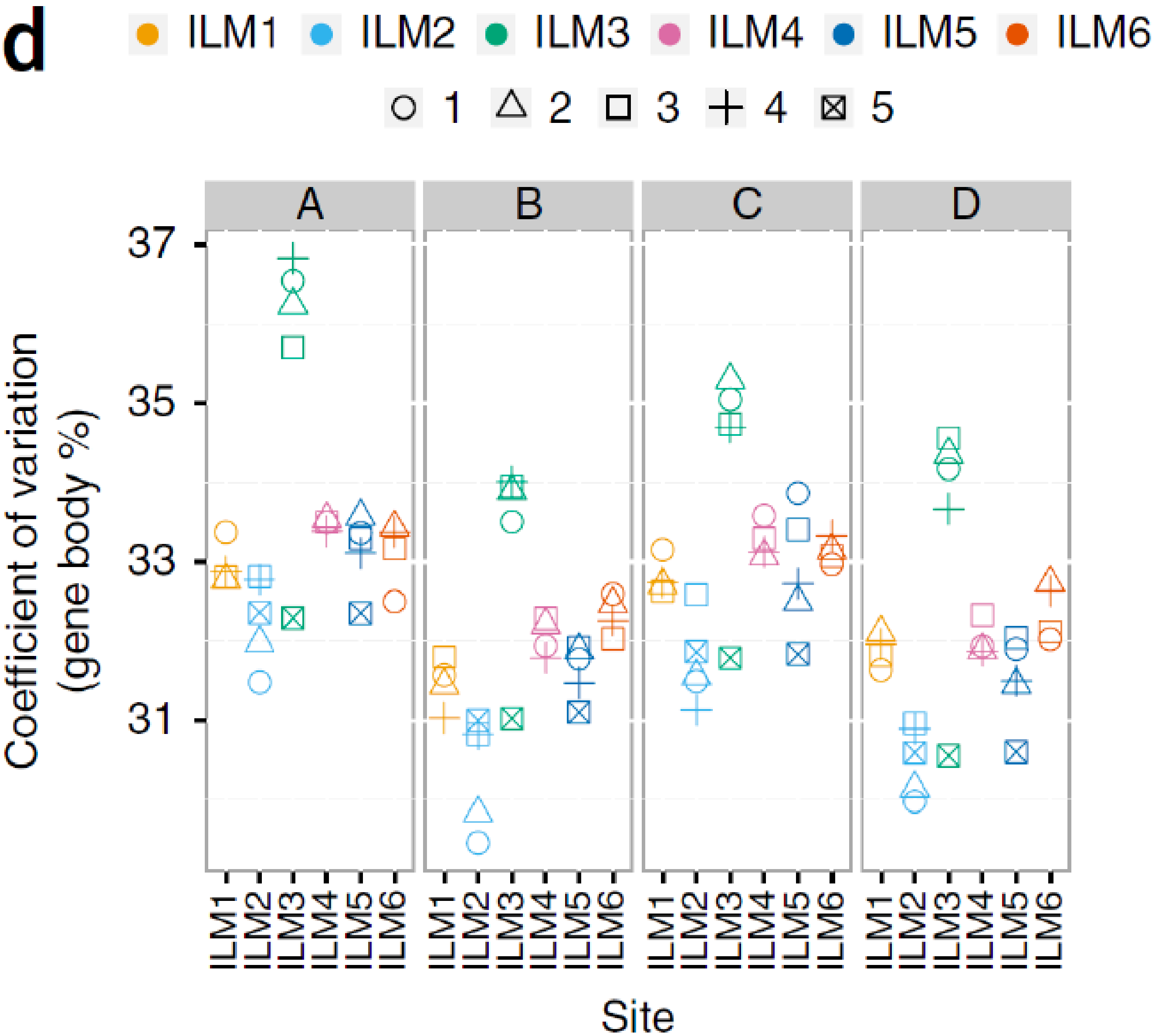


Outlier site:
ILM 3

5th replicate
→ *affected*

Paweł Łabaj

Identification of the underlying sources of variation – gene body



Outlier site:
ILM 3

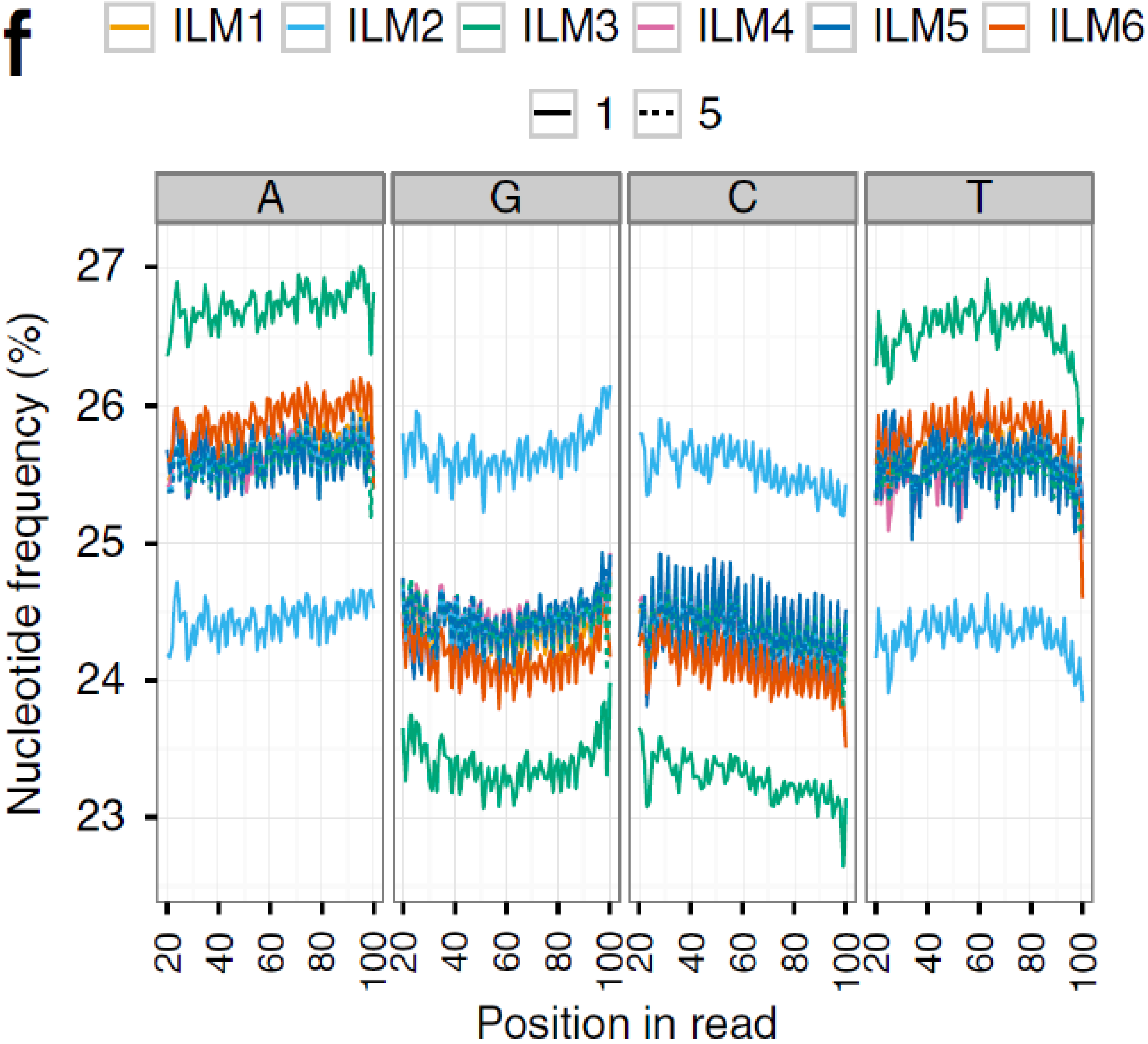
5th replicate
→ **not** affected

Paweł Łabaj

Identification of the underlying sources of variation – nucleotide composition

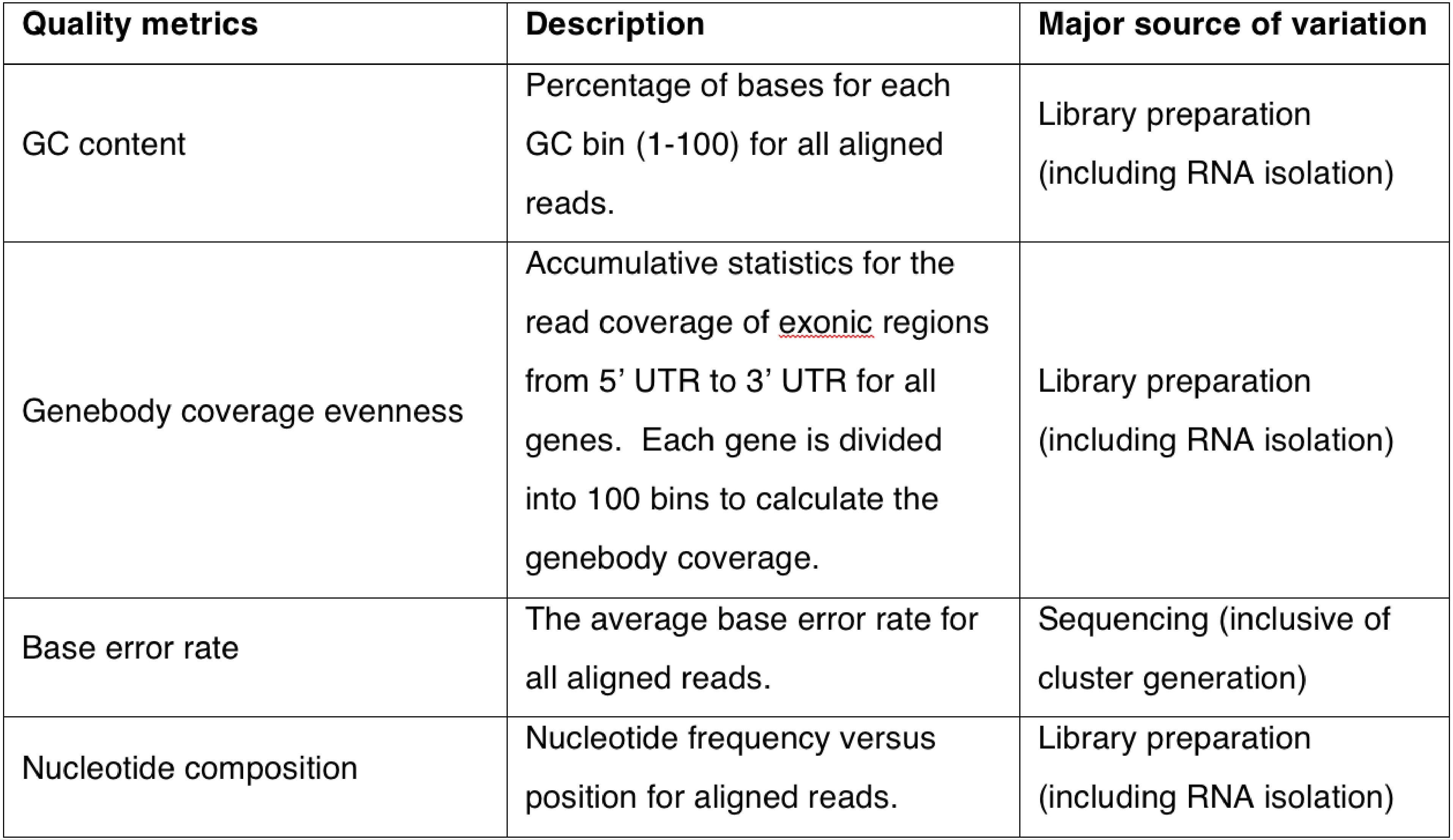
Outlier sites:
ILM 2 and ILM 3

5th replicate
→ **not** affected



Paweł Łabaj

Determine sequencing variation sources



Quality metrics	Description	Major source of variation
GC content	Percentage of bases for each GC bin (1-100) for all aligned reads.	Library preparation (including RNA isolation)
Genebody coverage evenness	Accumulative statistics for the read coverage of exonic regions from 5' UTR to 3' UTR for all genes. Each gene is divided into 100 bins to calculate the genebody coverage.	Library preparation (including RNA isolation)
Base error rate	The average base error rate for all aligned reads.	Sequencing (inclusive of cluster generation)
Nucleotide composition	Nucleotide frequency versus position for aligned reads.	Library preparation (including RNA isolation)

(4)

Annotations

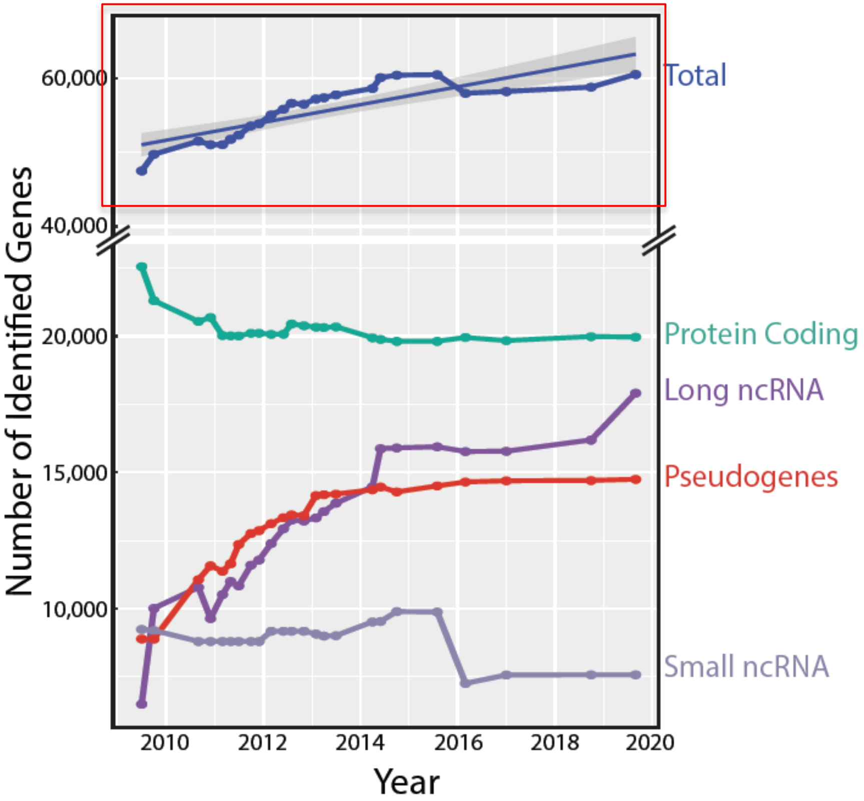
Your exome is not 62Mb

The 62 Mb "exome capture" is really the 1/3 exome capture

	Aceview	UCSC	Vega	ENSEMBL	Refseq
Aceview	178				
UCSC	76	81			
Vega	51	42	58		
ENSEMBL	64	60	43	70	
RefSeq	60	61	37	57	62

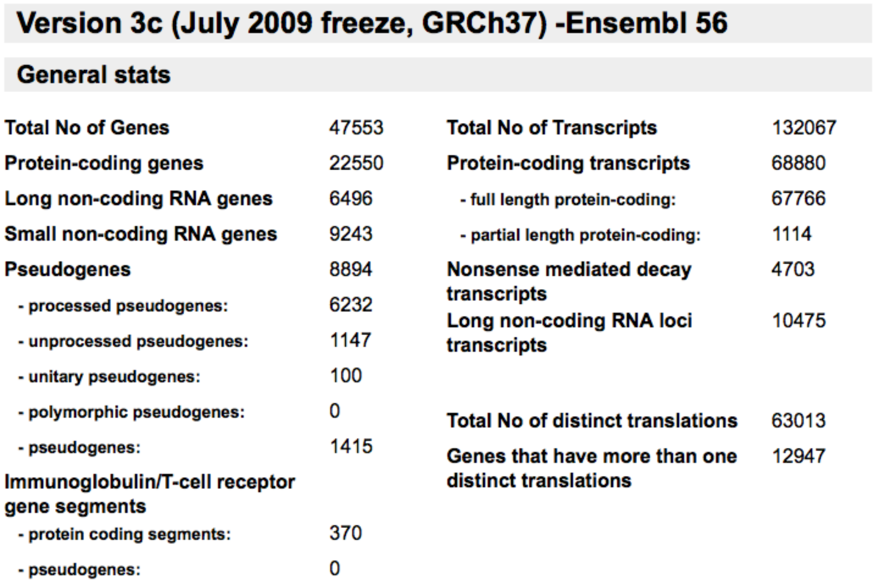
Zumbo and Mason, Genome Analysis: Current Procedures and Applications, 2013

New human genes are still being found



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

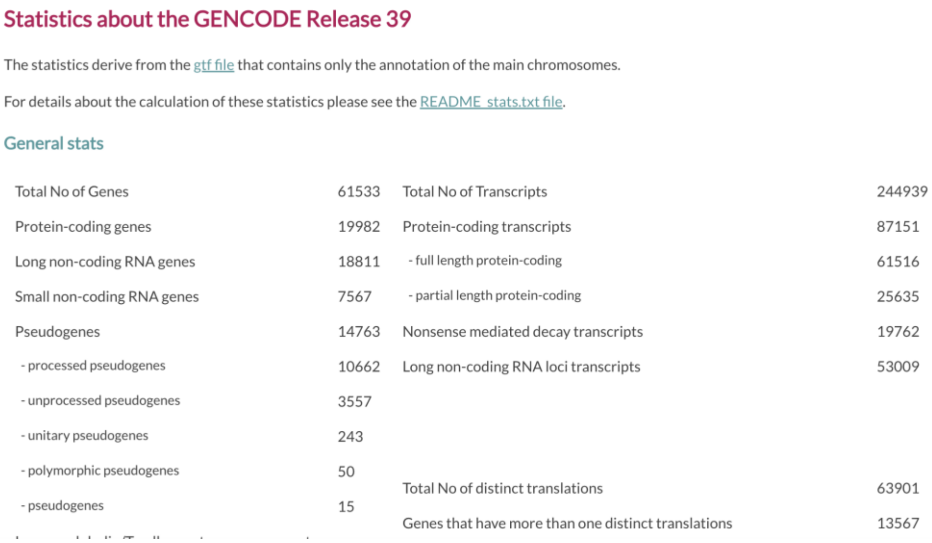
Annotations are a shifting sand, but so is the genome



Version 3c (July 2009 freeze, GRCh37) -Ensembl 56

General stats

Total No of Genes	47553	**Total No of Transcripts**	132067
Protein-coding genes	22550	**Protein-coding transcripts**	68880
Long non-coding RNA genes	6496	**- full length protein-coding:**	67766
Small non-coding RNA genes	9243	**- partial length protein-coding:**	1114
Pseudogenes	8894	**Nonsense mediated decay transcripts**	4703
- processed pseudogenes:	6232	**Long non-coding RNA loci transcripts**	10475
- unprocessed pseudogenes:	1147		
- unitary pseudogenes:	100		
- polymorphic pseudogenes:	0	**Total No of distinct translations**	63013
- pseudogenes:	1415	**Genes that have more than one distinct translations**	12947
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments:	370		
- pseudogenes:	0		



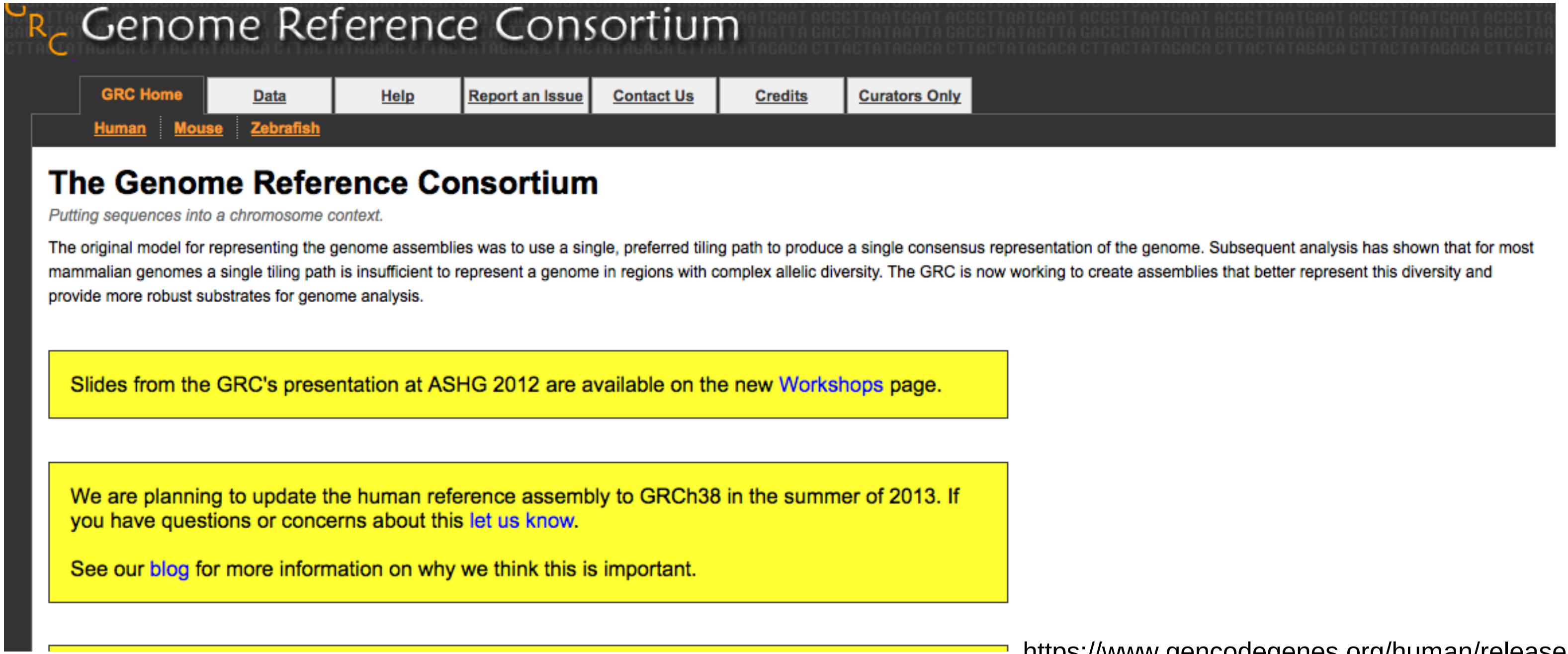
Statistics about the GENCODE Release 39

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	61533	Total No of Transcripts	244939
Protein-coding genes	19982	Protein-coding transcripts	87151
Long non-coding RNA genes	18811	- full length protein-coding	61516
Small non-coding RNA genes	7567	- partial length protein-coding	25635
Pseudogenes	14763	Nonsense mediated decay transcripts	19762
- processed pseudogenes	10662	Long non-coding RNA loci transcripts	53009
- unprocessed pseudogenes	3557		
- unitary pseudogenes	243		
- polymorphic pseudogenes	50	Total No of distinct translations	63901
- pseudogenes	15	Genes that have more than one distinct translations	13567



The Genome Reference Consortium

Putting sequences into a chromosome context.

The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

Slides from the GRC's presentation at ASHG 2012 are available on the new Workshops page.

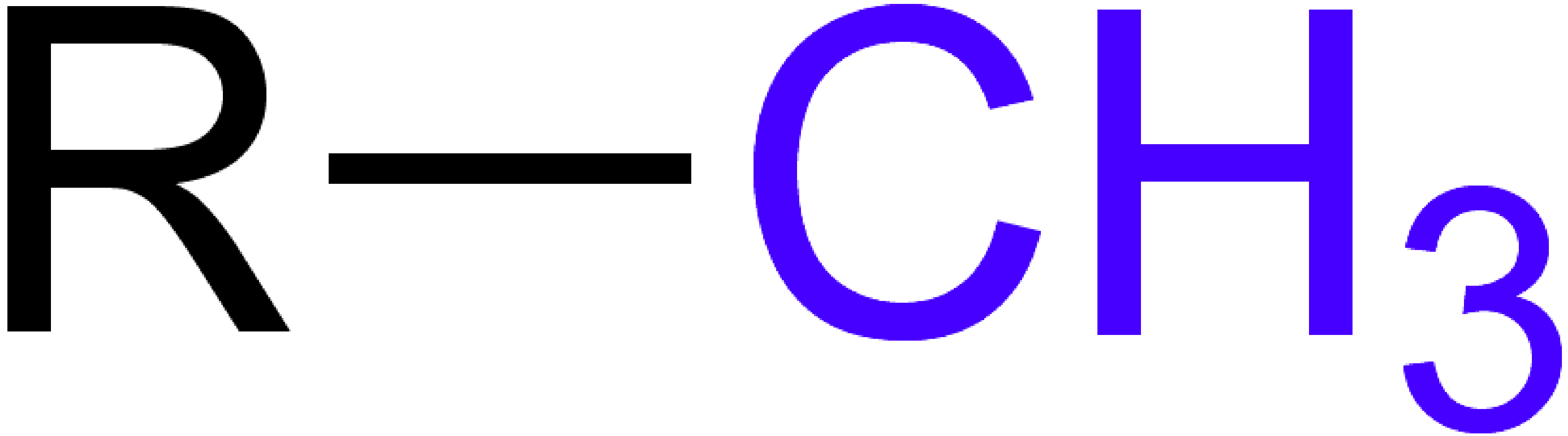
We are planning to update the human reference assembly to GRCh38 in the summer of 2013. If you have questions or concerns about this let us know.

See our blog for more information on why we think this is important.

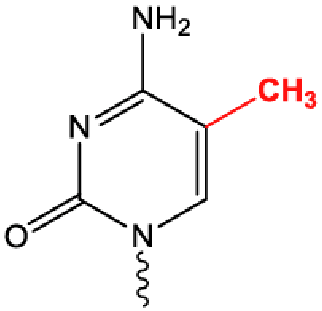
https://www.gencodegenes.org/human/releases.html

(5)

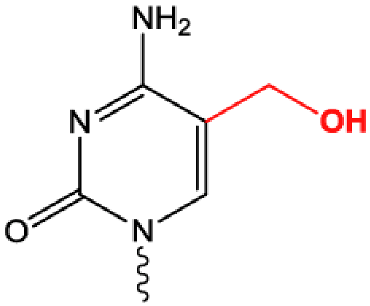
Epitranscriptome


R
CH3

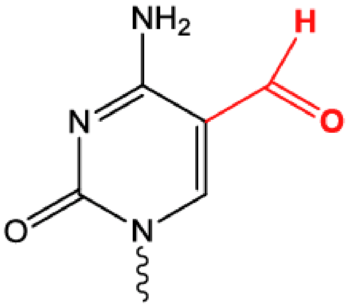
The four-base genome is just the beginning



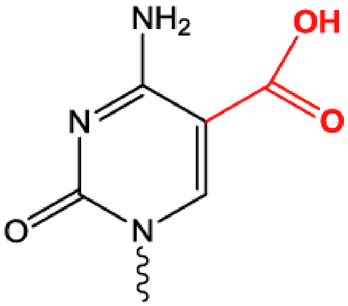
5-mC



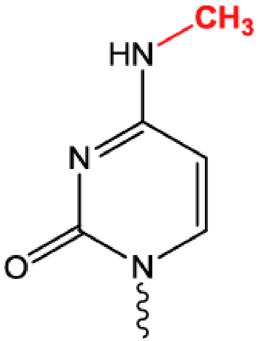
5-hmC



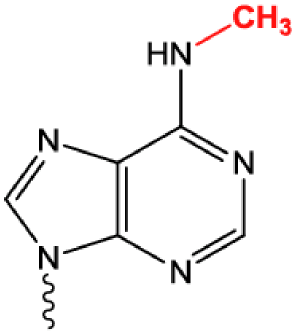
5-fC



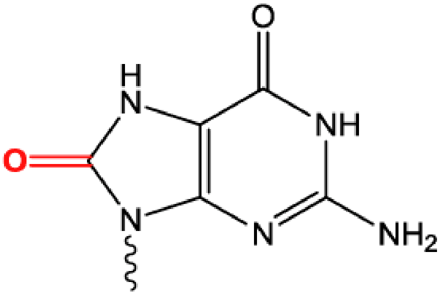
5-caC



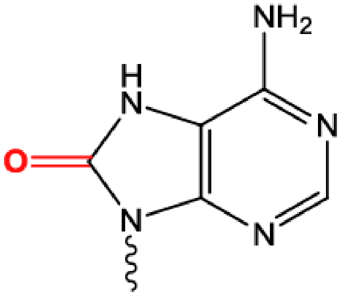
4-mC



6-mA



8-oxoG



8-oxoA

There are many RNA-mods as well:

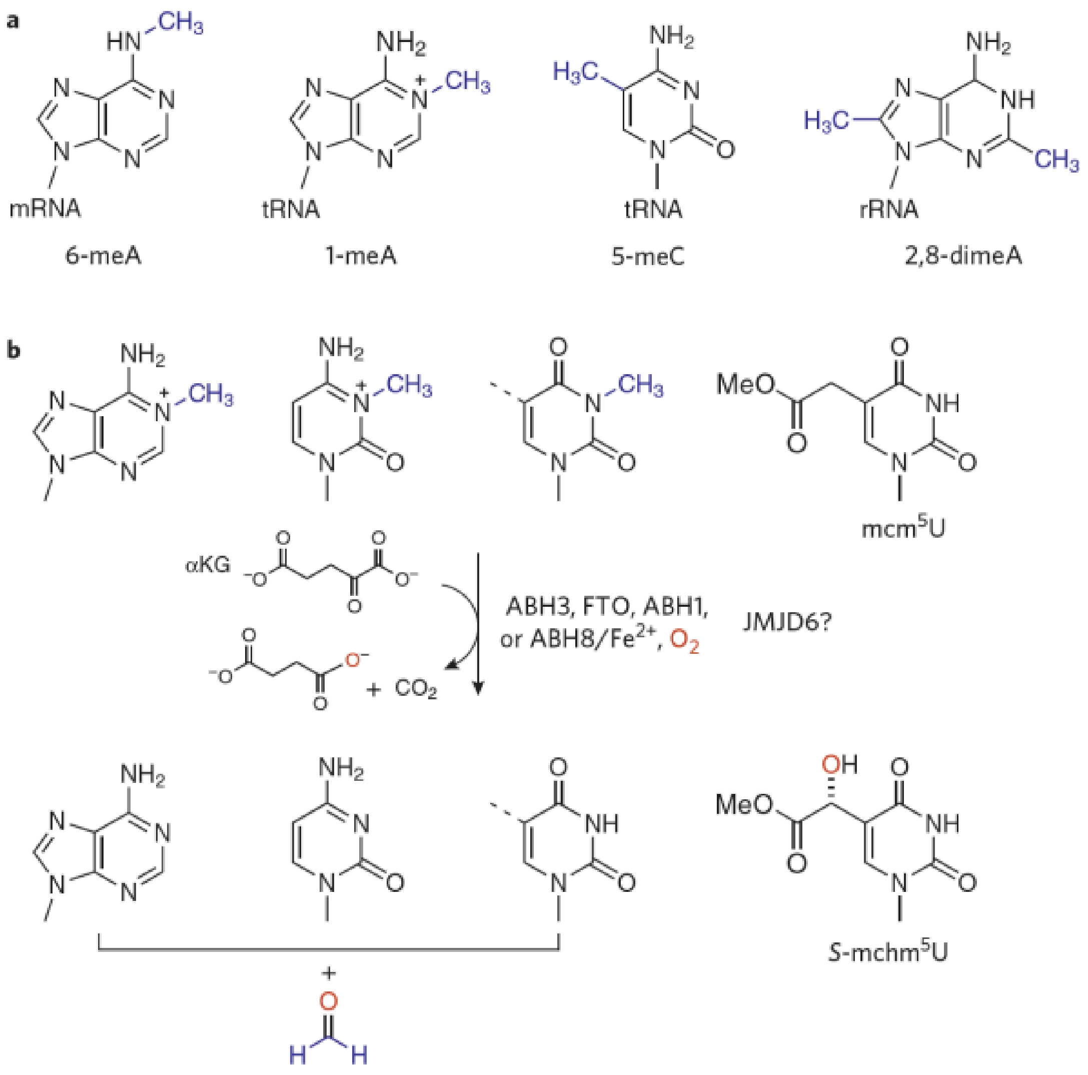
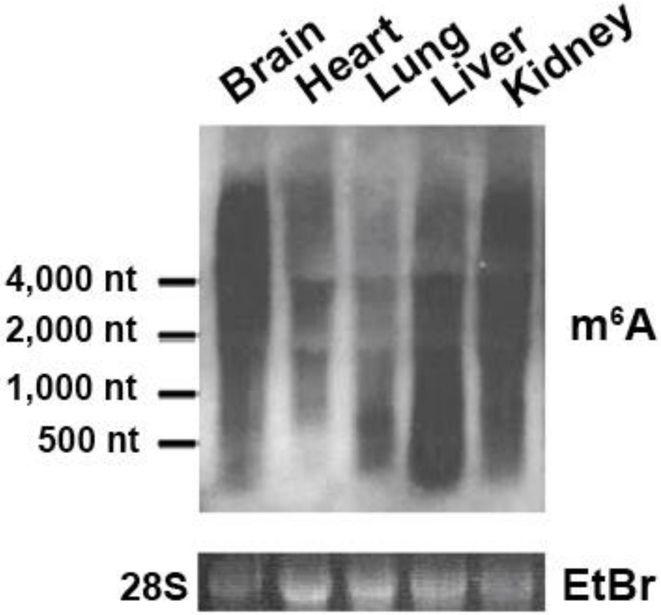
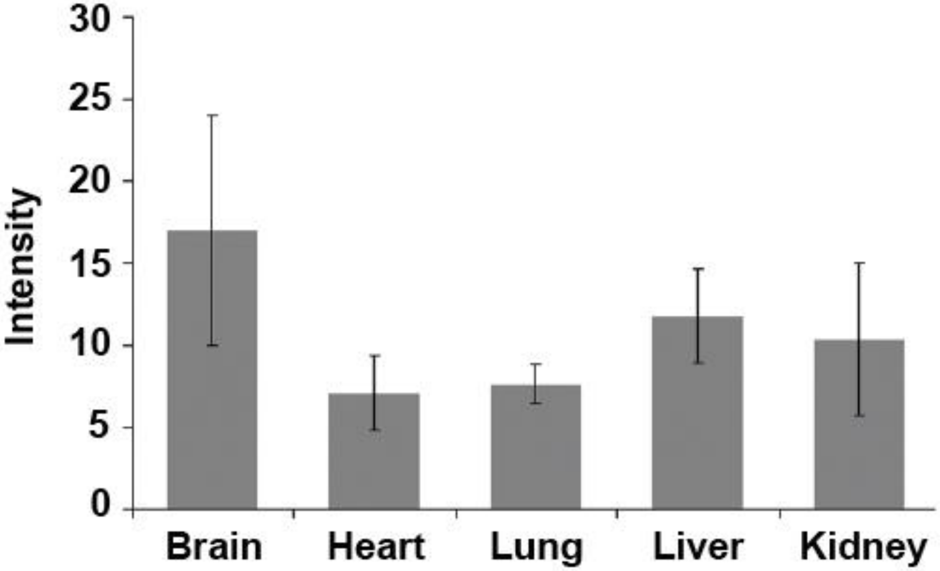


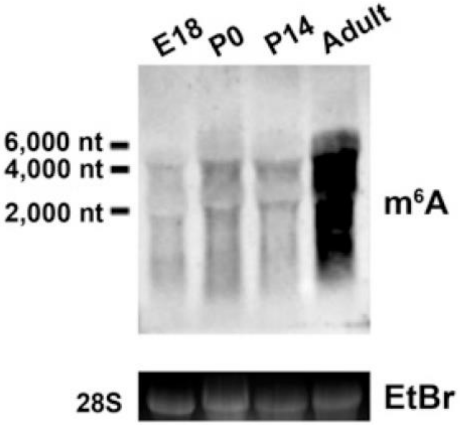
Figure 1 | Examples of RNA modification and demodification that may impact biological regulation. (**a**) Selected examples of RNA base methylation. (**b**) A group of dioxygenases that use iron, α-ketoglutarate and dioxygen to perform oxidation of modified RNA bases for demethylation or hypermodification.

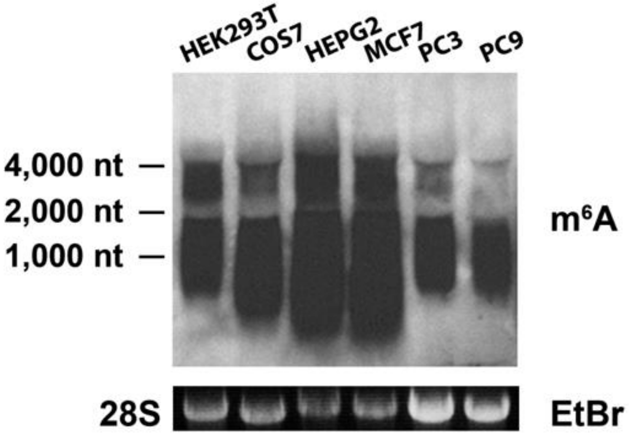
Chuan He

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



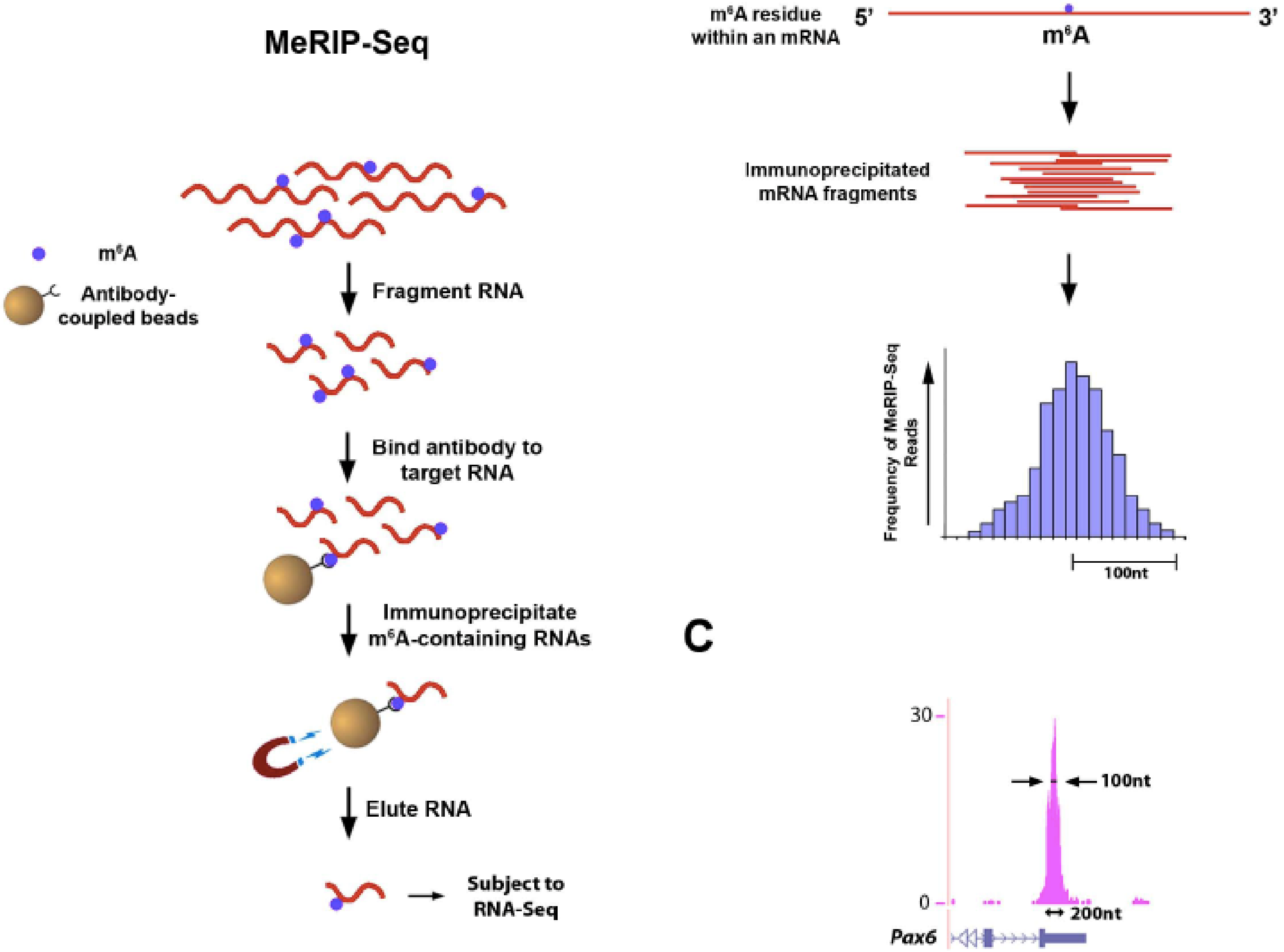






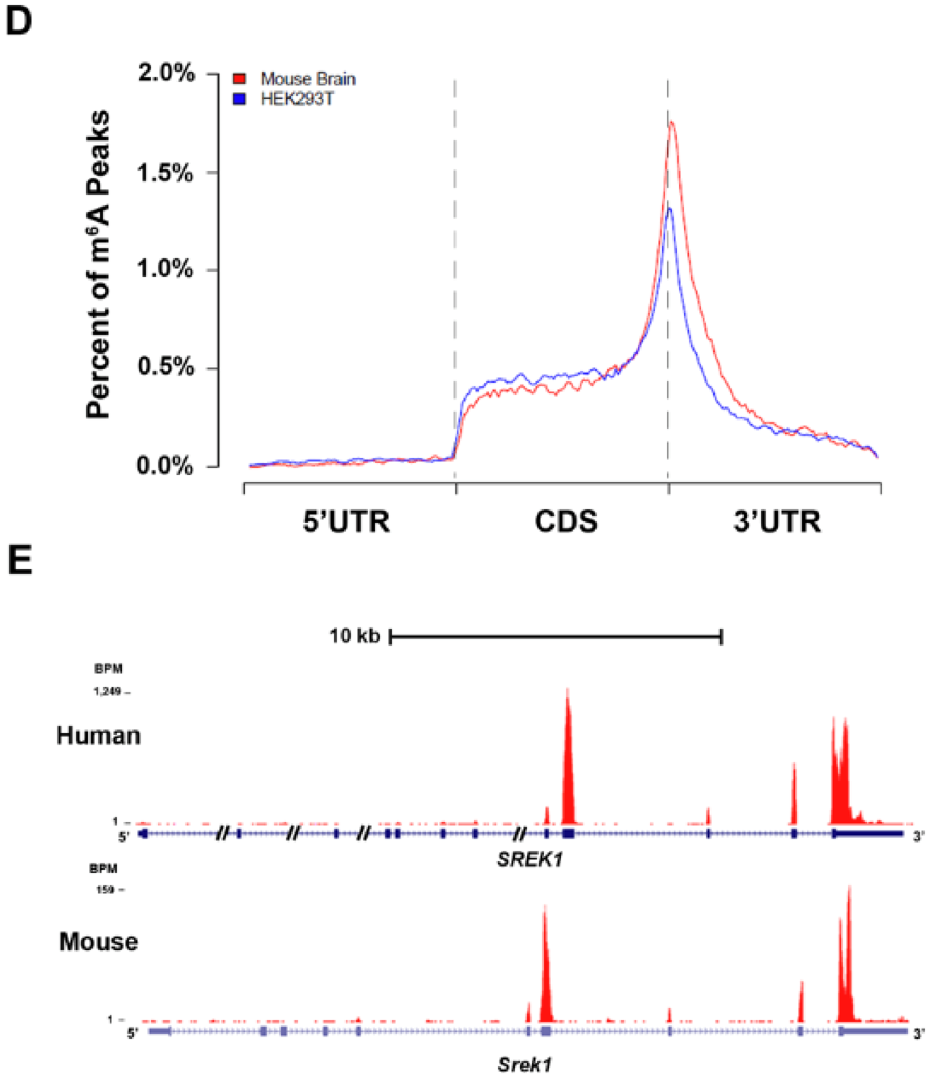
Meyer et al., Cell, 201

A new method: MeRIP-Seq


MeRIP-Seq
m^{6}A
Antibody-coupled beads
Fragment RNA
Bind antibody to target RNA
Immunoprecipitate m^{6}A-containing RNAs
Elute RNA
Subject to RNA-Seq
m^{6}A residue within an mRNA
5'
3'
m^{6}A
Immunoprecipitated mRNA fragments
Frequency of MeRIP-Seq Reads
100nt
C
30
0
100nt
200nt
Pax6

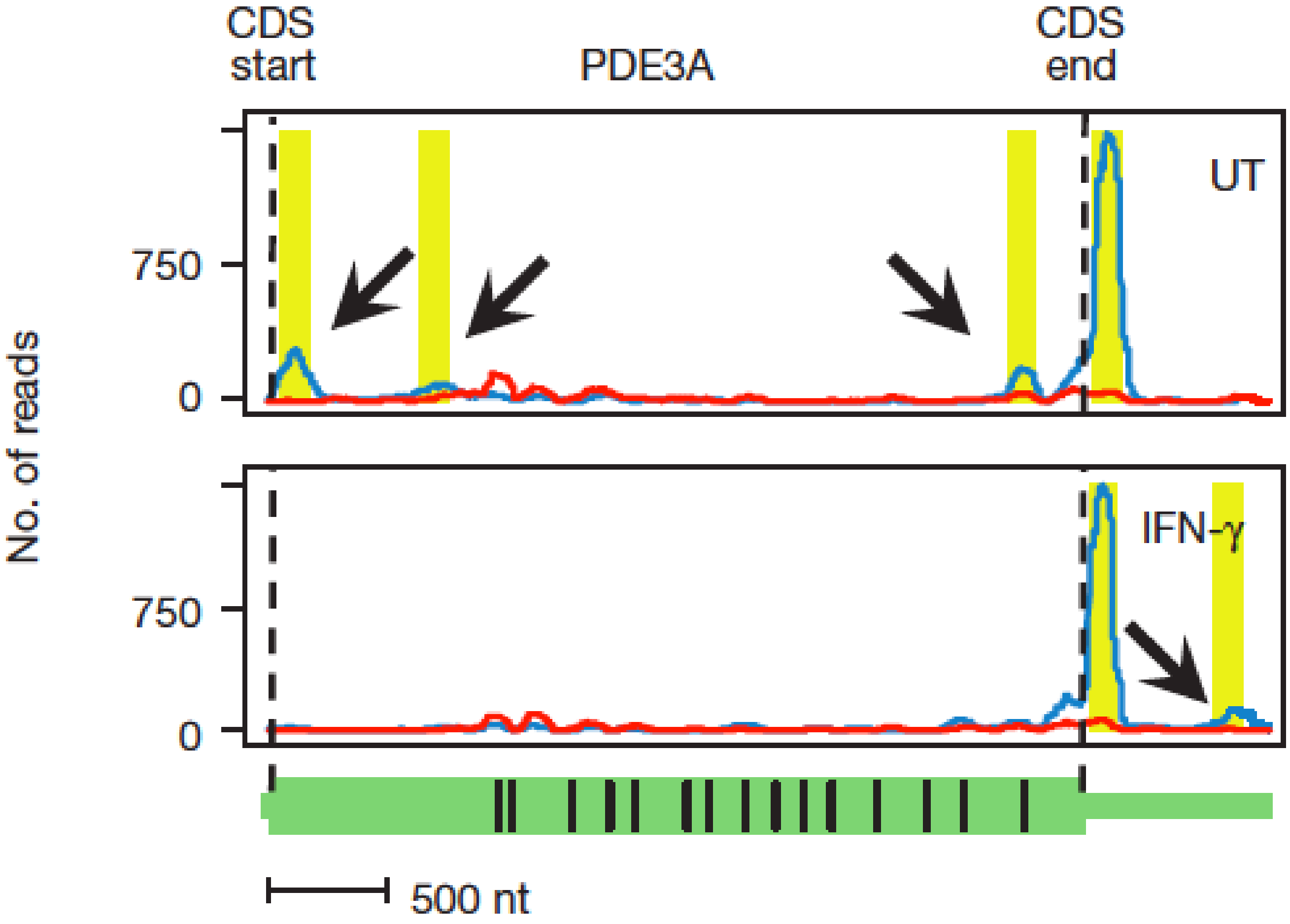
Meyer et al., Cell, 2012

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



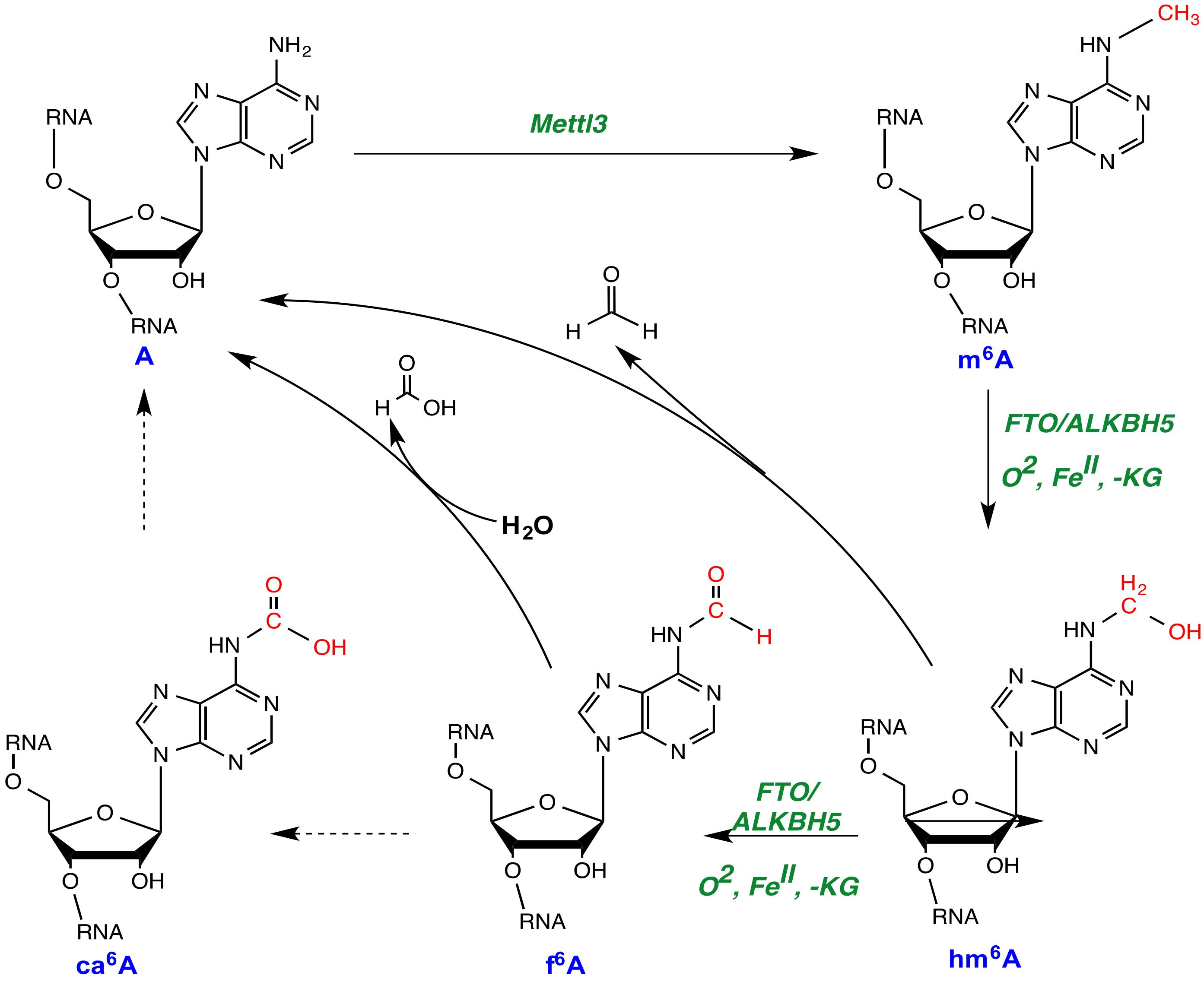
Meyer et al., Cell, 2012

m^6A levels may also change splicing patterns in genes



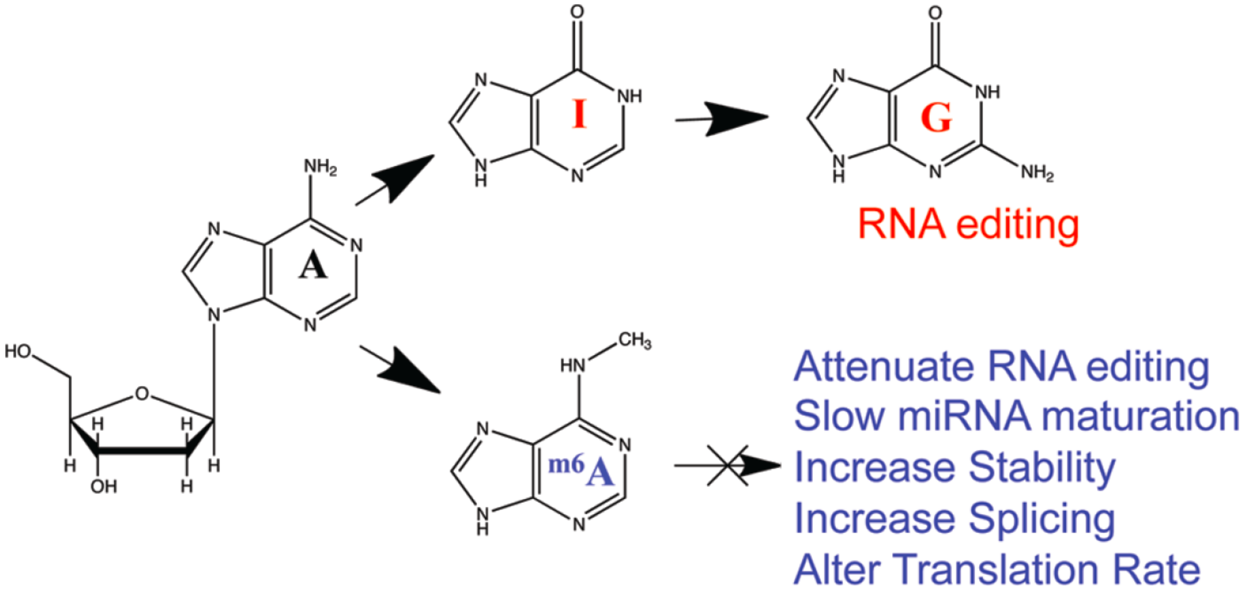
Dominissini *et al.*, Nature, 2012

RNA modifications give a new layer of cellular regulation



Li and Mason, ARGHG, 2014

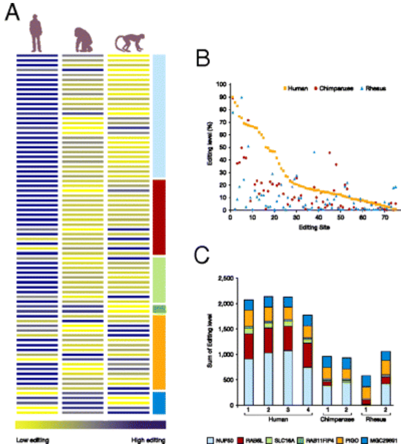
Many putative roles for m^{6}A in RNA



RNA editing

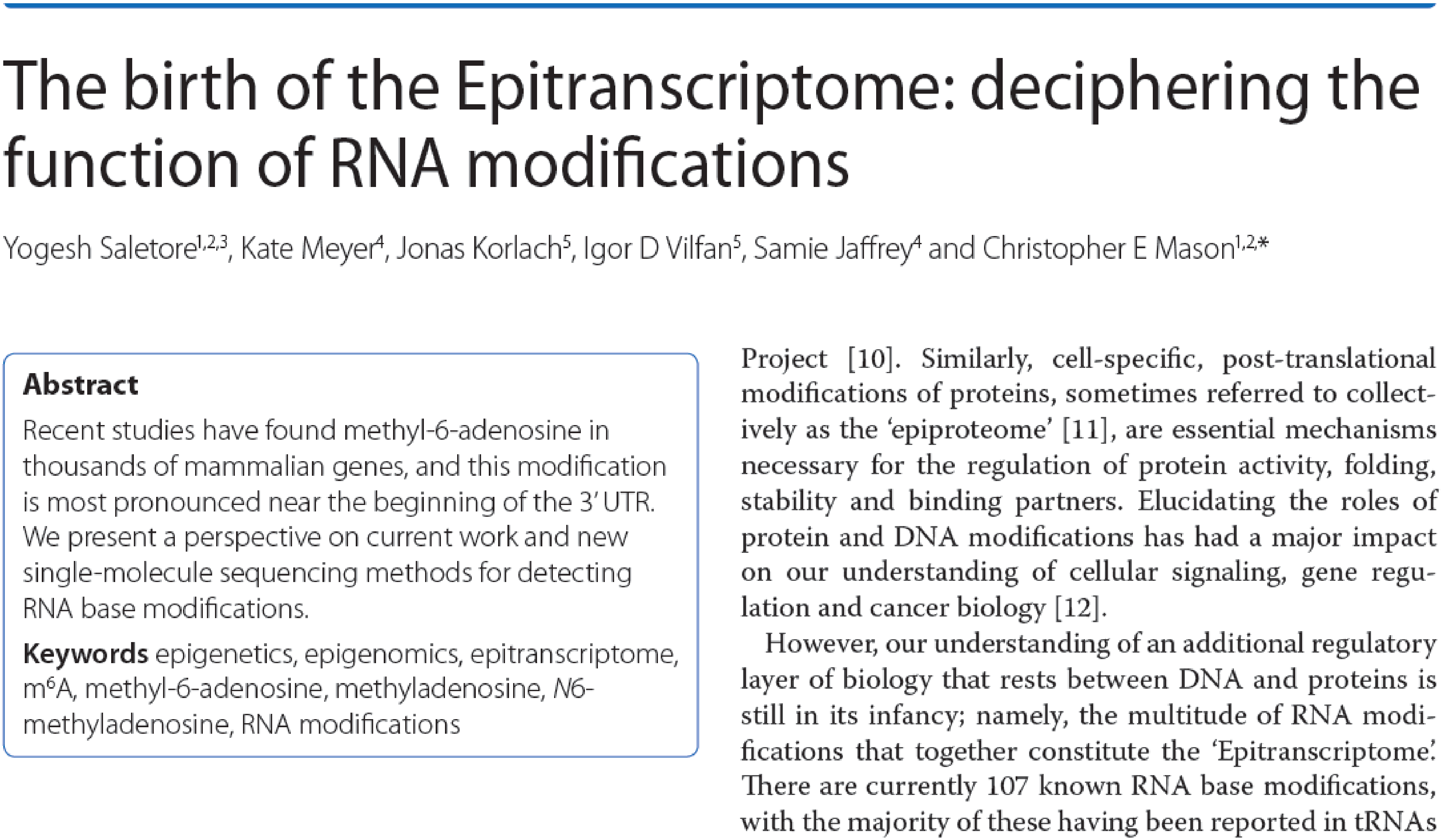
Attenuate RNA editing
Slow miRNA maturation
Increase Stability
Increase Splicing
Alter Translation Rate

Saletore, Chen-Kiang, and Mason, RNA Biology, 2013.



Paz-Yaakov et al, 2010

New layer of regulation to study



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^6A, 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

m^{6}A is just 1 of the 107
known RNA modifications
from the
RNA Modification Database

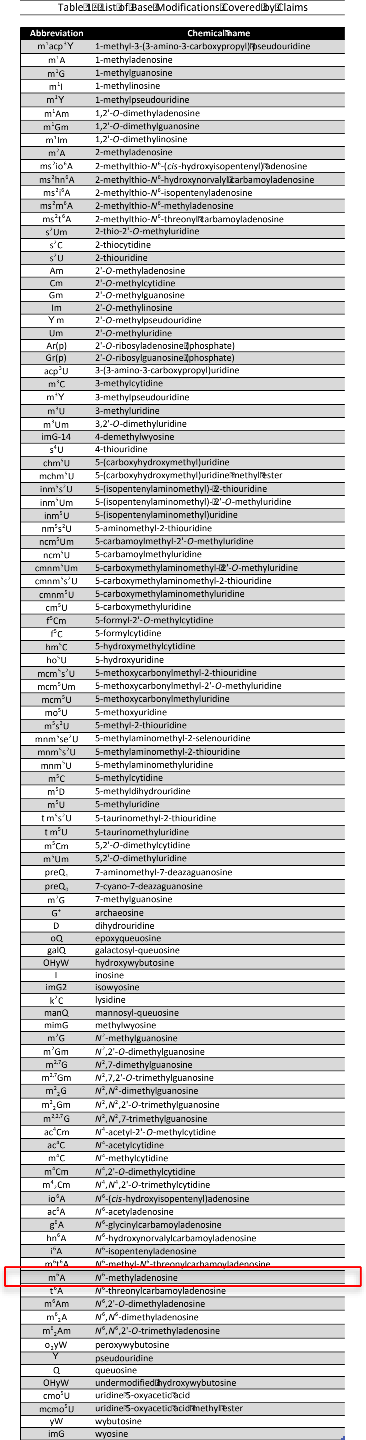
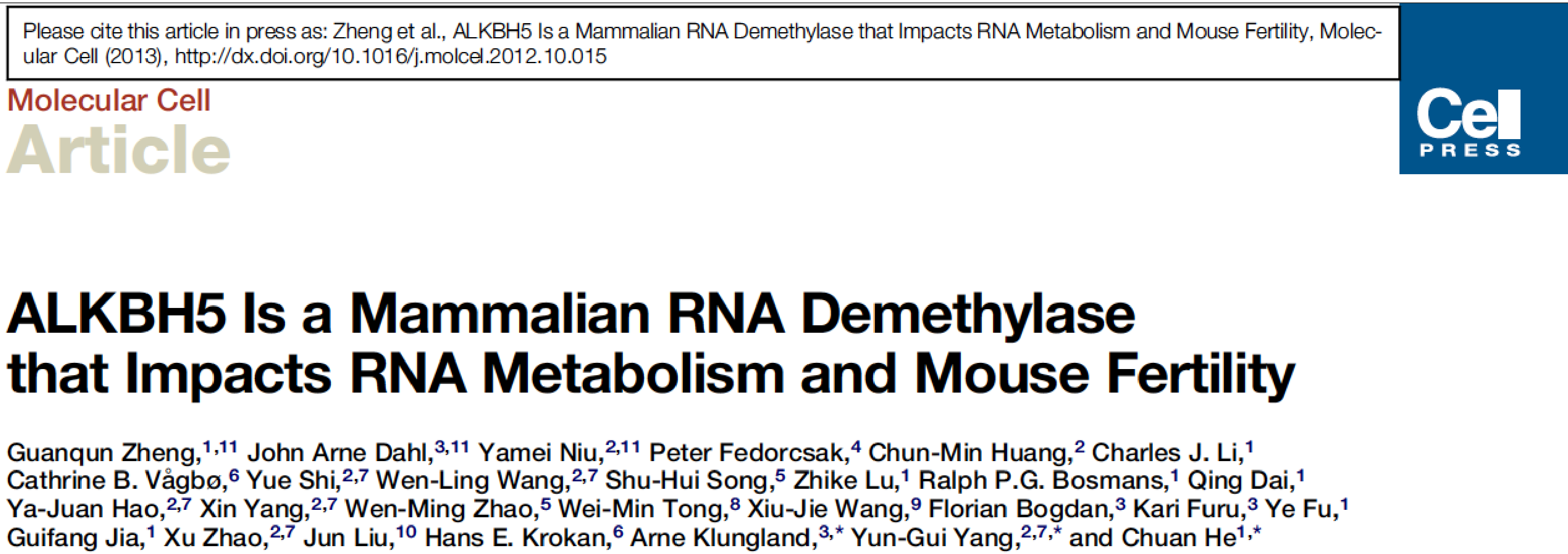


Table 1 List of Base Modifications Covered by Claims

Abbreviation	Chemical name
m^{1}acp^{3}Y	1-methyl-3-(3-amino-3-carboxypropyl) pseudouridine
m^{1}A	1-methyladenosine
m^{1}G	1-methylguanosine
m^{1}I	1-methylinosine
m^{1}Y	1-methylpseudouridine
m^{1}Am	1,2'-*O*-dimethyladenosine
m^{1}Gm	1,2'-*O*-dimethylguanosine
m^{1}Im	1,2'-*O*-dimethylinosine
m^{2}A	2-methyladenosine
ms^{2}io^{6}A	2-methylthio-N^{6}-(*cis*-hydroxyisopentenyl) adenosine
ms^{2}hn^{6}A	2-methylthio-N^{6}-hydroxynorvalyl carbamoyladenosine
ms^{2}i^{6}A	2-methylthio-N^{6}-isopentenyladenosine
ms^{2}m^{6}A	2-methylthio-N^{6}-methyladenosine
ms^{2}t^{6}A	2-methylthio-N^{6}-threonyl carbamoyladenosine
s^{2}Um	2-thio-2'-*O*-methyluridine
s^{2}C	2-thiocytidine
s^{2}U	2-thiouridine
Am	2'-*O*-methyladenosine
Cm	2'-*O*-methylcytidine
Gm	2'-*O*-methylguanosine
Im	2'-*O*-methylinosine
Y m	2'-*O*-methylpseudouridine
Um	2'-*O*-methyluridine
Ar(p)	2'-*O*-ribosyladenosine (phosphate)
Gr(p)	2'-*O*-ribosylguanosine (phosphate)
acp^{3}U	3-(3-amino-3-carboxypropyl)uridine
m^{3}C	3-methylcytidine
m^{3}Y	3-methylpseudouridine
m^{3}U	3-methyluridine
m^{3}Um	3,2'-*O*-dimethyluridine
imG-14	4-demethylwyosine
s^{4}U	4-thiouridine
chm^{5}U	5-(carboxyhydroxymethyl)uridine
mchm^{5}U	5-(carboxyhydroxymethyl)uridine methyl ester
inm^{5}s^{2}U	5-(isopentenylaminomethyl)- 2-thiouridine
inm^{5}Um	5-(isopentenylaminomethyl)- 2'-*O*-methyluridine
inm^{5}U	5-(isopentenylaminomethyl)uridine
nm^{5}s^{2}U	5-aminomethyl-2-thiouridine
ncm^{5}Um	5-carbamoylmethyl-2'-*O*-methyluridine
ncm^{5}U	5-carbamoylmethyluridine
cmnm^{5}Um	5-carboxymethylaminomethyl- 2'-*O*-methyluridine
cmnm^{5}s^{2}U	5-carboxymethylaminomethyl-2-thiouridine
cmnm^{5}U	5-carboxymethylaminomethyluridine
cm^{5}U	5-carboxymethyluridine
f^{5}Cm	5-formyl-2'-*O*-methylcytidine
f^{5}C	5-formylcytidine
hm^{5}C	5-hydroxymethylcytidine
ho^{5}U	5-hydroxyuridine
mcm^{5}s^{2}U	5-methoxycarbonylmethyl-2-thiouridine
mcm^{5}Um	5-methoxycarbonylmethyl-2'-*O*-methyluridine
mcm^{5}U	5-methoxycarbonylmethyluridine
mo^{5}U	5-methoxyuridine
m^{5}s^{2}U	5-methyl-2-thiouridine
mnm^{5}se^{2}U	5-methylaminomethyl-2-selenouridine
mnm^{5}s^{2}U	5-methylaminomethyl-2-thiouridine
mnm^{5}U	5-methylaminomethyluridine
m^{5}C	5-methylcytidine
m^{5}D	5-methyldihydrouridine
m^{5}U	5-methyluridine
t m^{5}s^{2}U	5-taurinomethyl-2-thiouridine
t m^{5}U	5-taurinomethyluridine
m^{5}Cm	5,2'-*O*-dimethylcytidine
m^{5}Um	5,2'-*O*-dimethyluridine
preQ$_{1}$	7-aminomethyl-7-deazaguanosine
preQ$_{0}$	7-cyano-7-deazaguanosine
m^{7}G	7-methylguanosine
G^{+}	archaeosine
D	dihydrouridine
oQ	epoxyqueuosine
galQ	galactosyl-queuosine
OHyW	hydroxywybutosine
I	inosine
imG2	isowyosine
k^{2}C	lysidine
manQ	mannosyl-queuosine
mimG	methylwyosine
m^{2}G	N^{2}-methylguanosine
m^{2}Gm	N^{2},2'-*O*-dimethylguanosine
m^{2,7}G	N^{2},7-dimethylguanosine
m^{2,7}Gm	N^{2},7,2'-*O*-trimethylguanosine
m$^{2}_{2}$G	N^{2},N^{2}-dimethylguanosine
m$^{2}_{2}$Gm	N^{2},N^{2},2'-*O*-trimethylguanosine
m^{2,2,7}G	N^{2},N^{2},7-trimethylguanosine
ac^{4}Cm	N^{4}-acetyl-2'-*O*-methylcytidine
ac^{4}C	N^{4}-acetylcytidine
m^{4}C	N^{4}-methylcytidine
m^{4}Cm	N^{4},2'-*O*-dimethylcytidine
m$^{4}_{2}$Cm	N^{4},N^{4},2'-*O*-trimethylcytidine
io^{6}A	N^{6}-(*cis*-hydroxyisopentenyl)adenosine
ac^{6}A	N^{6}-acetyladenosine
g^{6}A	N^{6}-glycinylcarbamoyladenosine
hn^{6}A	N^{6}-hydroxynorvalylcarbamoyladenosine
i^{6}A	N^{6}-isopentenyladenosine
m^{6}t^{6}A	N^{6}-methyl-N^{6}-threonylcarbamoyladenosine
m^{6}A	N^{6}-methyladenosine
t^{6}A	N^{6}-threonylcarbamoyladenosine
m^{6}Am	N^{6},2'-*O*-dimethyladenosine
m$^{6}_{2}$A	N^{6},N^{6}-dimethyladenosine
m$^{6}_{2}$Am	N^{6},N^{6},2'-*O*-trimethyladenosine
o$_{2}$yW	peroxywybutosine
Y	pseudouridine
Q	queuosine
OHyW	undermodified hydroxywybutosine
cmo^{5}U	uridine 5-oxyacetic acid
mcmo^{5}U	uridine 5-oxyacetic acid methyl ester
yW	wybutosine
imG	wyosine



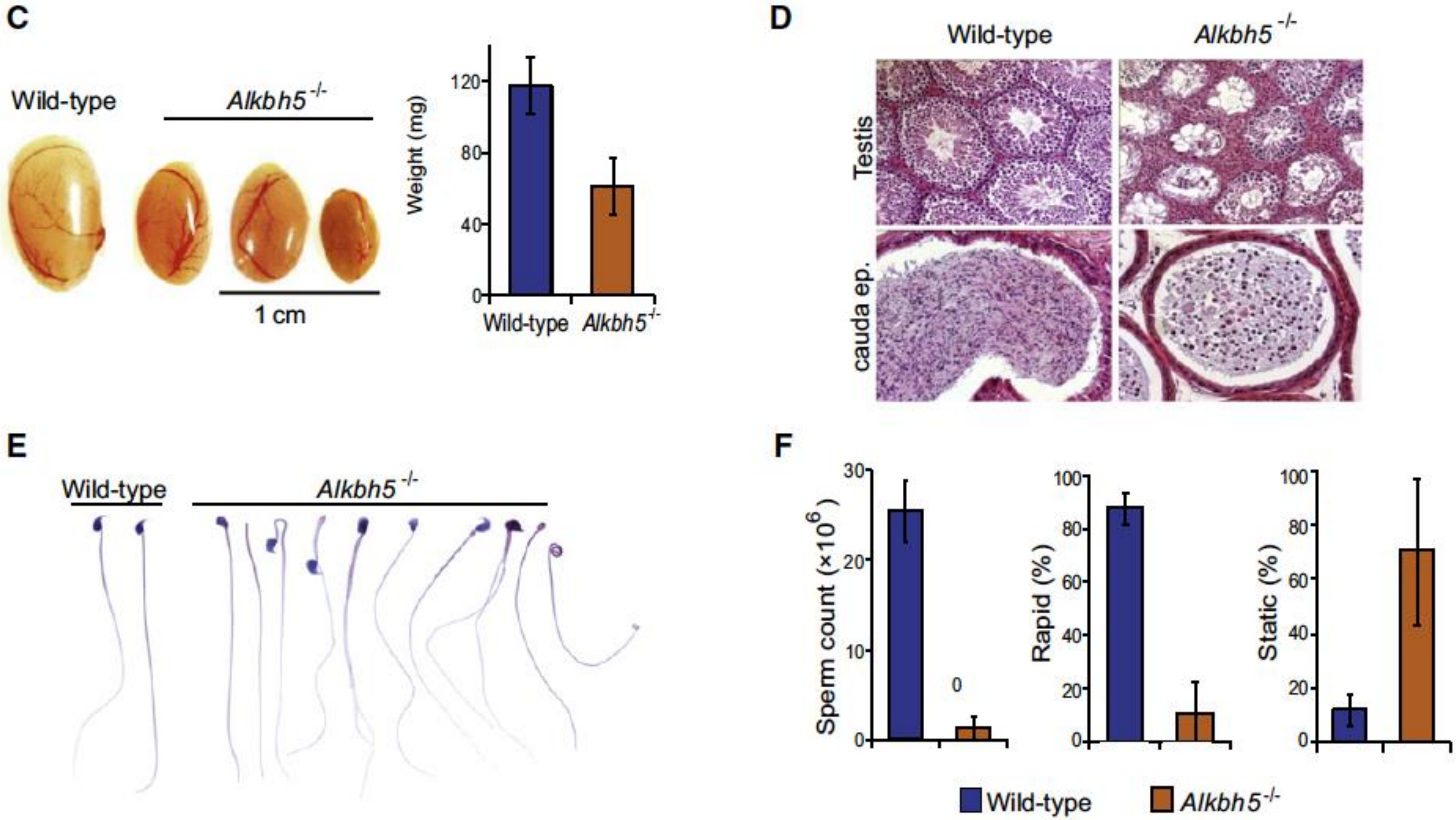
Please cite this article in press as: Zheng et al., ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2012.10.015

Molecular Cell
Article

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

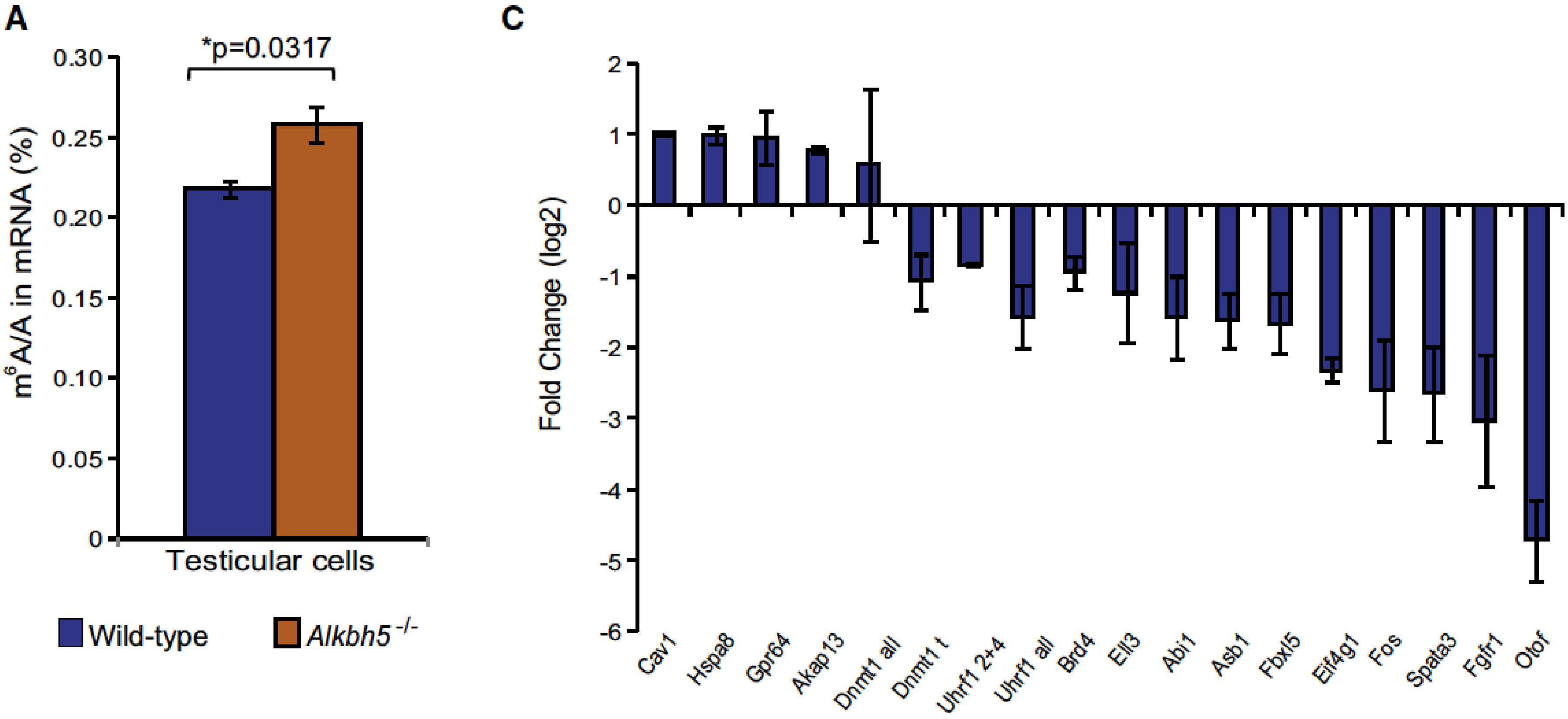
Guanqun Zheng,[1,11] John Arne Dahl,[3,11] Yamei Niu,[2,11] Peter Fedorcsak,[4] Chun-Min Huang,[2] Charles J. Li,[1] Cathrine B. Vågbø,[6] Yue Shi,[2,7] Wen-Ling Wang,[2,7] Shu-Hui Song,[5] Zhike Lu,[1] Ralph P.G. Bosmans,[1] Qing Dai,[1] Ya-Juan Hao,[2,7] Xin Yang,[2,7] Wen-Ming Zhao,[5] Wei-Min Tong,[8] 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

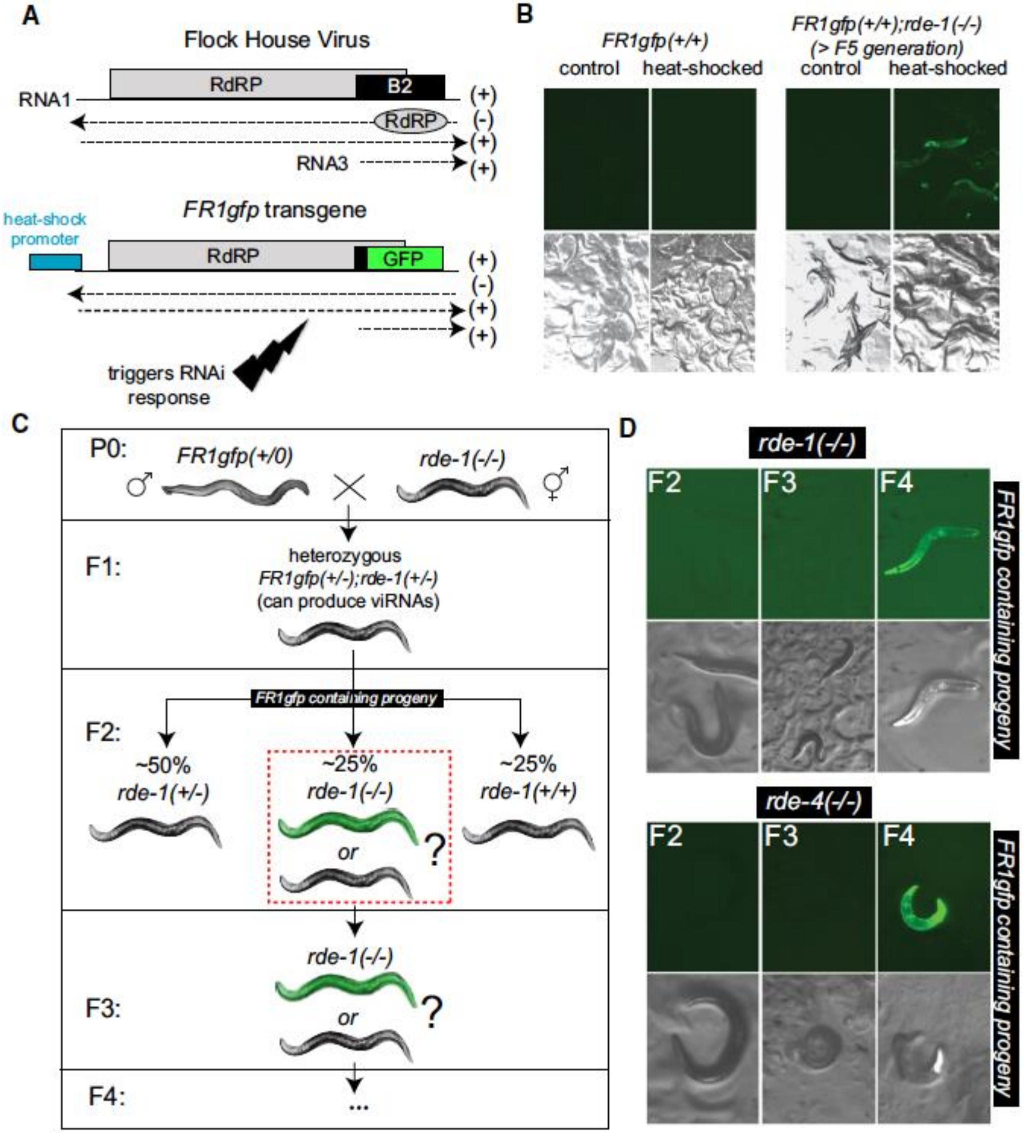


Zheng et al., Mol. Cell, 2013

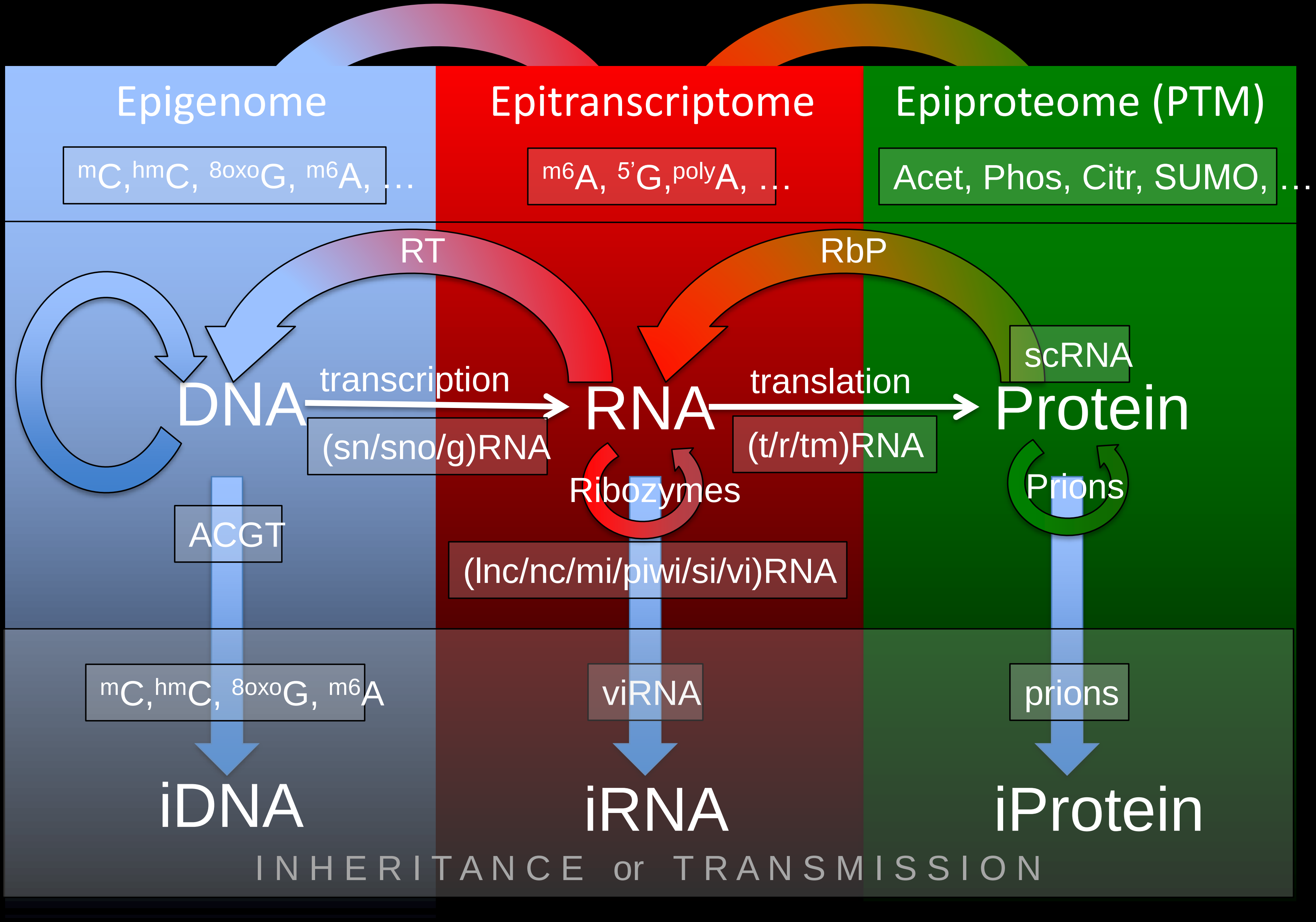
Dysregulated m^{6}A affects many epigenetic modifiers



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



Rechavi et al, 2011



Epigenome

^{m}C, ^{hm}C, ^{8oxo}G, ^{m6}A, …

Epitranscriptome

^{m6}A, $^{5'}$G, ^{poly}A, …

Epiproteome (PTM)

Acet, Phos, Citr, SUMO, …

RT

RbP

DNA

transcription

RNA

translation

Protein

scRNA

(sn/sno/g)RNA

(t/r/tm)RNA

Ribozymes

Prions

ACGT

(lnc/nc/mi/piwi/si/vi)RNA

^{m}C, ^{hm}C, ^{8oxo}G, ^{m6}A

viRNA

prions

iDNA

iRNA

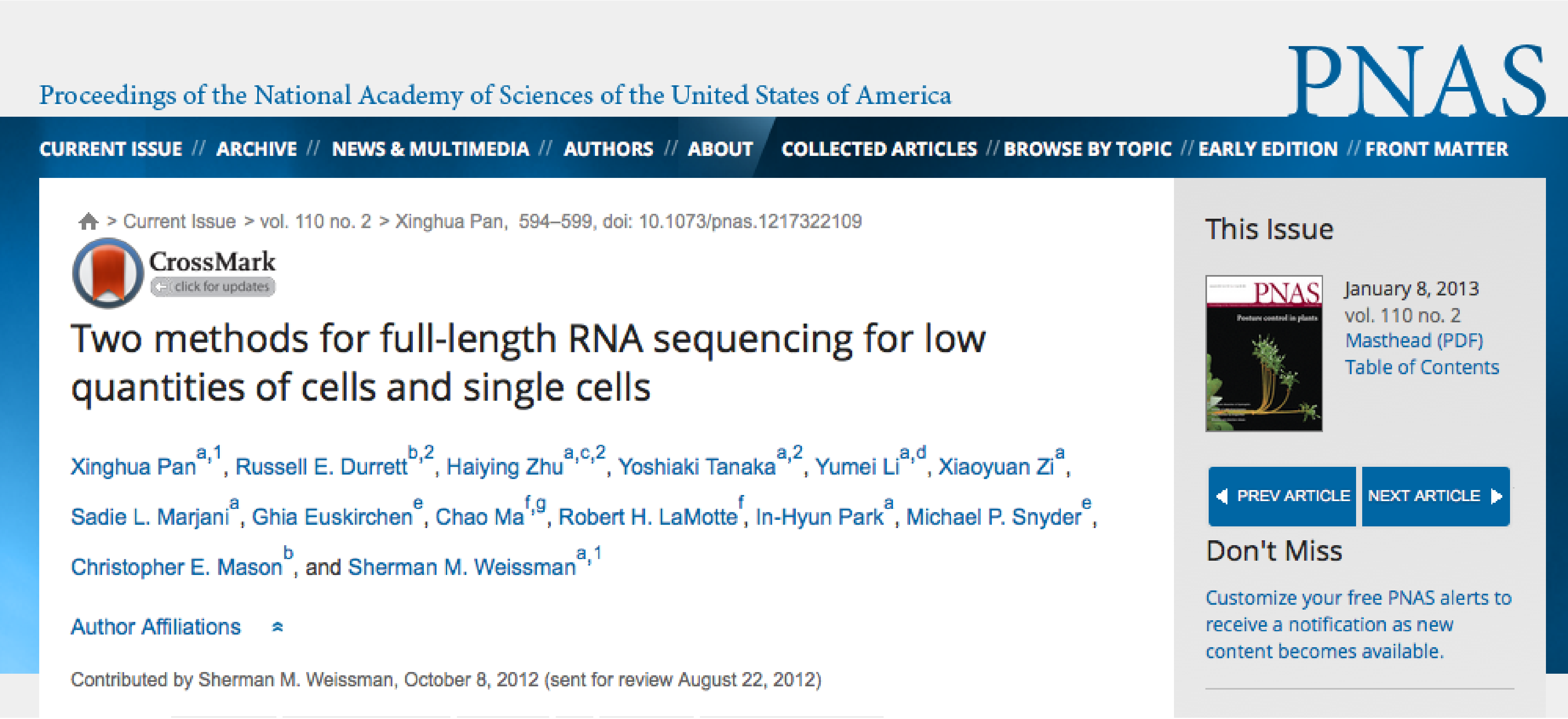
iProtein

INHERITANCE or TRANSMISSION

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

The Era of Single Cells

It used to be very hard to look at individual cells


Proceedings of the National Academy of Sciences of the United States of America
PNAS
CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // AUTHORS // ABOUT
COLLECTED ARTICLES // BROWSE BY TOPIC // EARLY EDITION // FRONT MATTER
CrossMark
click for updates
PREV ARTICLE
NEXT ARTICLE

> Current Issue > vol. 110 no. 2 > Xinghua Pan, 594–599, doi: 10.1073/pnas.1217322109

Two methods for full-length RNA sequencing for low quantities of cells and single cells

Xinghua Pan[a,1], Russell E. Durrett[b,2], Haiying Zhu[a,c,2], Yoshiaki Tanaka[a,2], Yumei Li[a,d], Xiaoyuan Zi[a], Sadie L. Marjani[a], Ghia Euskirchen[e], Chao Ma[f,g], Robert H. LaMotte[f], In-Hyun Park[a], Michael P. Snyder[e], Christopher E. Mason[b], and Sherman M. Weissman[a,1]

Author Affiliations

Contributed by Sherman M. Weissman, October 8, 2012 (sent for review August 22, 2012)

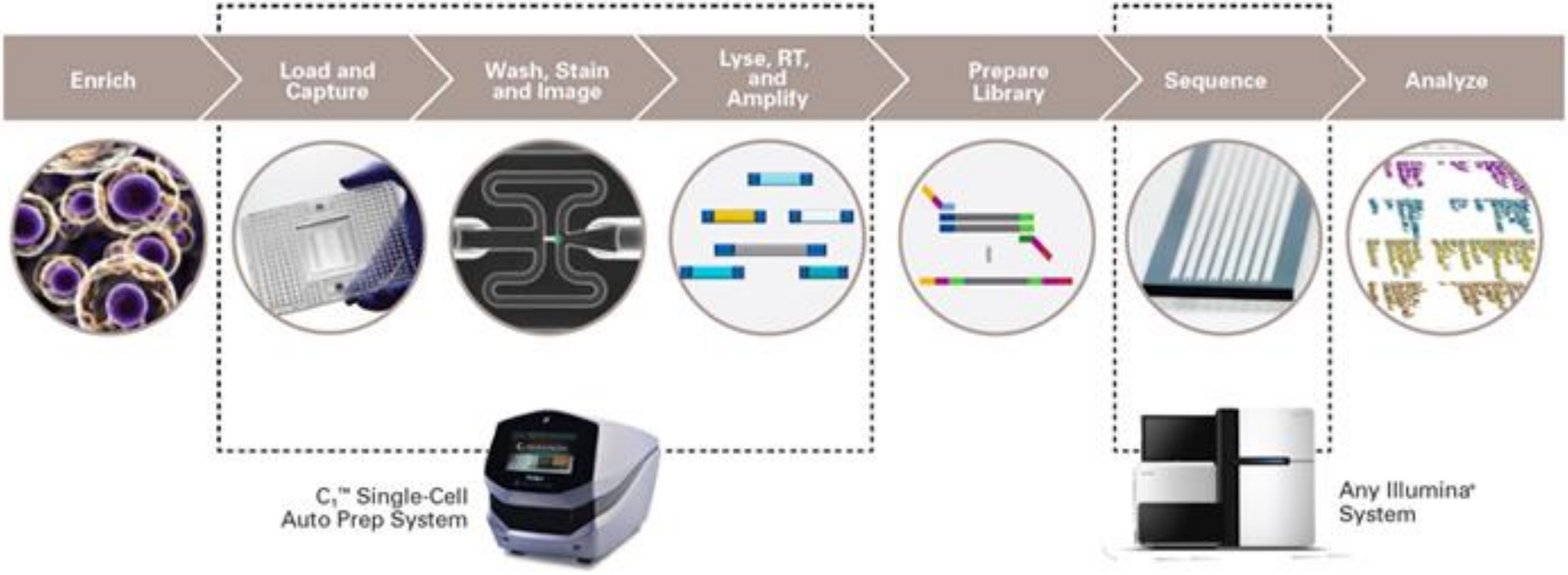
This Issue

January 8, 2013
vol. 110 no. 2
Masthead (PDF)
Table of Contents

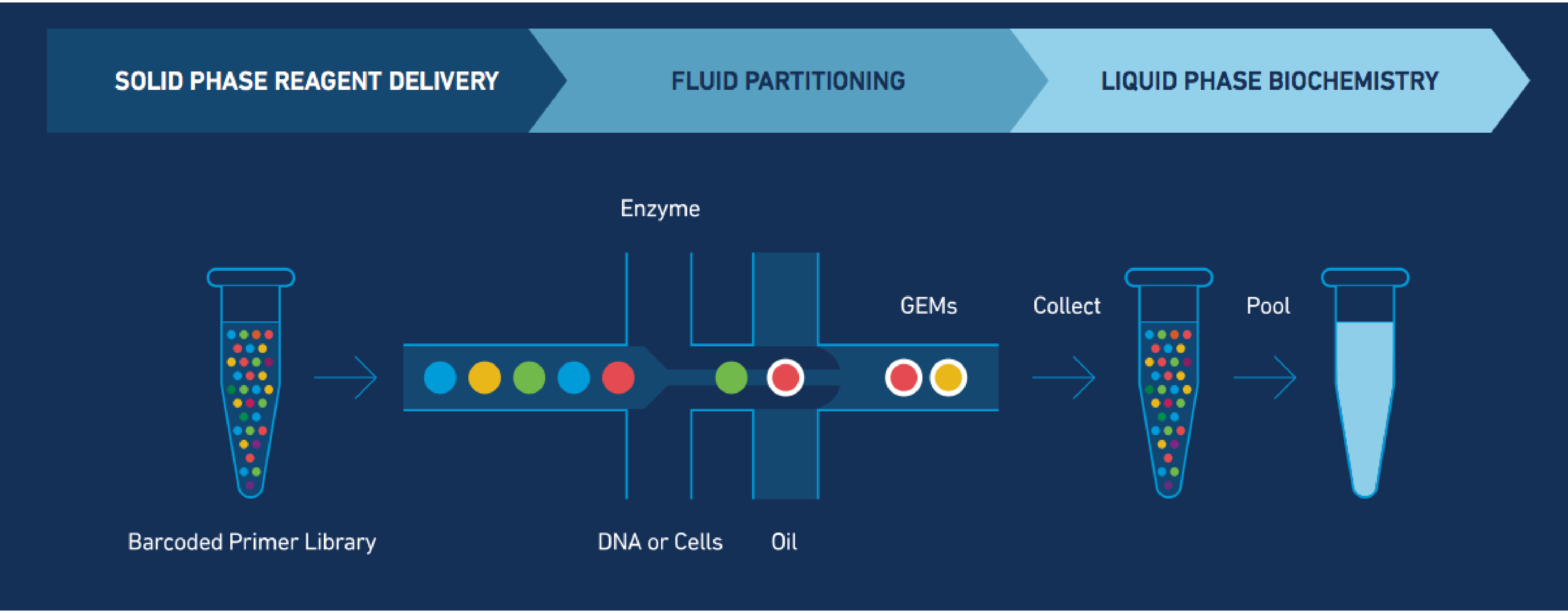
Don't Miss

Customize your free PNAS alerts to receive a notification as new content becomes available.

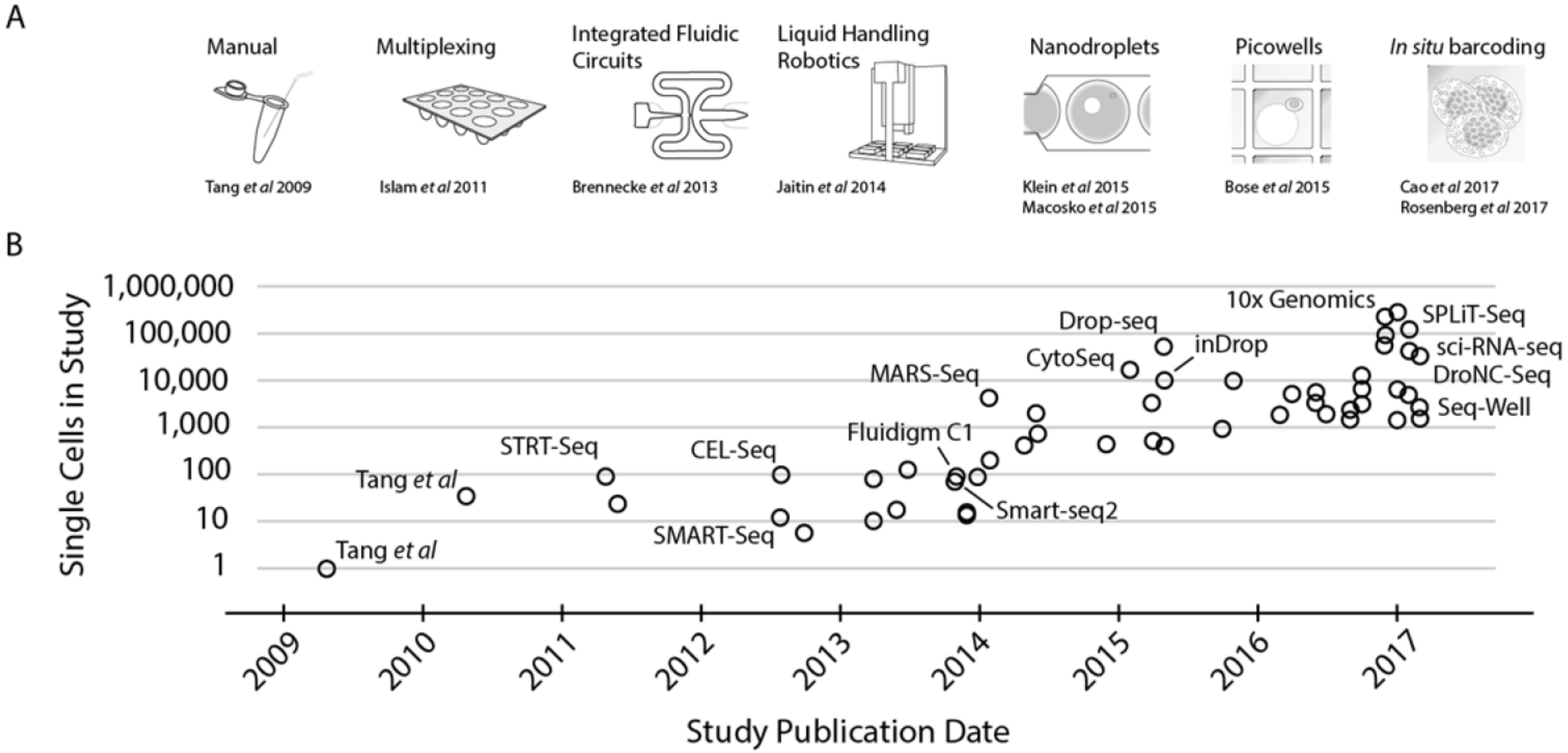
But now it’s very easy – Fluidigm C1



10X Genomics Single-Cell



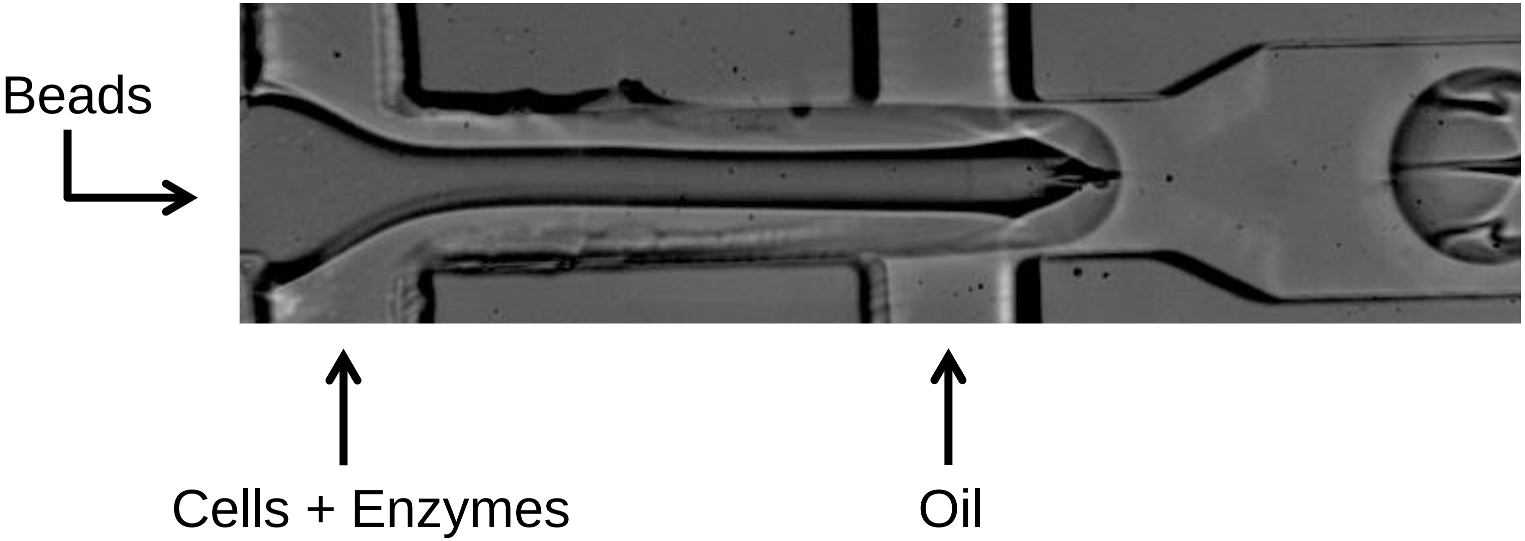
The explosion of scRNA-seq experiments



Svennson *et al*., 2017

Single cell capture and RNA chemistry using nanodroplets

- Drop-seq



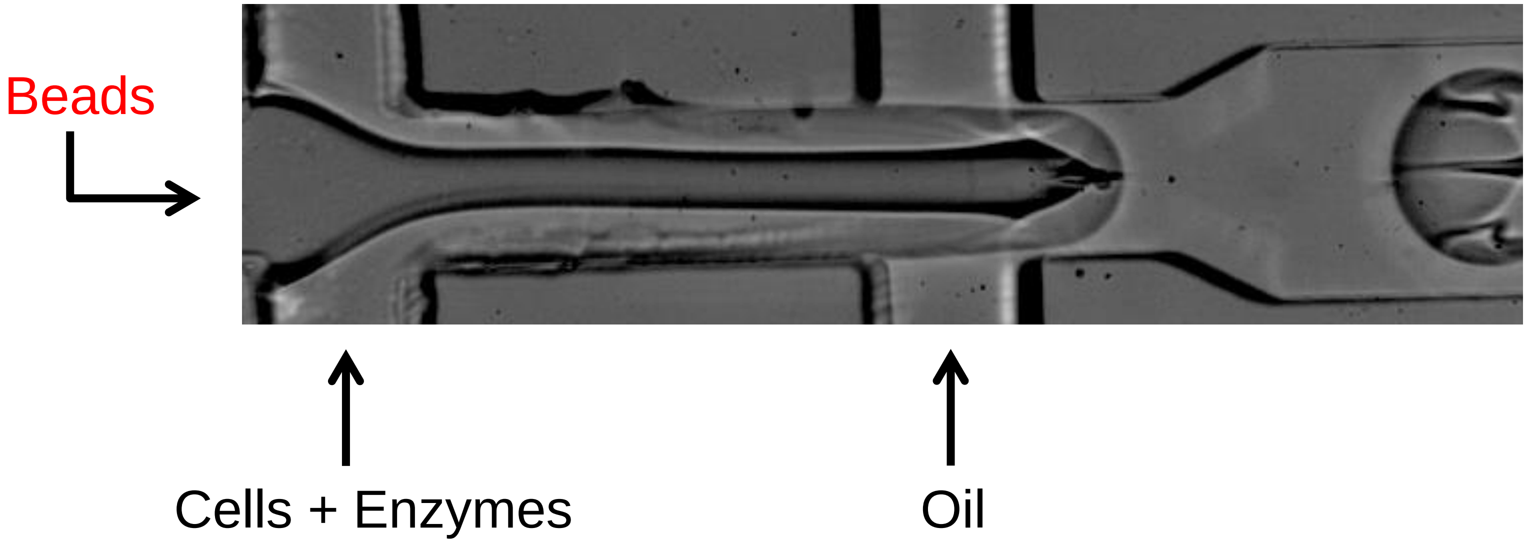
Beads

Cells + Enzymes

Oil

http://mccarrolllab.com/dropseq/
Macosko *et al.*, Cell 2015

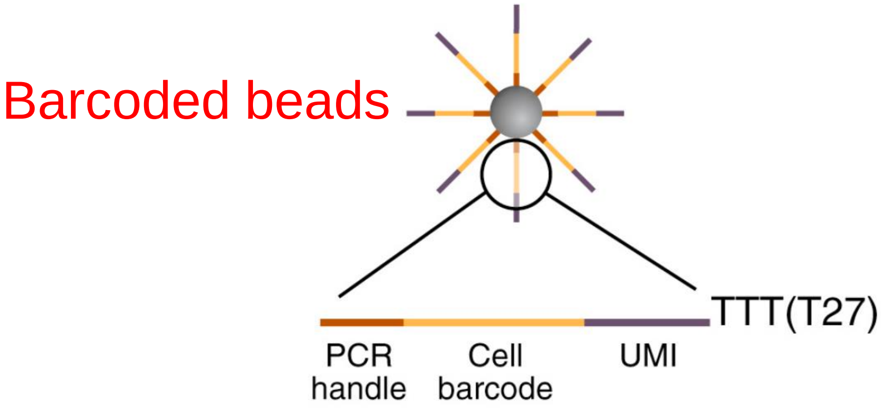
Single cell capture and RNA chemistry using nanodroplets



Beads

Cells + Enzymes

Oil



Barcoded beads

TTT(T27)

PCR handle

Cell barcode

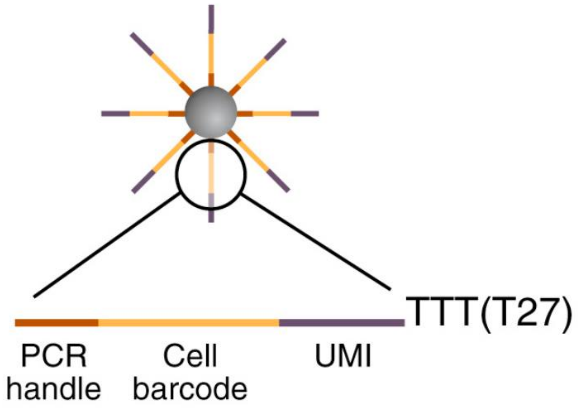
UMI

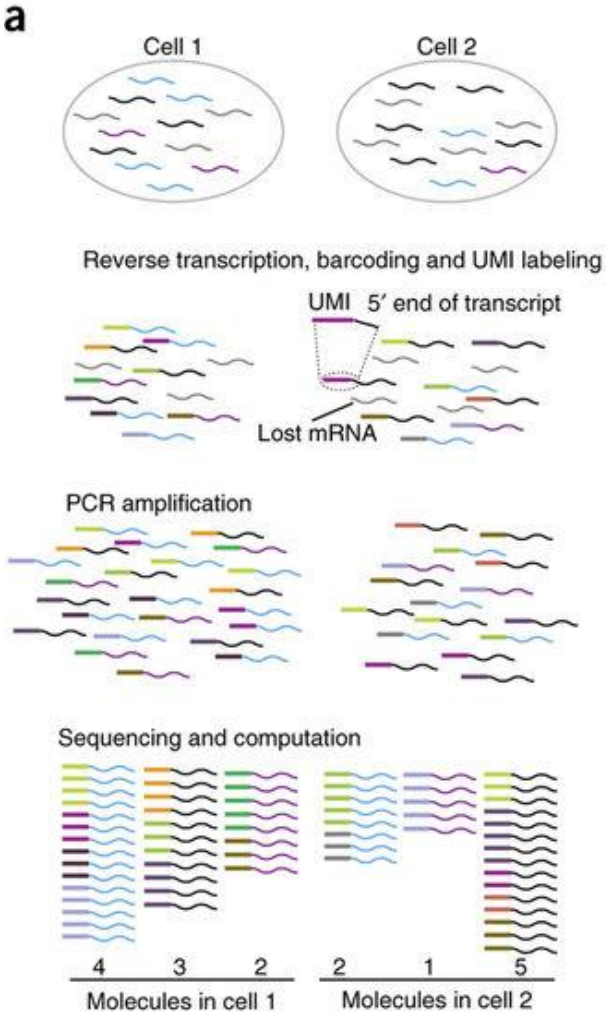
http://mccarrolllab.com/dropseq/

Macosko *et al.*, Cell 2015

Unique Molecular Identifiers (UMIs)

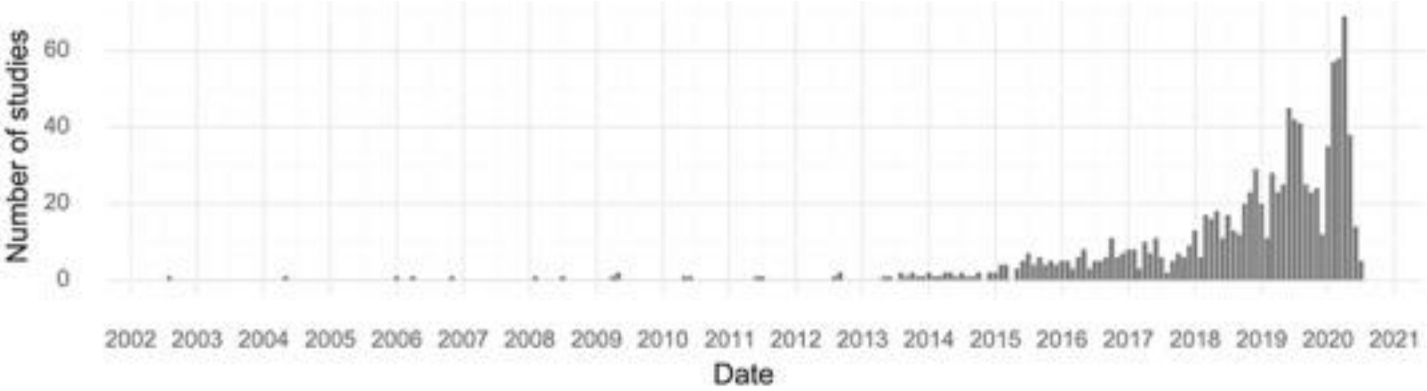
Barcoded beads

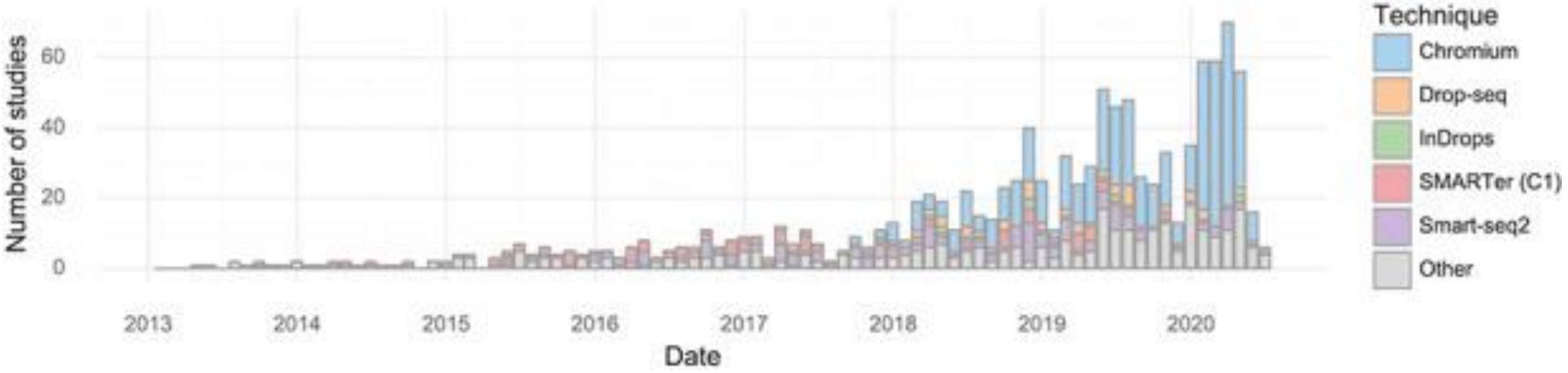




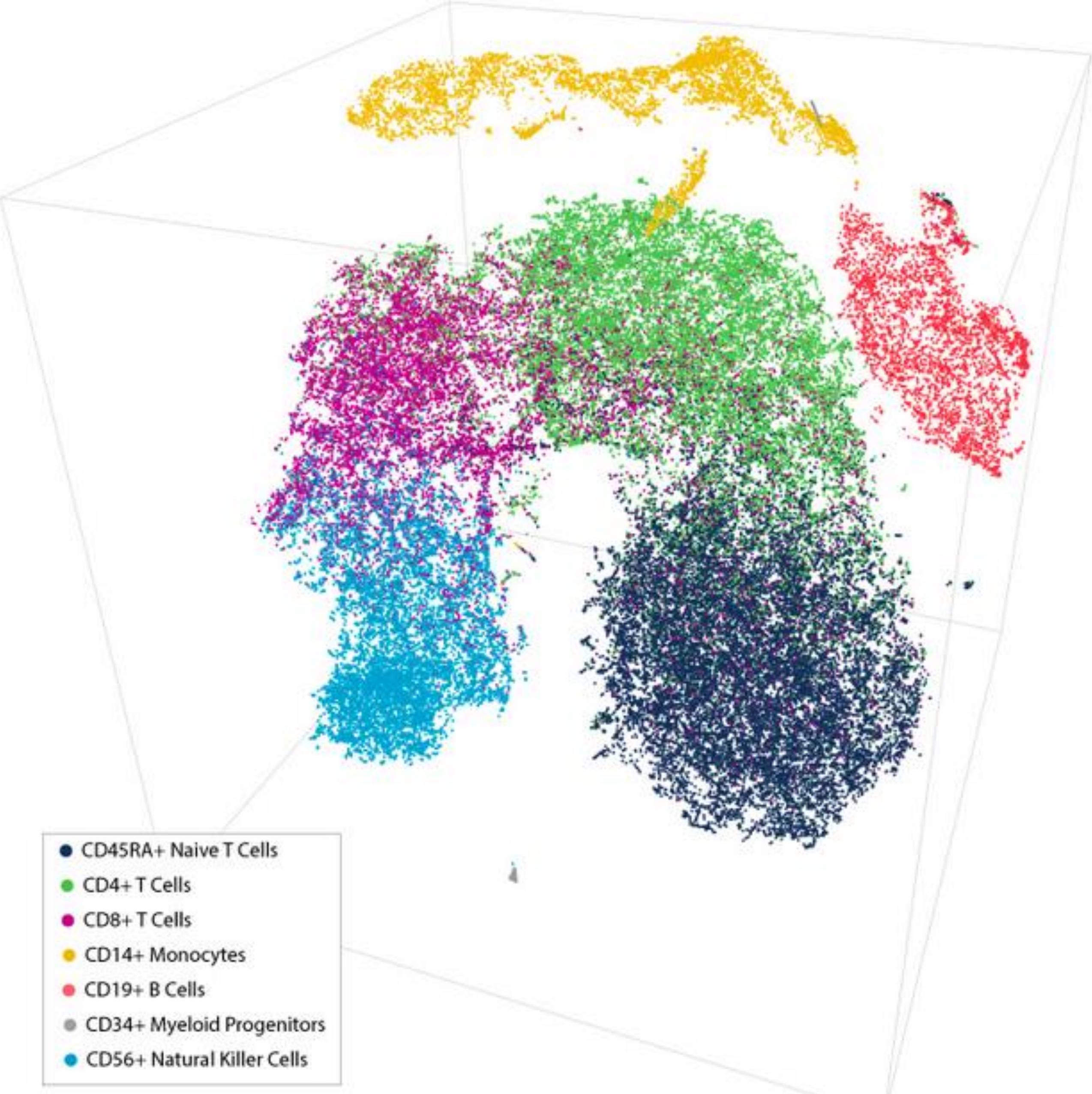
Islam *et al.*, Nature Methods 2014

Clear increase over time

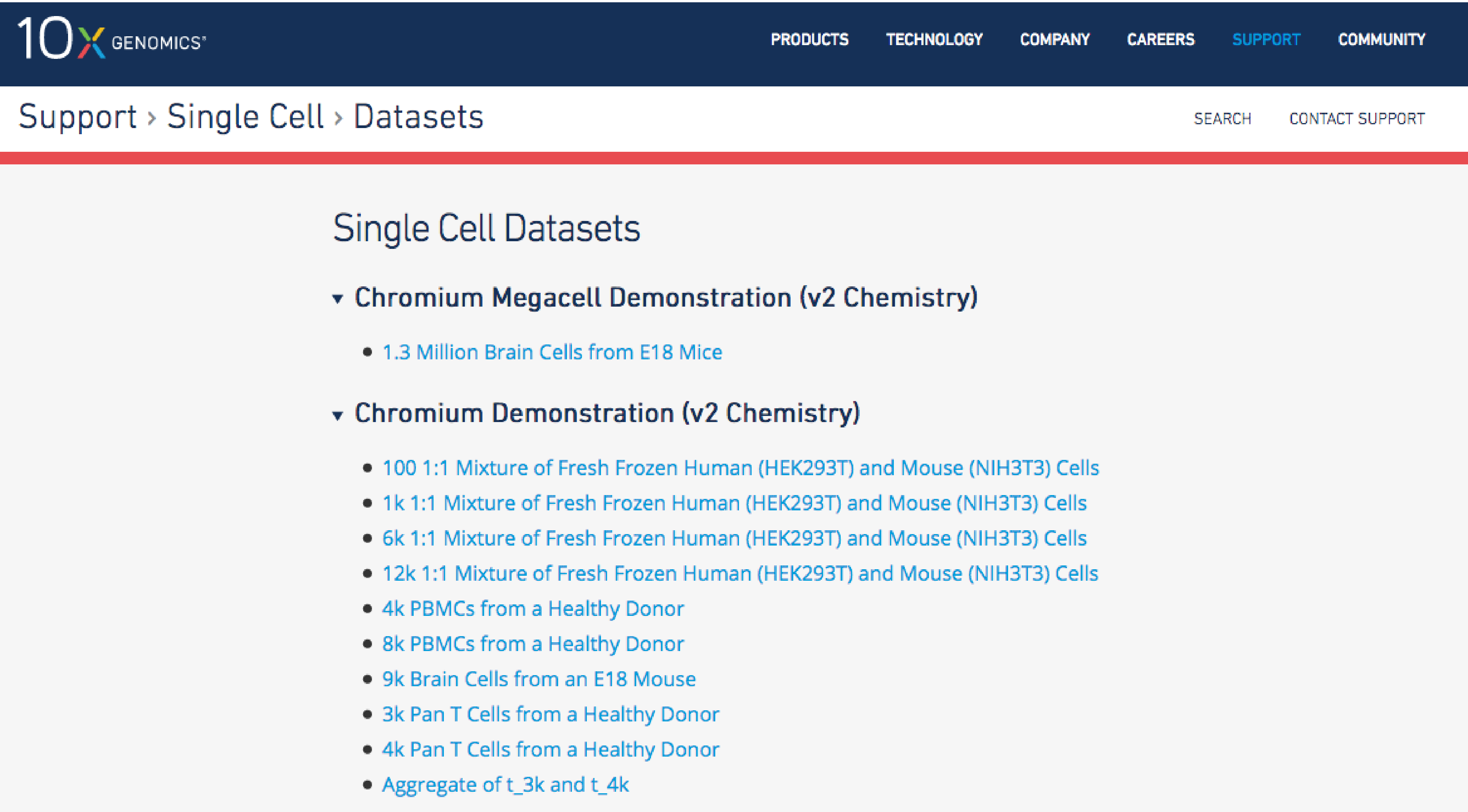




https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7698659/


● CD45RA+ Naive T Cells
● CD4+ T Cells
● CD8+ T Cells
● CD14+ Monocytes
● CD19+ B Cells
● CD34+ Myeloid Progenitors
● CD56+ Natural Killer Cells

1.3 million neurons catalogued



Single Cell Datasets

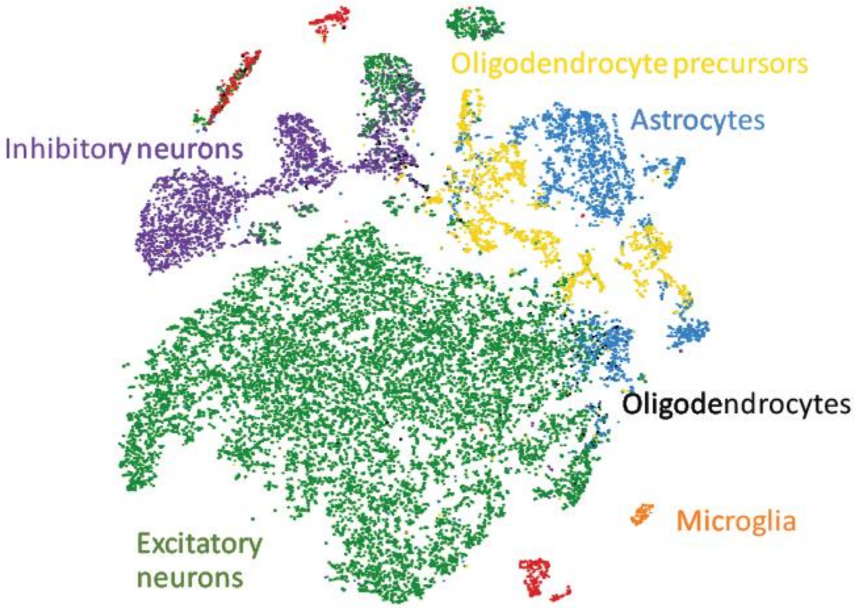
▾ Chromium Megacell Demonstration (v2 Chemistry)

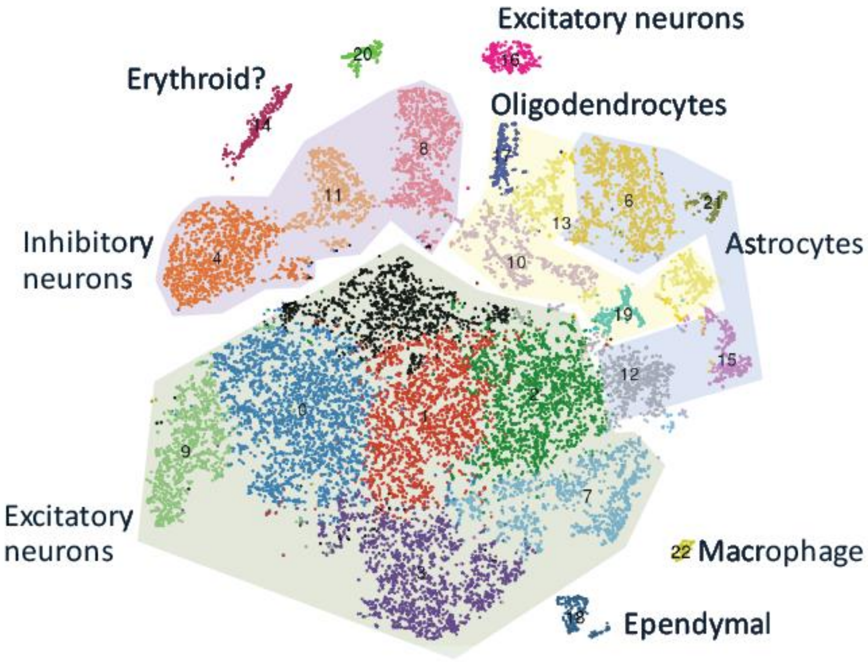
- 1.3 Million Brain Cells from E18 Mice

▾ Chromium Demonstration (v2 Chemistry)

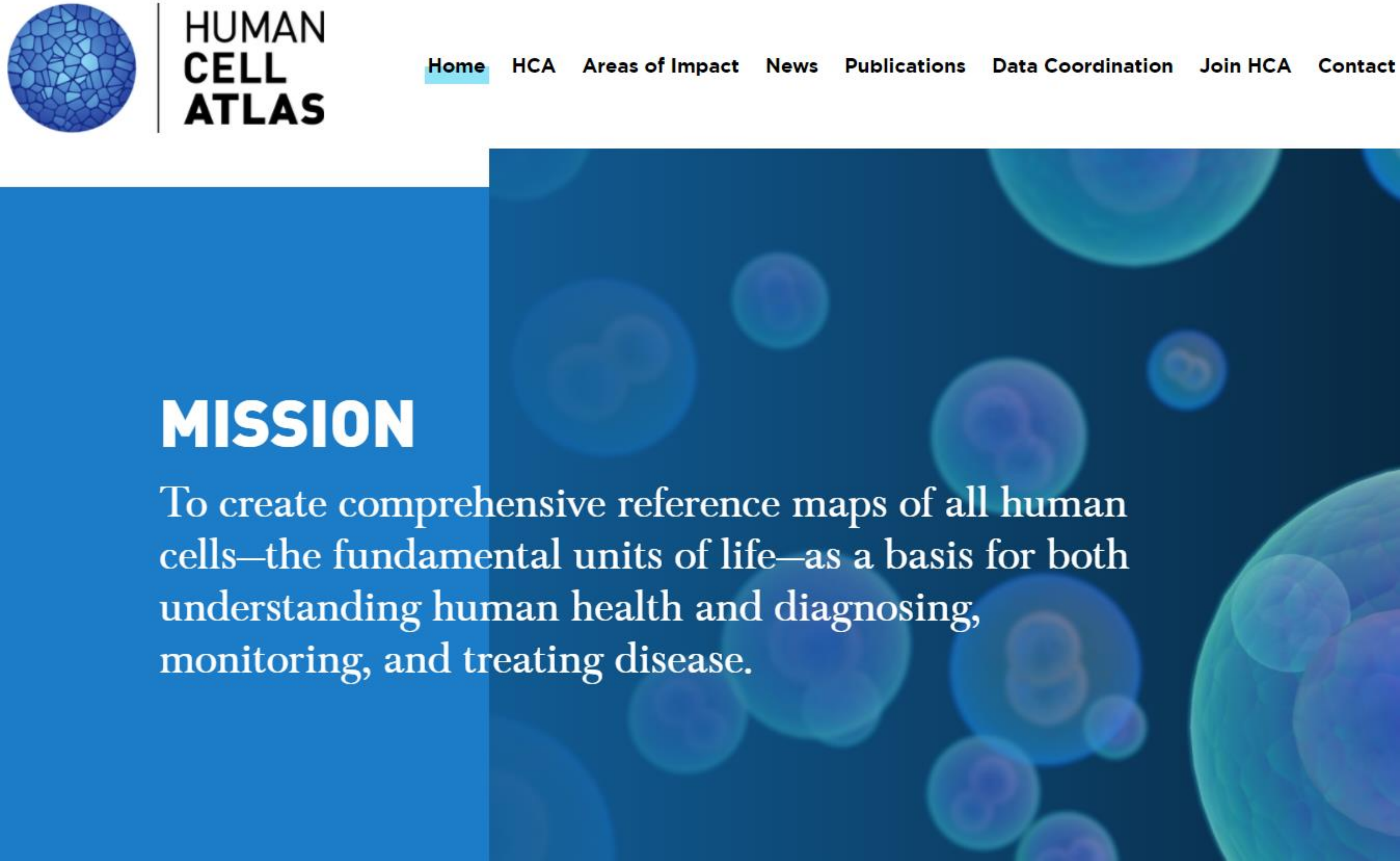
- 100 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 6k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 12k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 4k PBMCs from a Healthy Donor
- 8k PBMCs from a Healthy Donor
- 9k Brain Cells from an E18 Mouse
- 3k Pan T Cells from a Healthy Donor
- 4k Pan T Cells from a Healthy Donor
- Aggregate of t_3k and t_4k

1.3 million mouse embryonic brain cells, 10X Chromium





10x Genomics | LIT000015 Chromium™ Million Brain Cells Application Note



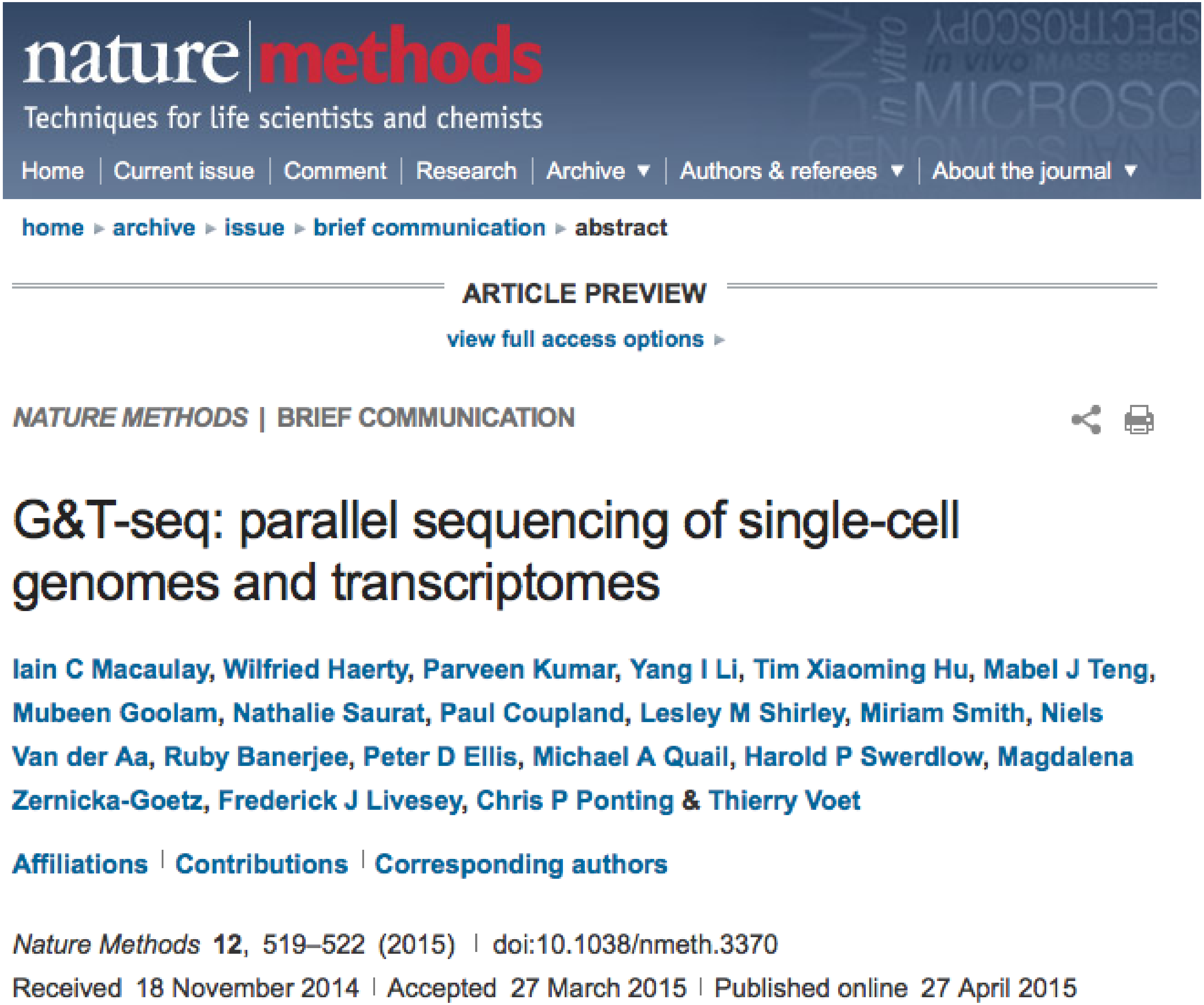
MISSION

To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

www.humancellatlas.org

Beyond single cell RNA-seq

Single nuclei sequencing	scNuc-seq
Epigenomics	scBS-seq, scRRBS-seq, scCHIP-seq, scATAC-seq, scDNase-seq
Genomics	Whole genome, exome
Multiple simultaneous measurements	
RNA + DNA	DR-seq, G&T-seq
RNA + methylation	scM&T-seq, scMT-seq
RNA + DNA + methylation	scTrio-seq
RNA + protein + chromatin	DOGMA-seq
RNA + protein	index sorting, CITE-seq
RNA + genome editing	Perturb-seq, CRISP-seq, CROP-seq


nature methods
Techniques for life scientists and chemists
Home | Current issue | Comment | Research | Archive ▼ | Authors & referees ▼ | About the journal ▼
home ▸ archive ▸ issue ▸ brief communication ▸ abstract
ARTICLE PREVIEW
view full access options ▸

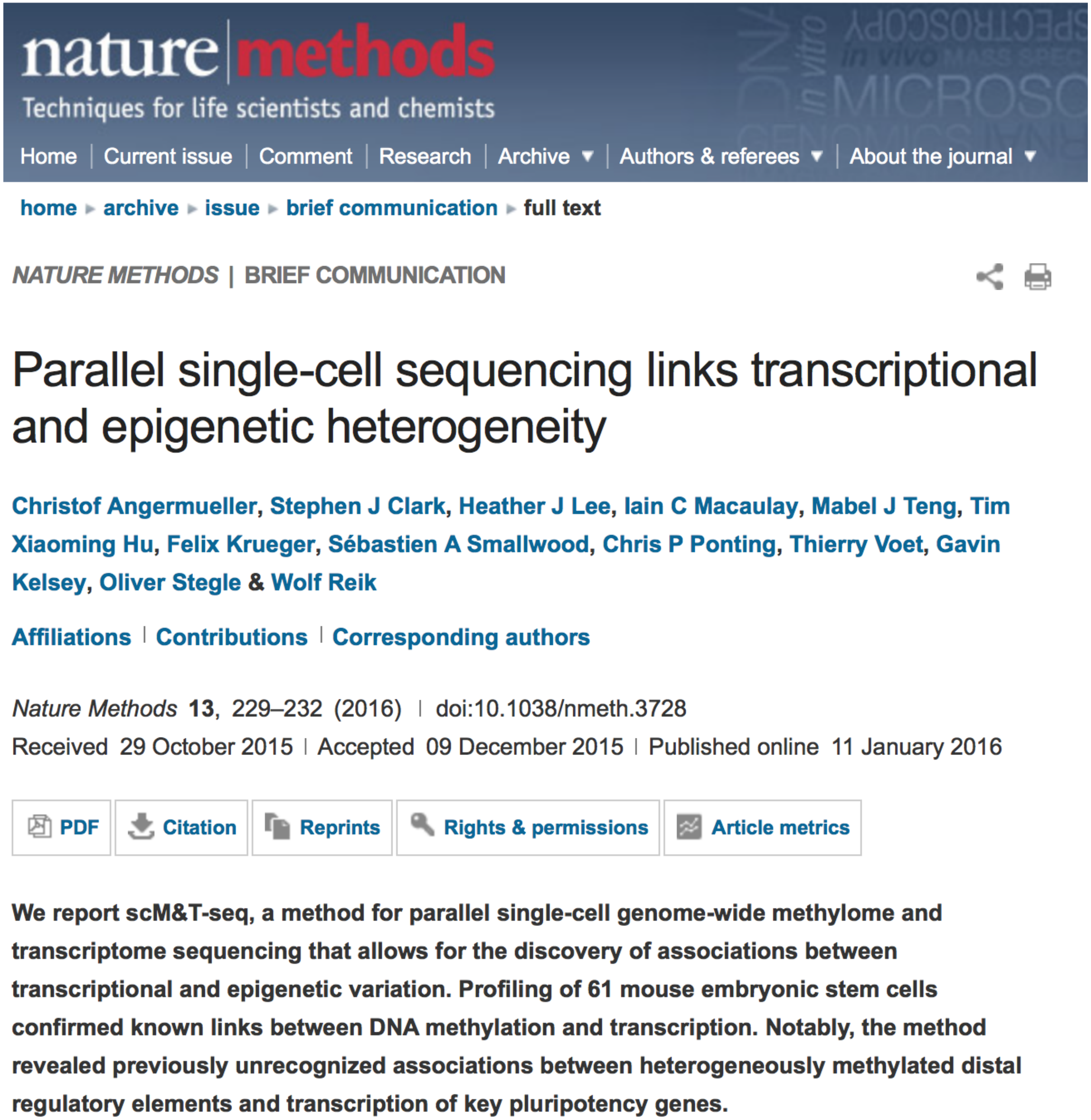
NATURE METHODS | BRIEF COMMUNICATION

G&T-seq: parallel sequencing of single-cell genomes and transcriptomes

Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, Miriam Smith, Niels Van der Aa, Ruby Banerjee, Peter D Ellis, Michael A Quail, Harold P Swerdlow, Magdalena Zernicka-Goetz, Frederick J Livesey, Chris P Ponting & Thierry Voet

Nature Methods **12**, 519–522 (2015) | doi:10.1038/nmeth.3370
Received 18 November 2014 | Accepted 27 March 2015 | Published online 27 April 2015

BOMBAY SAPPHIRE
Distilled
DRY GIN
EVER-TREE
PREMIUM INDIAN
TONIC
WATER
6.8FL OZ (200ML)



NATURE METHODS | BRIEF COMMUNICATION

Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity

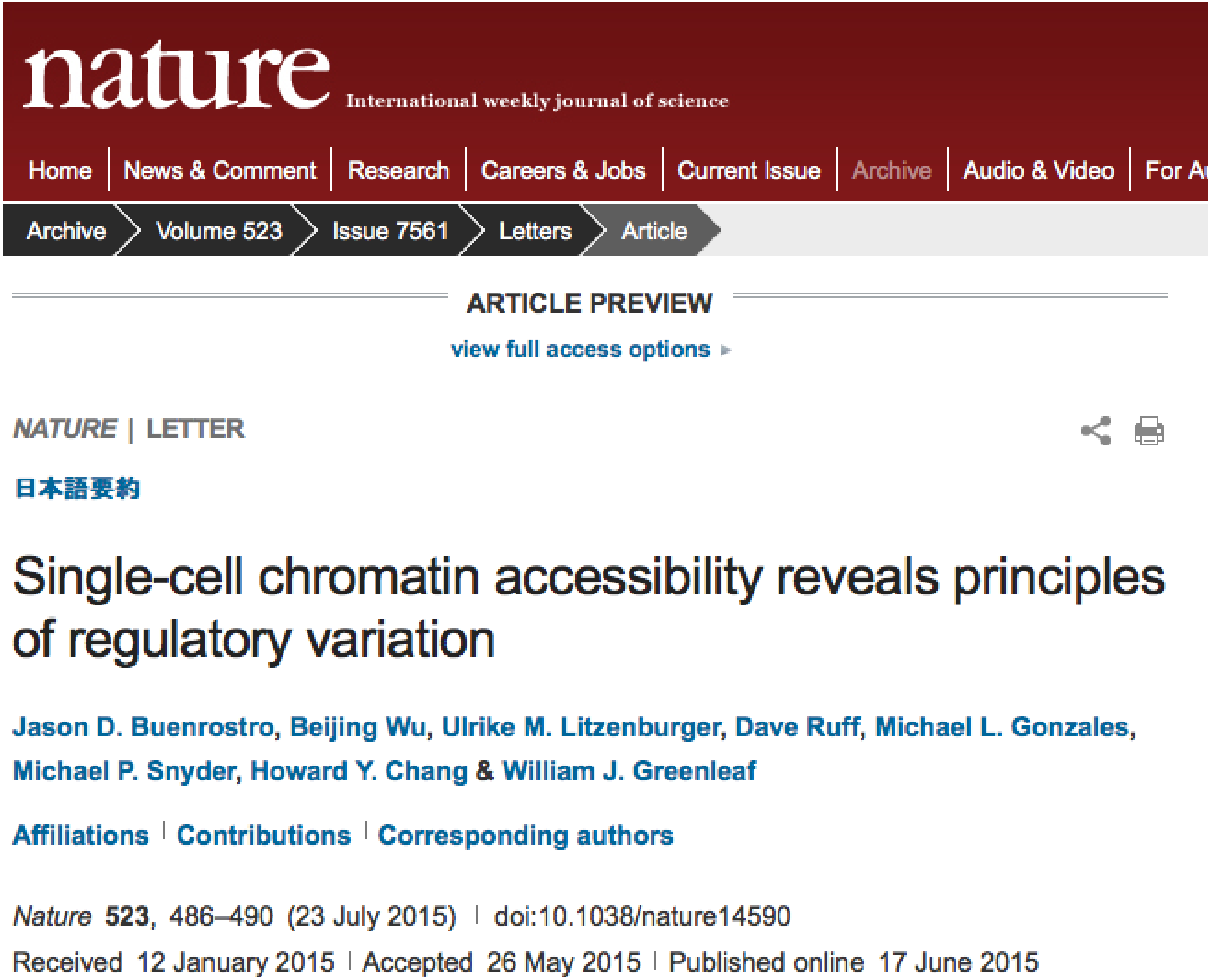
Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle & Wolf Reik

Affiliations | Contributions | Corresponding authors

Nature Methods **13**, 229–232 (2016) | doi:10.1038/nmeth.3728

Received 29 October 2015 | Accepted 09 December 2015 | Published online 11 January 2016

We report scM&T-seq, a method for parallel single-cell genome-wide methylome and transcriptome sequencing that allows for the discovery of associations between transcriptional and epigenetic variation. Profiling of 61 mouse embryonic stem cells confirmed known links between DNA methylation and transcription. Notably, the method revealed previously unrecognized associations between heterogeneously methylated distal regulatory elements and transcription of key pluripotency genes.



nature International weekly journal of science

ARTICLE PREVIEW

NATURE | LETTER

日本語要約

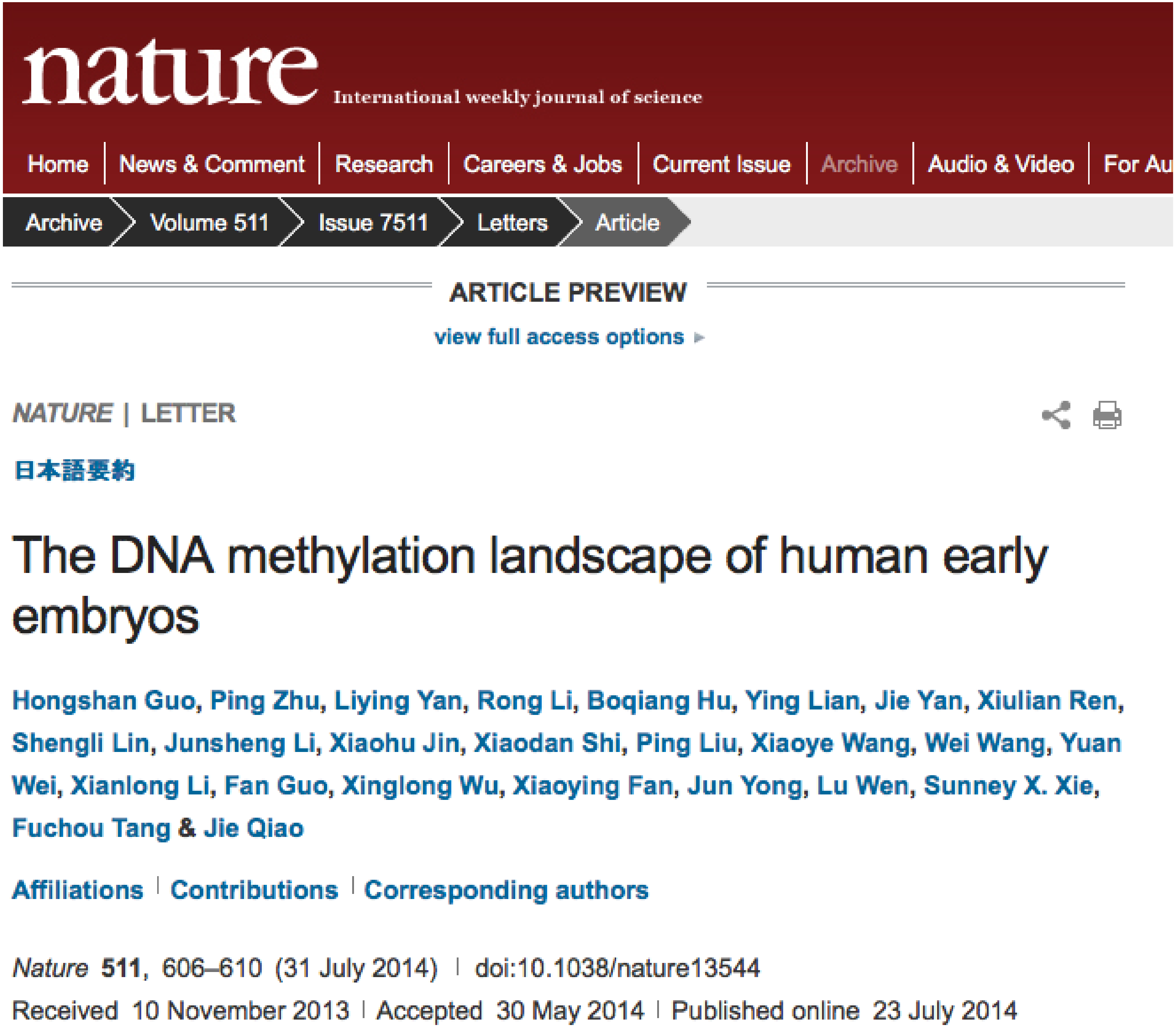
Single-cell chromatin accessibility reveals principles of regulatory variation

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang & William J. Greenleaf

Affiliations | Contributions | Corresponding authors

Nature **523**, 486–490 (23 July 2015) | doi:10.1038/nature14590

Received 12 January 2015 | Accepted 26 May 2015 | Published online 17 June 2015



nature International weekly journal of science

ARTICLE PREVIEW

view full access options ▸

NATURE | LETTER

日本語要約

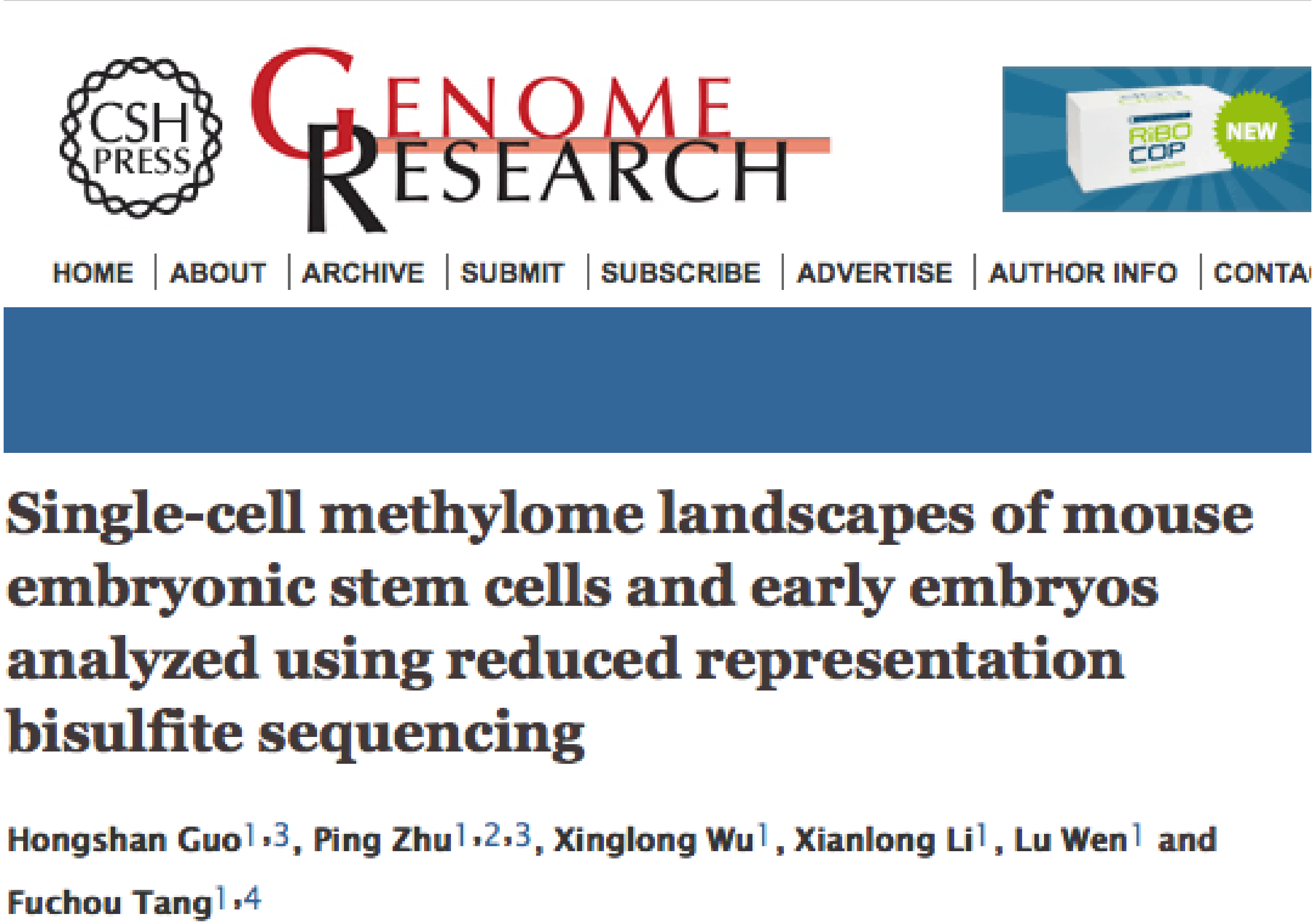
The DNA methylation landscape of human early embryos

Hongshan Guo, Ping Zhu, Liying Yan, Rong Li, Boqiang Hu, Ying Lian, Jie Yan, Xiulian Ren, Shengli Lin, Junsheng Li, Xiaohu Jin, Xiaodan Shi, Ping Liu, Xiaoye Wang, Wei Wang, Yuan Wei, Xianlong Li, Fan Guo, Xinglong Wu, Xiaoying Fan, Jun Yong, Lu Wen, Sunney X. Xie, Fuchou Tang & Jie Qiao

Affiliations | Contributions | Corresponding authors

Nature **511**, 606–610 (31 July 2014) | doi:10.1038/nature13544

Received 10 November 2013 | Accepted 30 May 2014 | Published online 23 July 2014



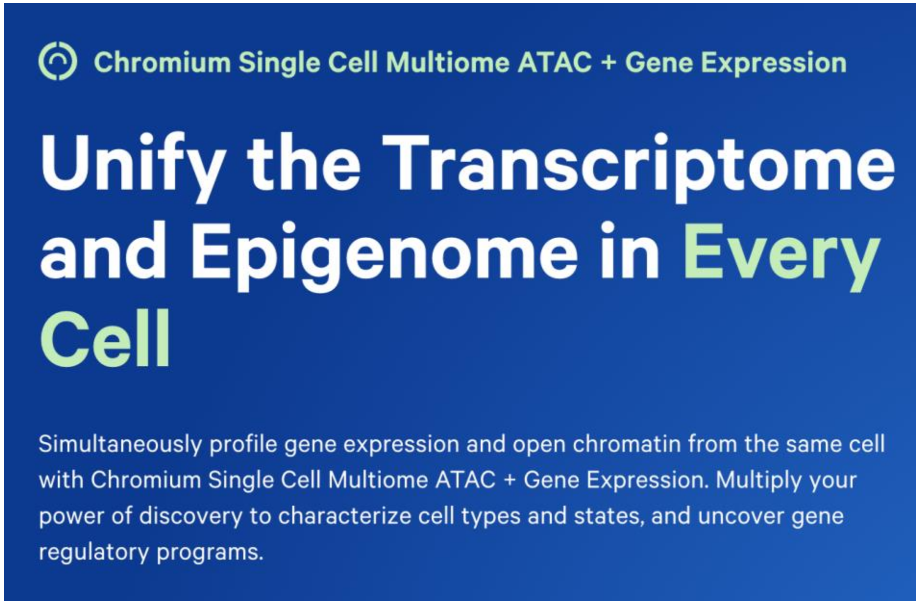
CSH PRESS

GENOME RESEARCH

Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing

Hongshan Guo[1,3], Ping Zhu[1,2,3], Xinglong Wu[1], Xianlong Li[1], Lu Wen[1] and Fuchou Tang[1,4]

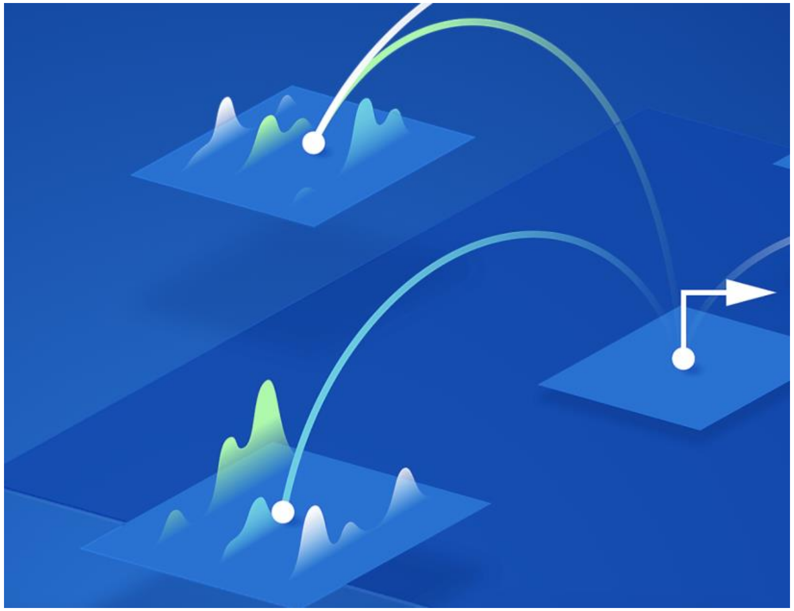
scATAC/RNA-seq



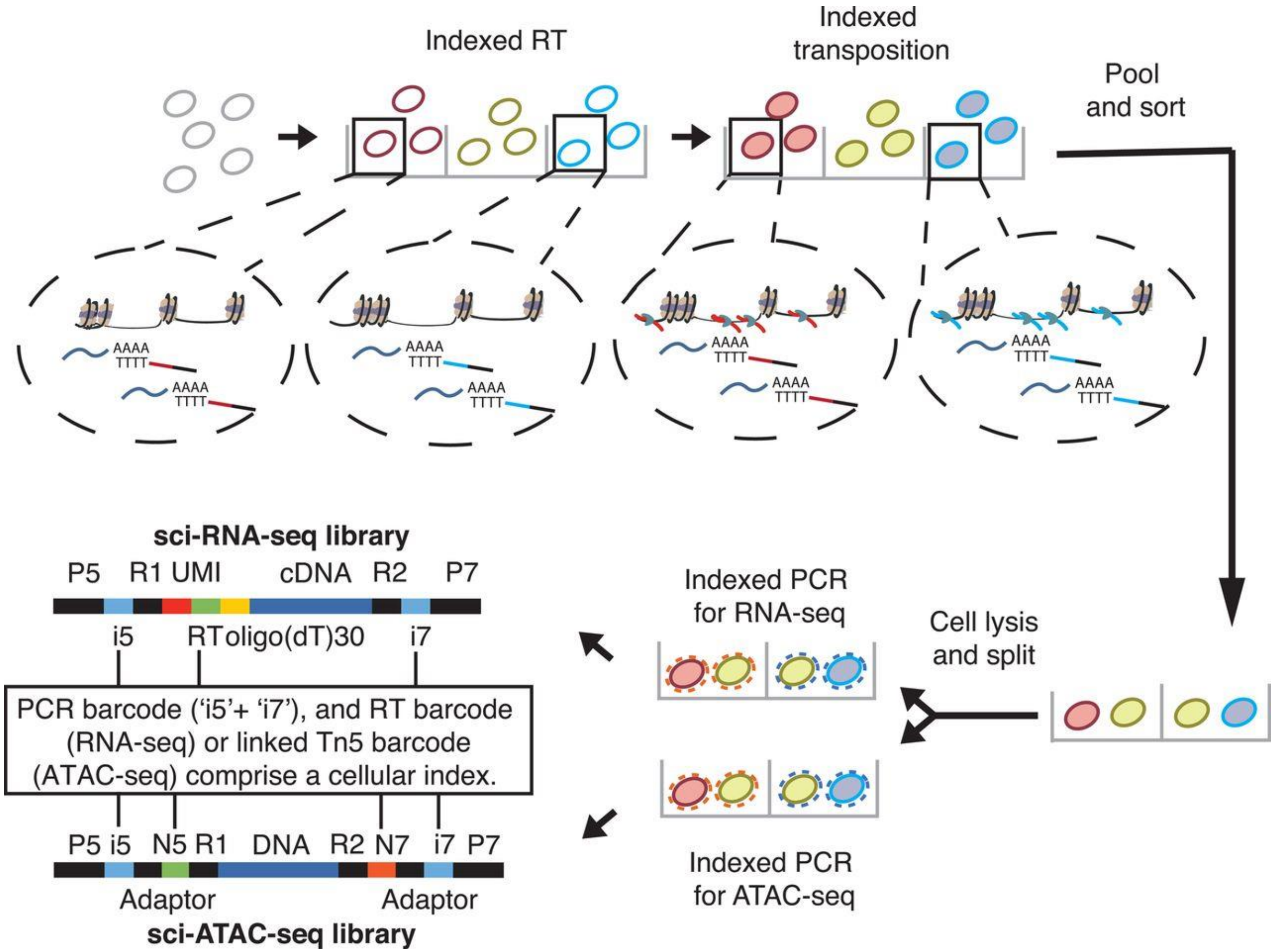
Chromium Single Cell Multiome ATAC + Gene Expression

Unify the Transcriptome and Epigenome in Every Cell

Simultaneously profile gene expression and open chromatin from the same cell with Chromium Single Cell Multiome ATAC + Gene Expression. Multiply your power of discovery to characterize cell types and states, and uncover gene regulatory programs.

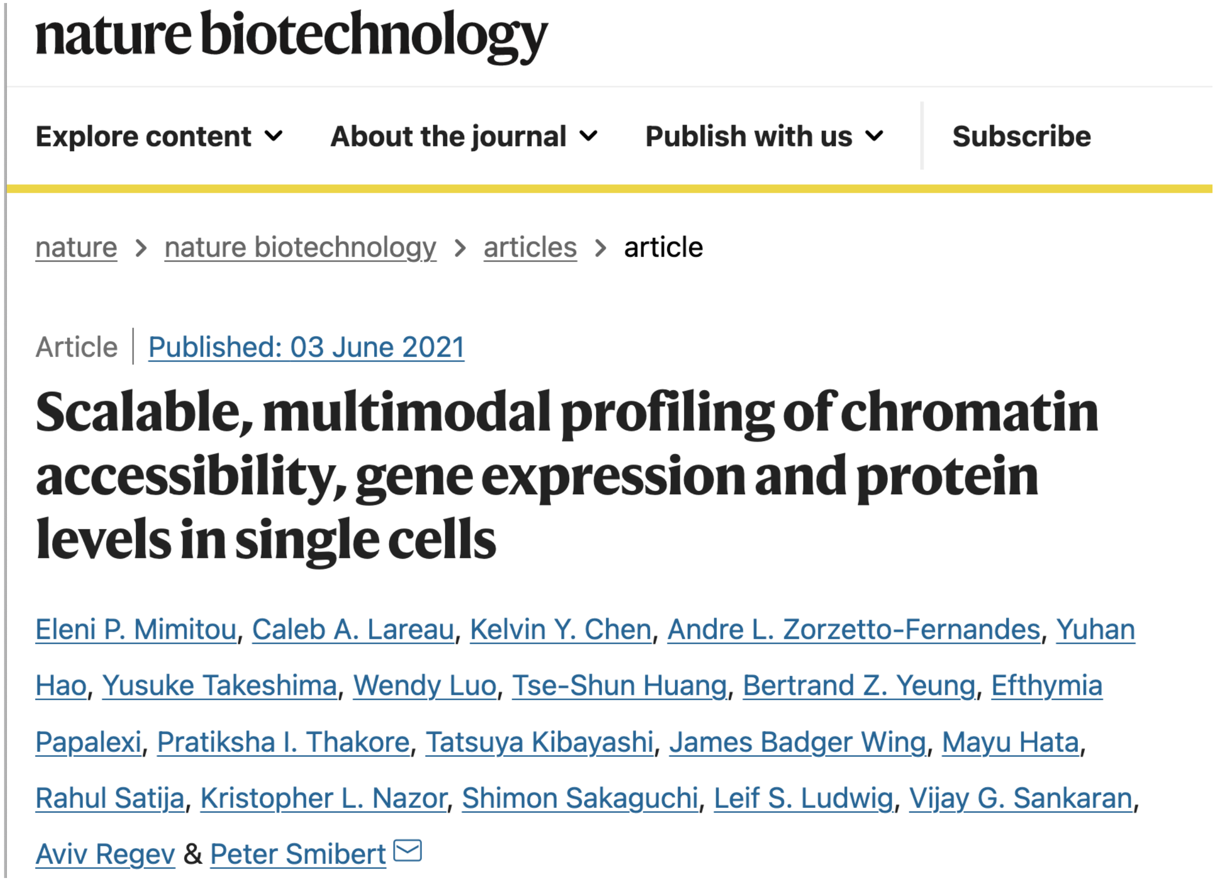


https://www.10xgenomics.com/products/single-cell-multiome-atac-plus-gene-expression



https://science.sciencemag.org/content/361/6409/1380

DOGMA-seq



Article | Published: 03 June 2021

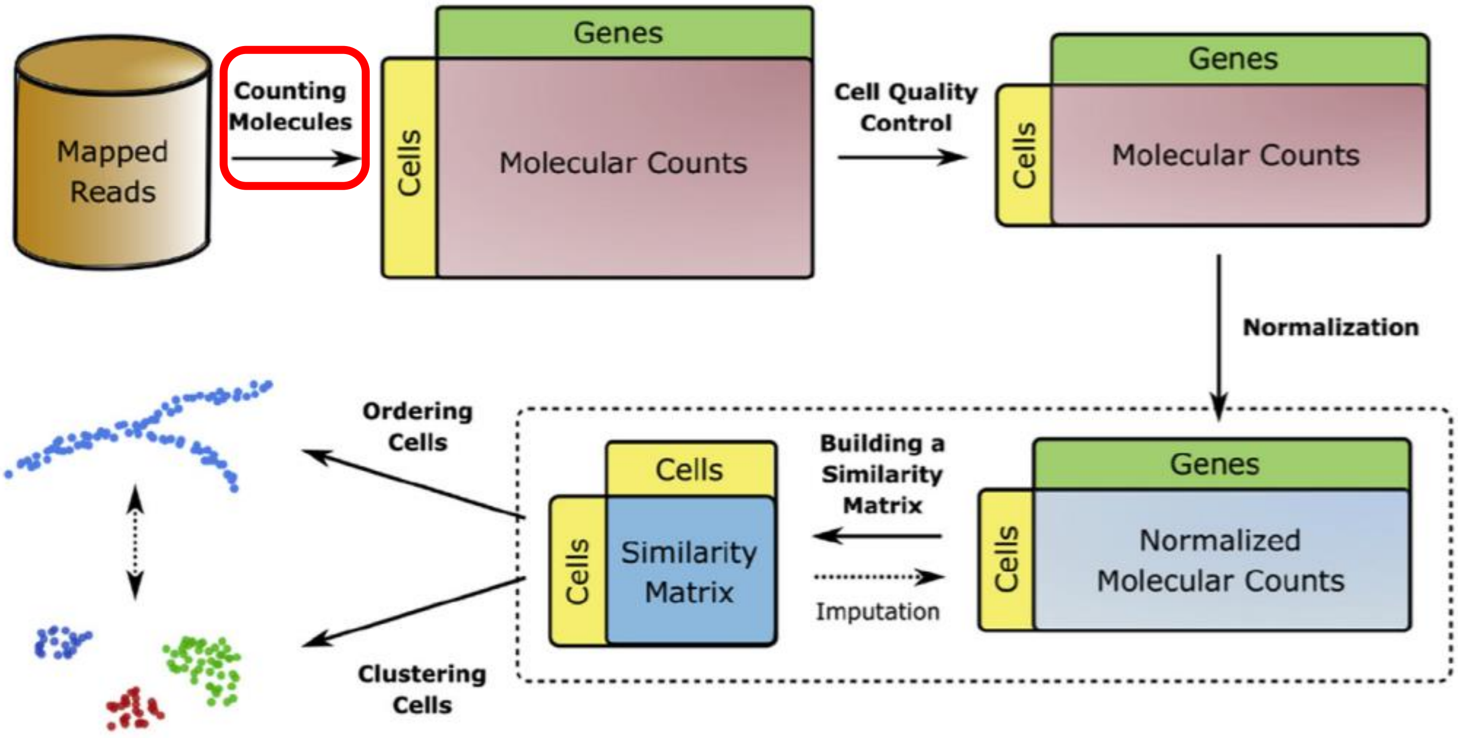
Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells

Eleni P. Mimitou, Caleb A. Lareau, Kelvin Y. Chen, Andre L. Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z. Yeung, Efthymia Papalexi, Pratiksha I. Thakore, Tatsuya Kibayashi, James Badger Wing, Mayu Hata, Rahul Satija, Kristopher L. Nazor, Shimon Sakaguchi, Leif S. Ludwig, Vijay G. Sankaran, Aviv Regev & Peter Smibert

scATAC-seq (single-cell assay for transposase accessible chromatin by sequencing), plus select antigen profiling by sequencing (ASAP-seq), and optional capture of mitochondrial DNA for clonal tracking.

https://www.nature.com/articles/s41587-021-00927-2

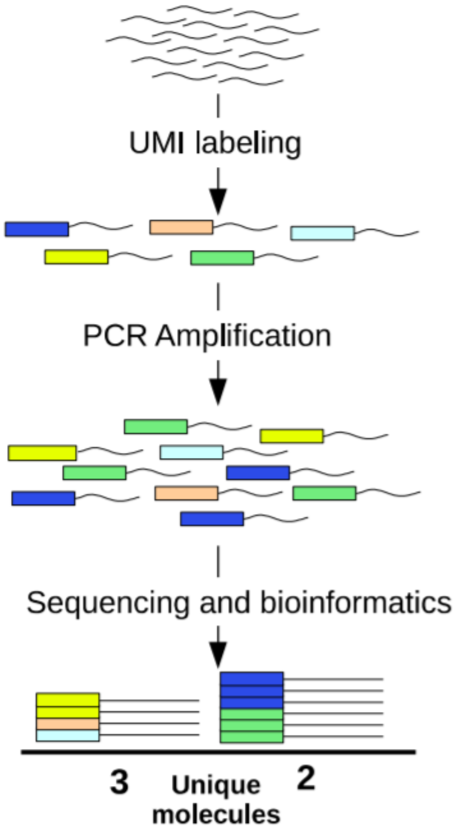
Analysis:
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Counting Molecules

- Counting reads
  - featureCounts, etc.
- Counting UMIs
  - Unique
    - does not account for PCR and sequencing errors
  - Directional adjacency graph (UMI-tools)
  - Bayesian (dropEst)
  - Proprietary (SevenBridges for BD Precise)

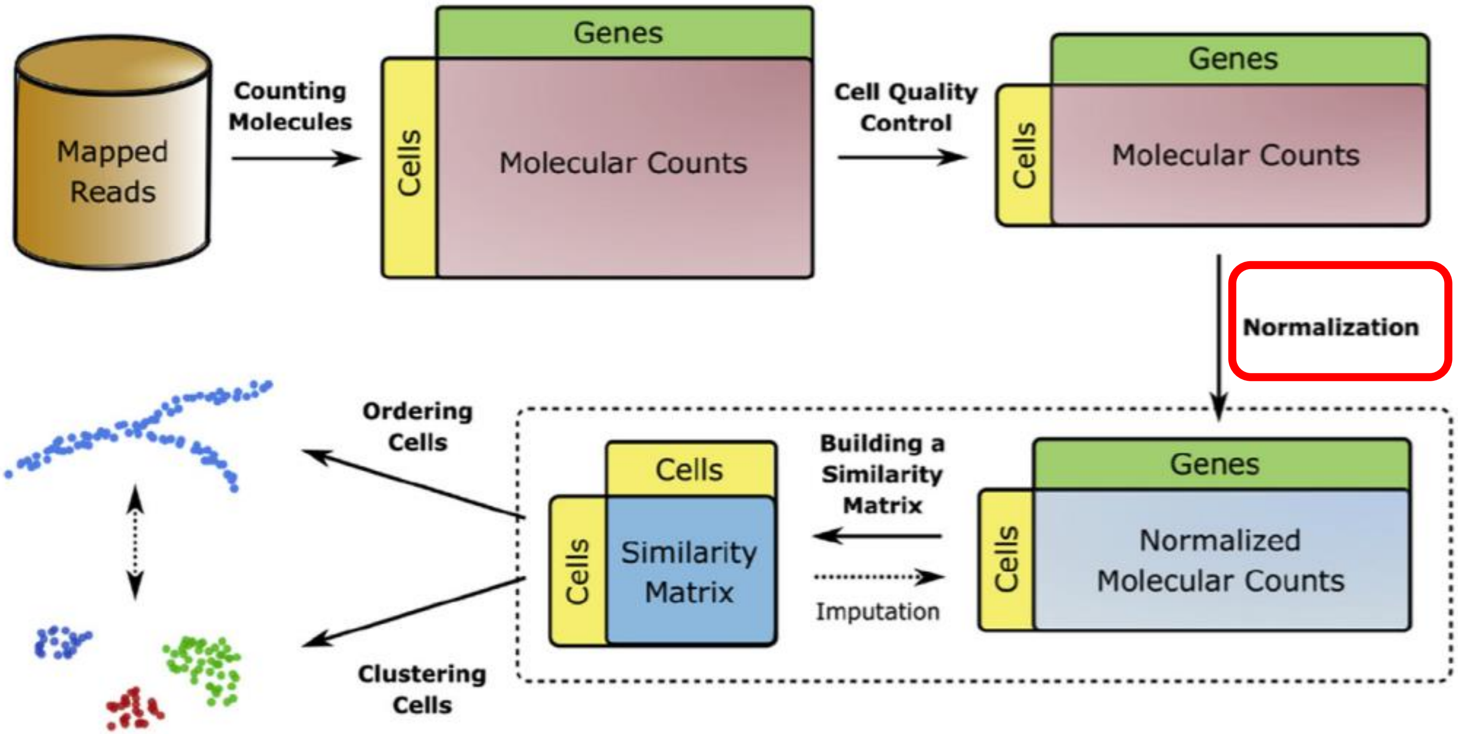


Smith *et al*., Genome Research 2017

Commonly used open-source tools

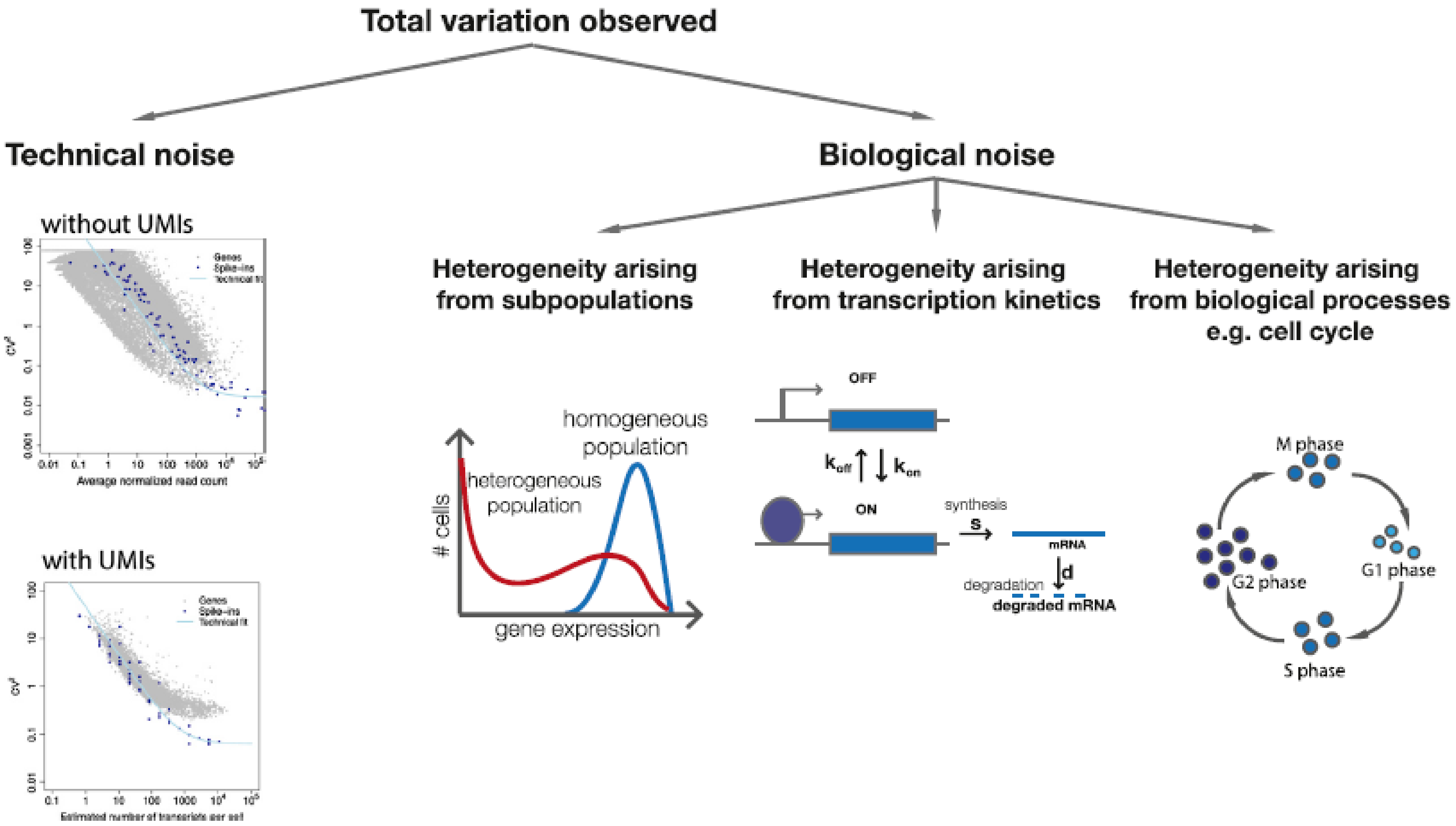
1. Infer which barcodes come from valid cells – **UMI-tools**
2. Extract cell barcodes and UMIs from R1 and add to R2 – **UMI-tools**
3. Align to reference genome (GRCh38) – **STAR**
4. Assign reads to genes (Ensembl/gencode) – **featureCounts**
5. Count unique UMIs per gene – **UMI-tools**
6. QC – **fastqc, picard, multiqc, custom scripts**

Structure of a generic pipeline



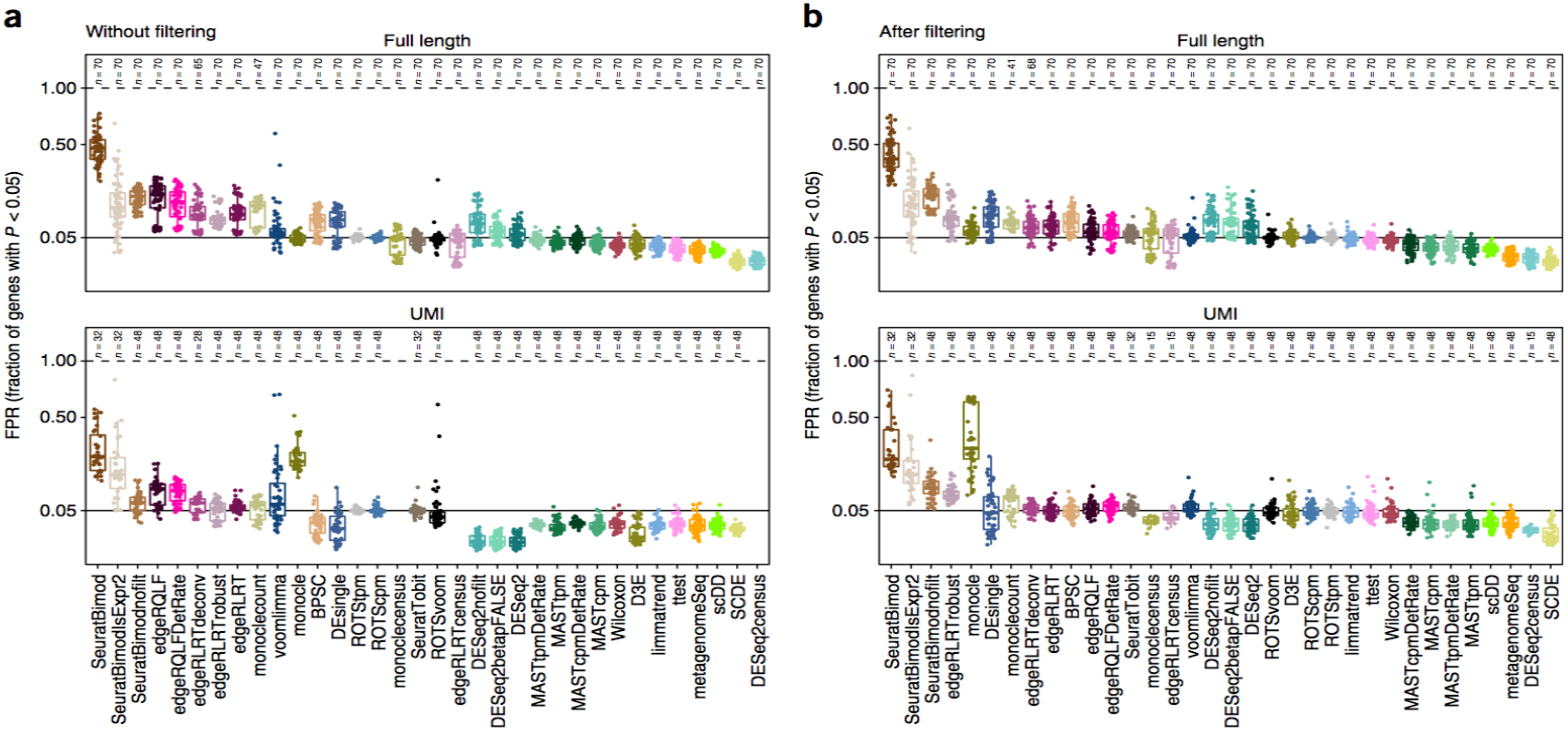
Camara, Current Opinions in Systems Biology, 2018

Normalization challenges



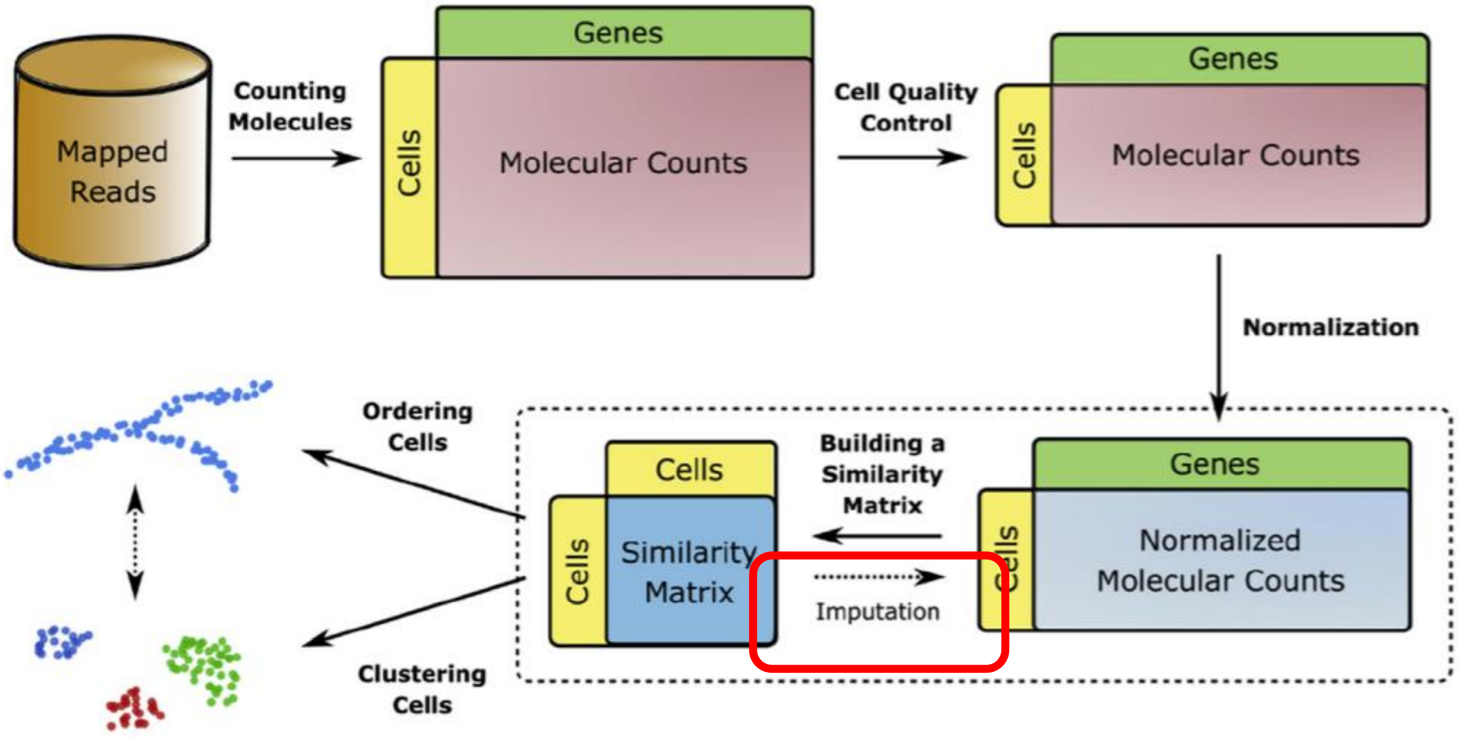
Kolodziejczyk *et al.*, Briefings in Functional Genomics 2017

Normalization + Differential Expression Analysis



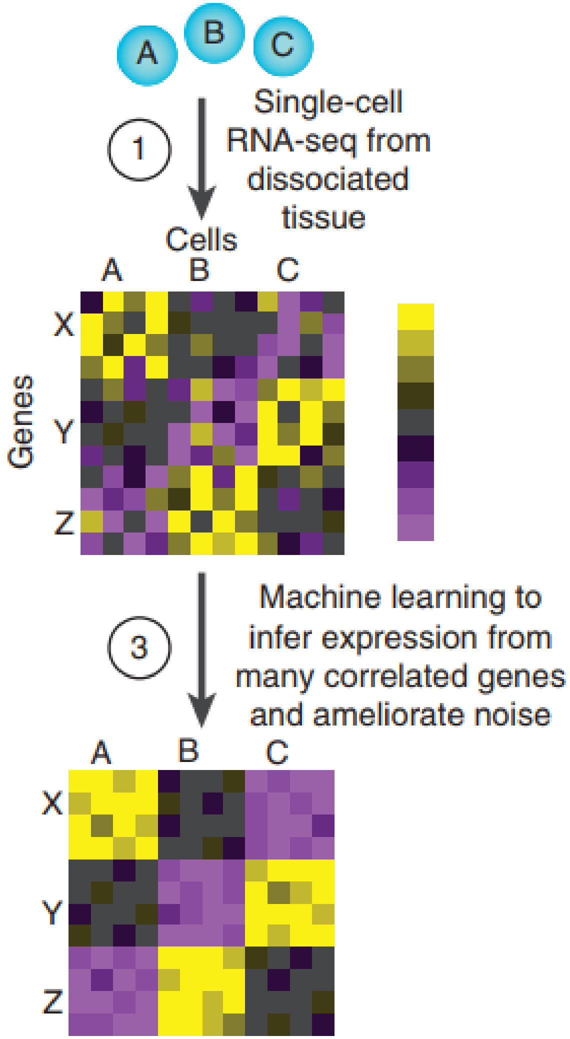
Soneson and Robinson, Nature Methods 2018

Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Gene Expression Imputation


A
B
C
1
Single-cell RNA-seq from dissociated tissue
Cells
A
B
C
Genes
X
Y
Z
3
Machine learning to infer expression from many correlated genes and ameliorate noise
A
B
C
X
Y
Z

Satija *et al*., Nature Biotechnology 2015

Gene Expression Imputation

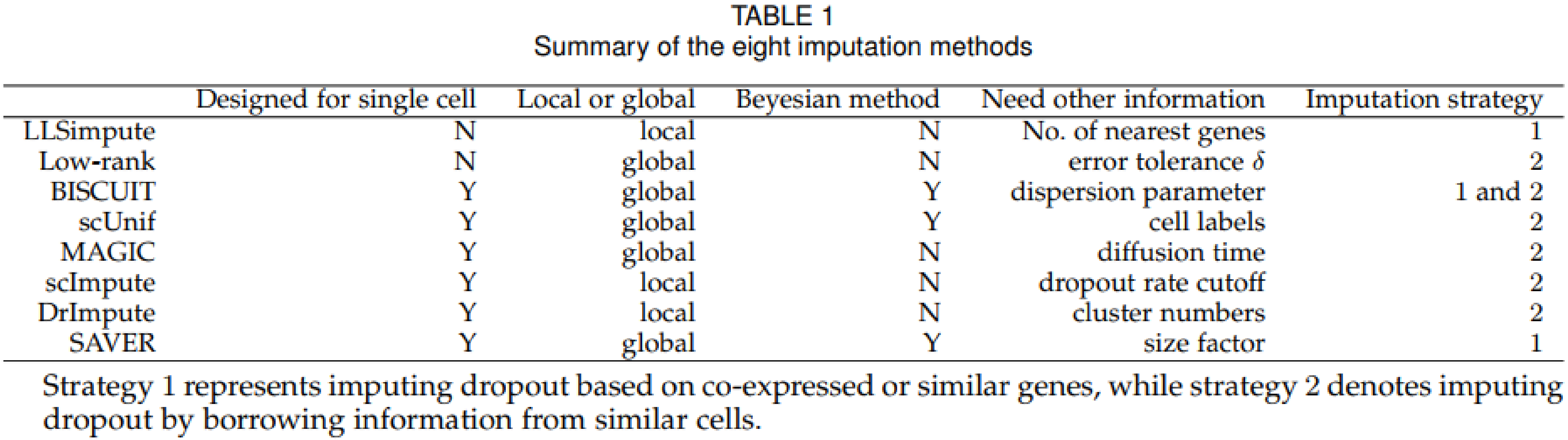


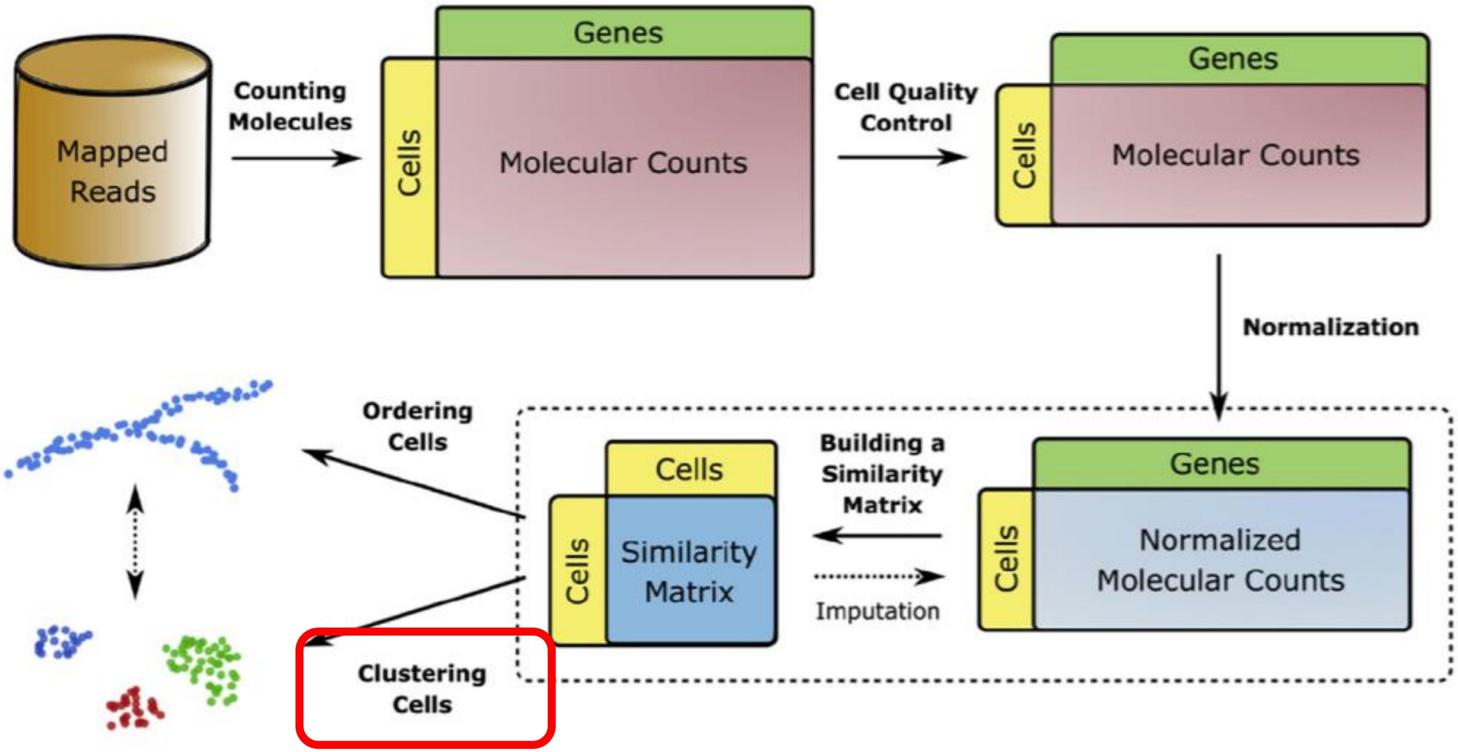
TABLE 1
Summary of the eight imputation methods

	Designed for single cell	Local or global	Beyesian method	Need other information	Imputation strategy
LLSimpute	N	local	N	No. of nearest genes	1
Low-rank	N	global	N	error tolerance δ	2
BISCUIT	Y	global	Y	dispersion parameter	1 and 2
scUnif	Y	global	Y	cell labels	2
MAGIC	Y	global	N	diffusion time	2
scImpute	Y	local	N	dropout rate cutoff	2
DrImpute	Y	local	N	cluster numbers	2
SAVER	Y	global	Y	size factor	1

Strategy 1 represents imputing dropout based on co-expressed or similar genes, while strategy 2 denotes imputing dropout by borrowing information from similar cells.

Zhang and Zhang, Biorxiv 2017

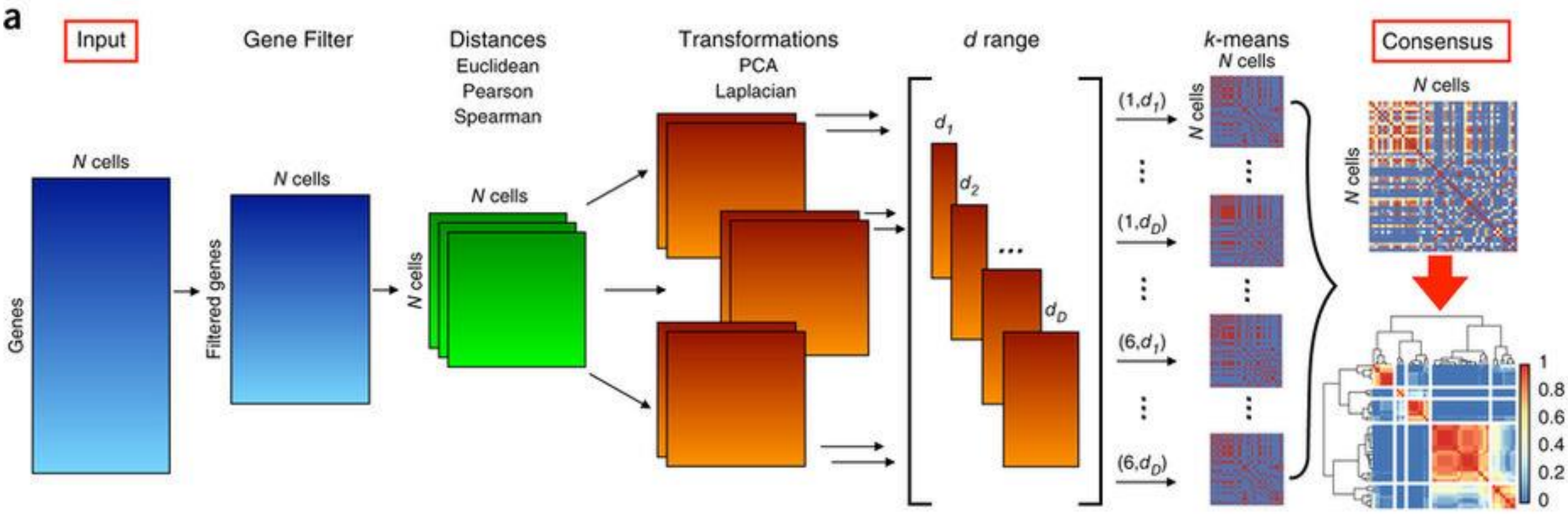
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Clustering Cells

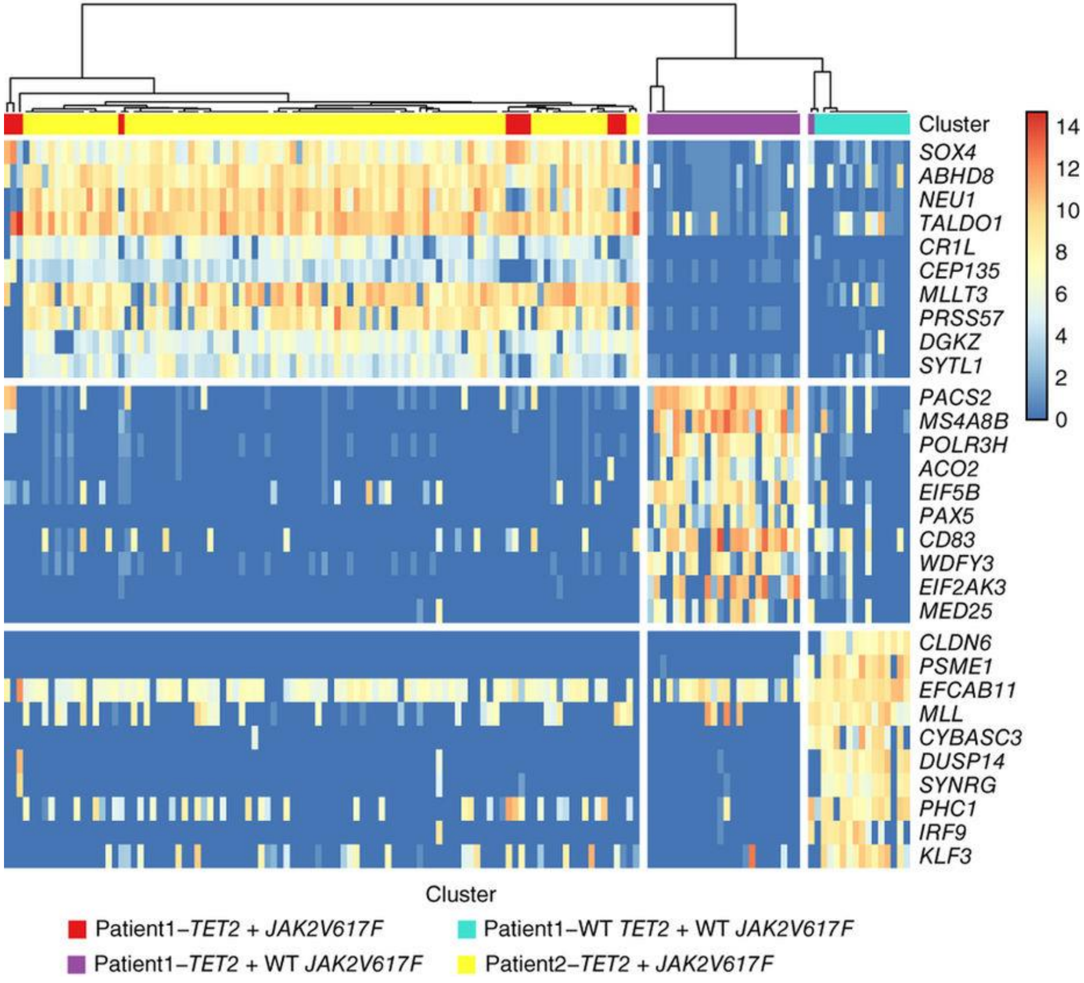
SC3: consensus clustering of single-cell RNA-seq data



Kiselev *et al.*, Nature Methods 2017

Differential Expression Analysis

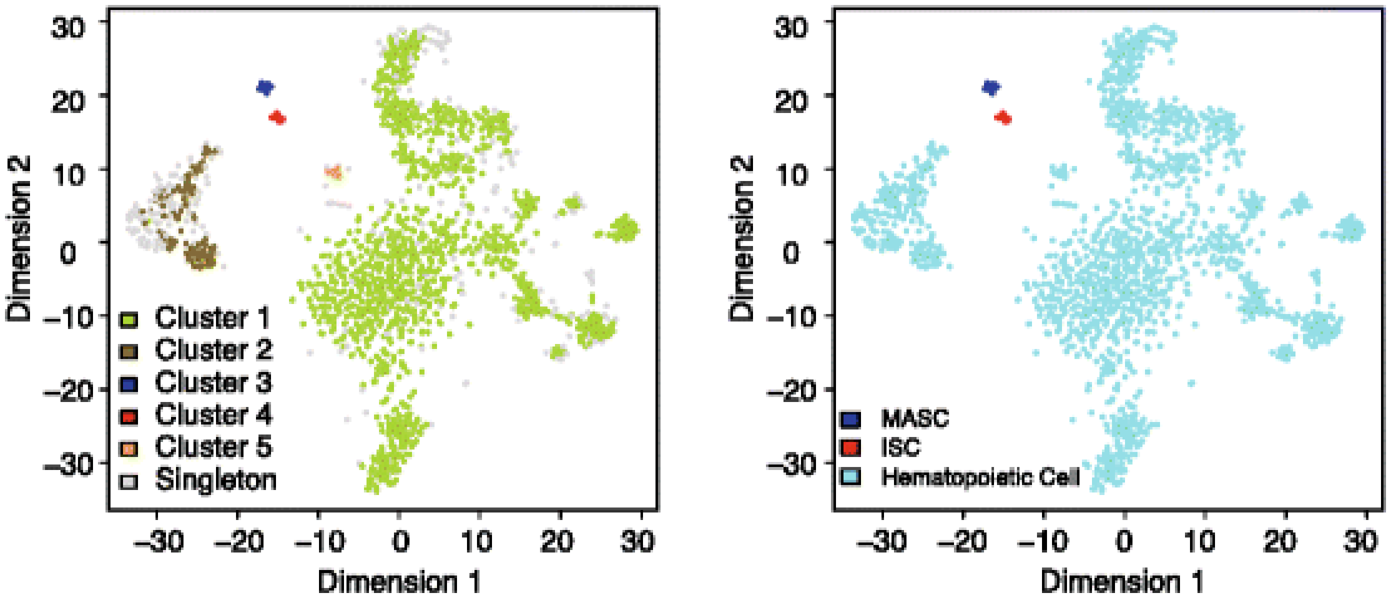
SC3: consensus clustering of single-cell RNA-seq data



Kiselev *et al*., Nature Methods 2017

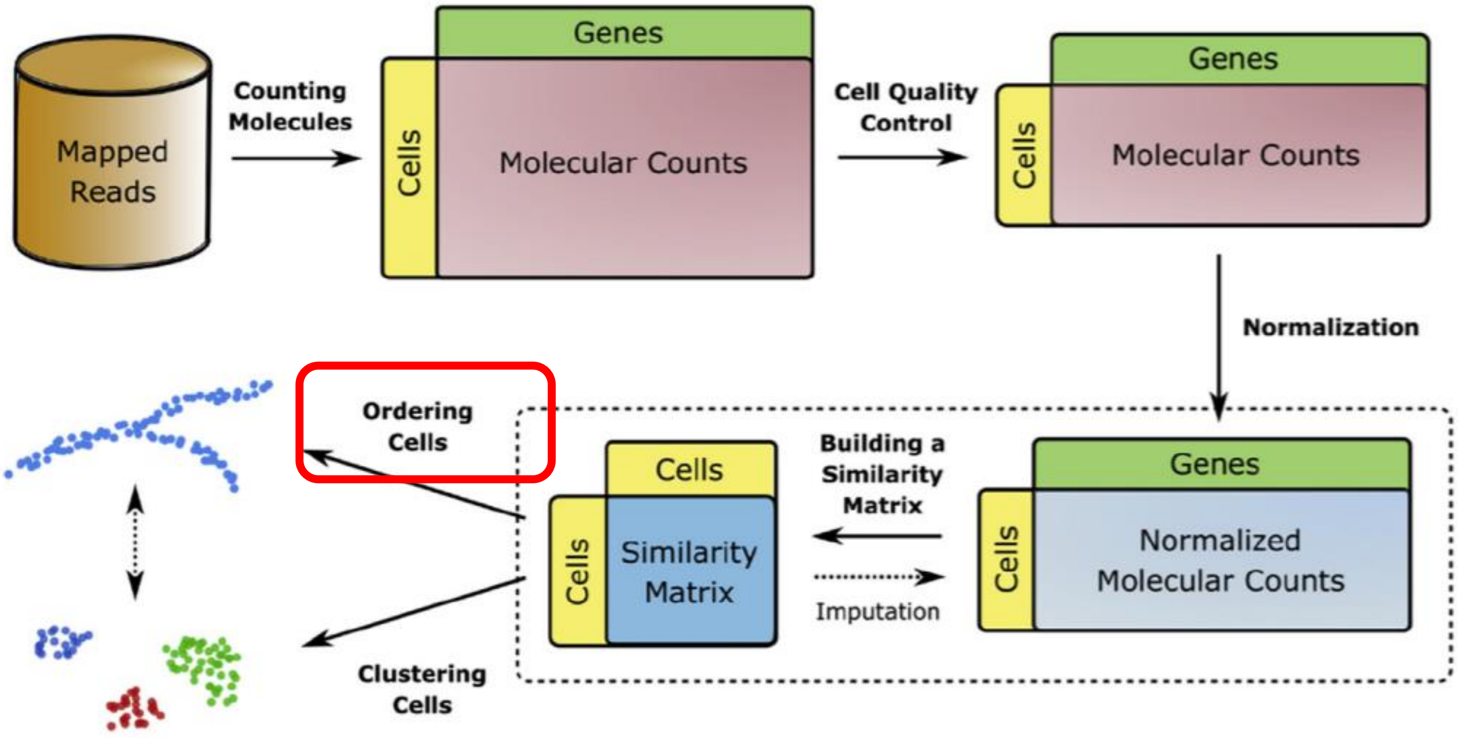
Clustering Cells

GiniClust: detecting rare cell types from single-cell gene expression data with Gini index



Jiang *et al*., Genome Biology 2016

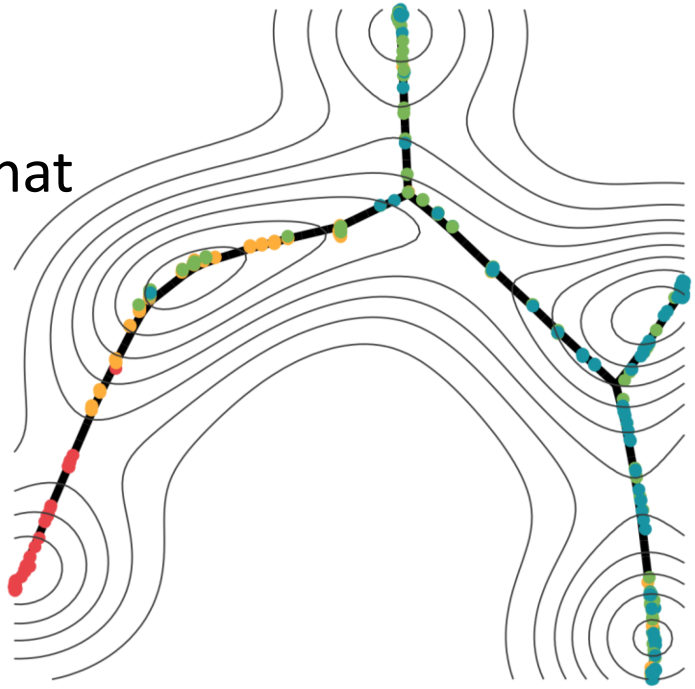
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

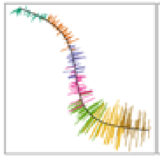
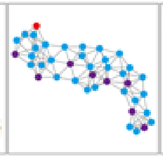
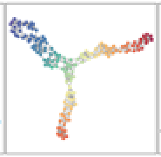
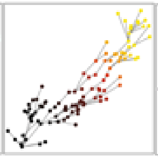
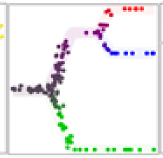
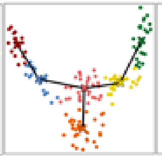
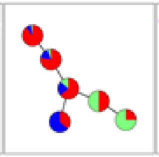
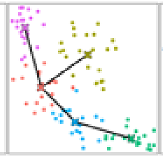
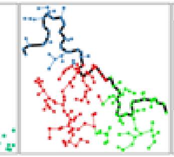
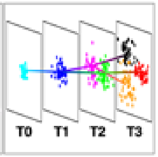
Single Cell Trajectory Inference

- “Pseudotime” introduced in Trapnell *et al.*, *Nature Biotechnology* 2014 (Monocle)
- Steps:
  1. (Optional) Choose genes that define a biological process
  2. Reduce dimensionality
  3. Order cells



http://cole-trapnell-lab.github.io/monocle-release/

Single Cell Trajectory Inference

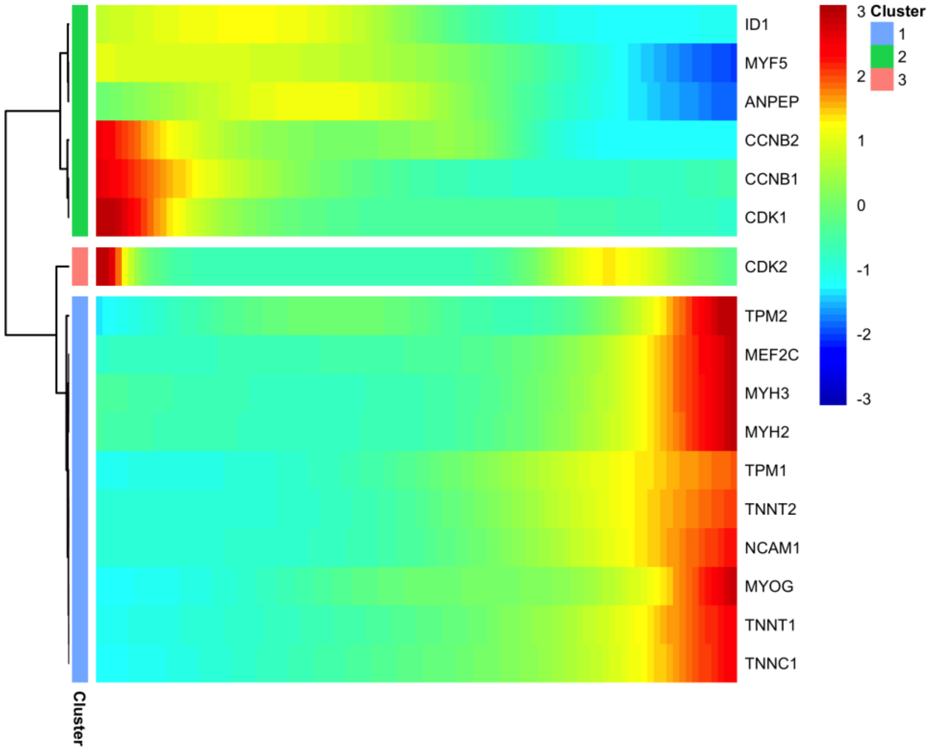
Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	−	+	−	−
Scalability w.r.t. cells	−	−	±	±	−	±	+	+	−	±
Scalability w.r.t. genes	+	+	+	+	−	+	±	±	±	+
Code and documentation	−	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	−	±	−	+	+	+

Cannoodt *et al.*, 2016

Single Cell Trajectory Inference

- “Pseudotime” introduced in Trapnell *et al.*, Nature Biotechnology 2014 (Monocle)
- Steps:
    1. (Optional) Choose genes that define a biological process
    2. Reduce dimensionality

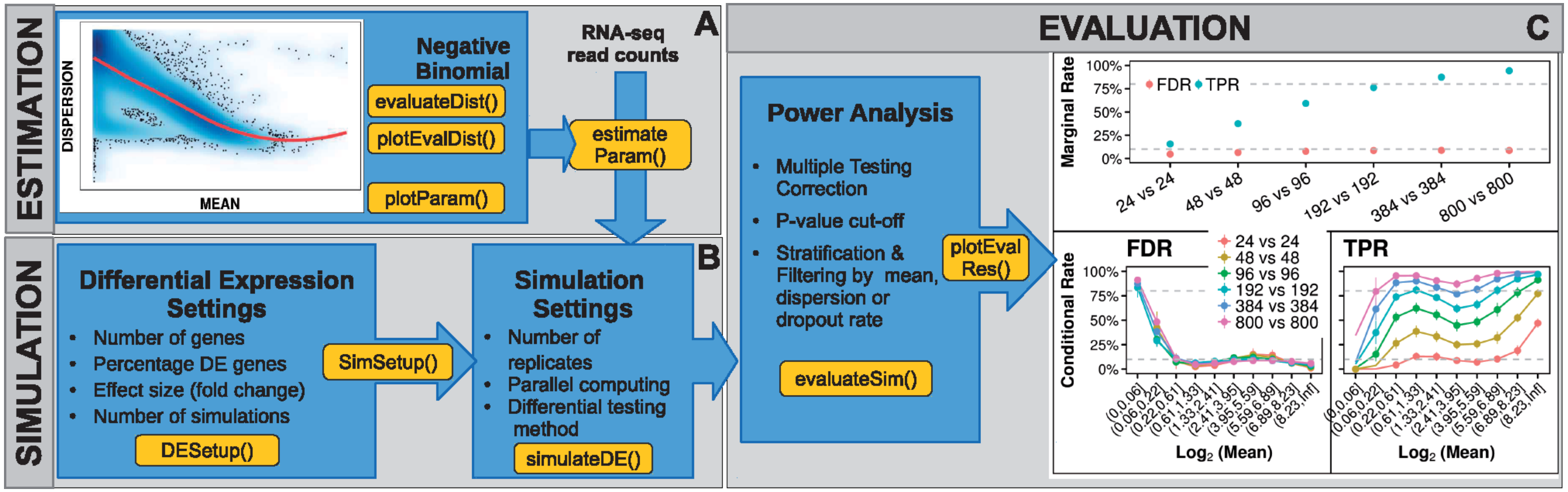
Differential Expression Analysis using Monocle



Qiu *et al.*, Nature Methods 2017

Simulating scRNA-seq data

PowSimR



ESTIMATION

Negative Binomial

evaluateDist()

plotEvalDist()

plotParam()

RNA-seq read counts

estimate Param()

SIMULATION

Differential Expression Settings

- Number of genes
- Percentage DE genes
- Effect size (fold change)
- Number of simulations

DESetup()

SimSetup()

Simulation Settings

- Number of replicates
- Parallel computing
- Differential testing method

simulateDE()

EVALUATION

Power Analysis

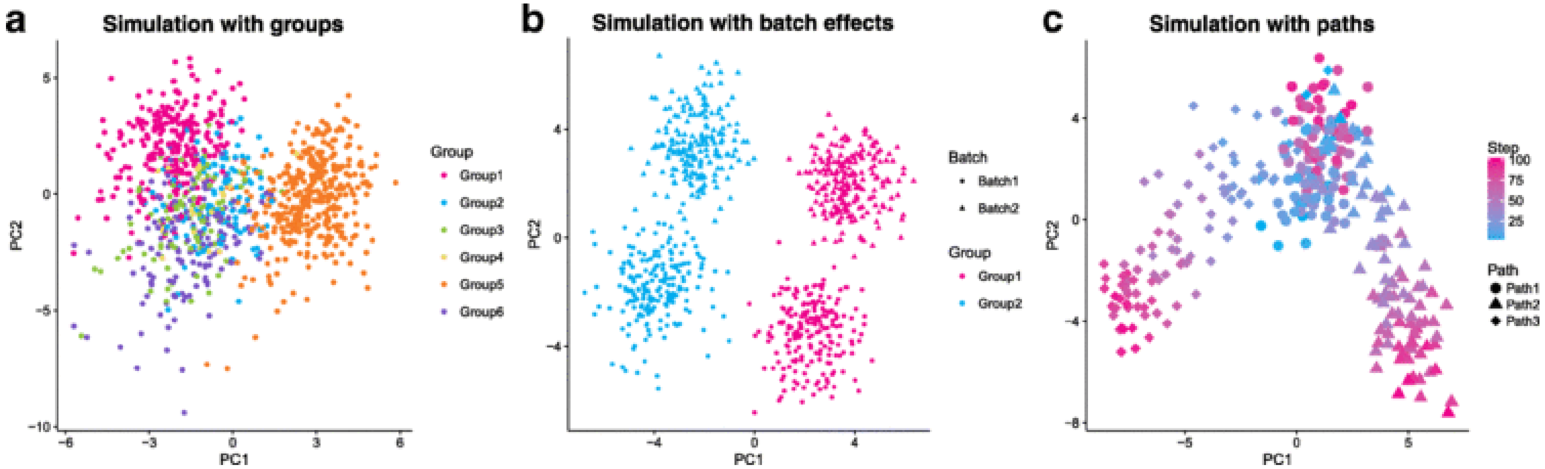
- Multiple Testing Correction
- P-value cut-off
- Stratification & Filtering by mean, dispersion or dropout rate

plotEvalRes()

evaluateSim()

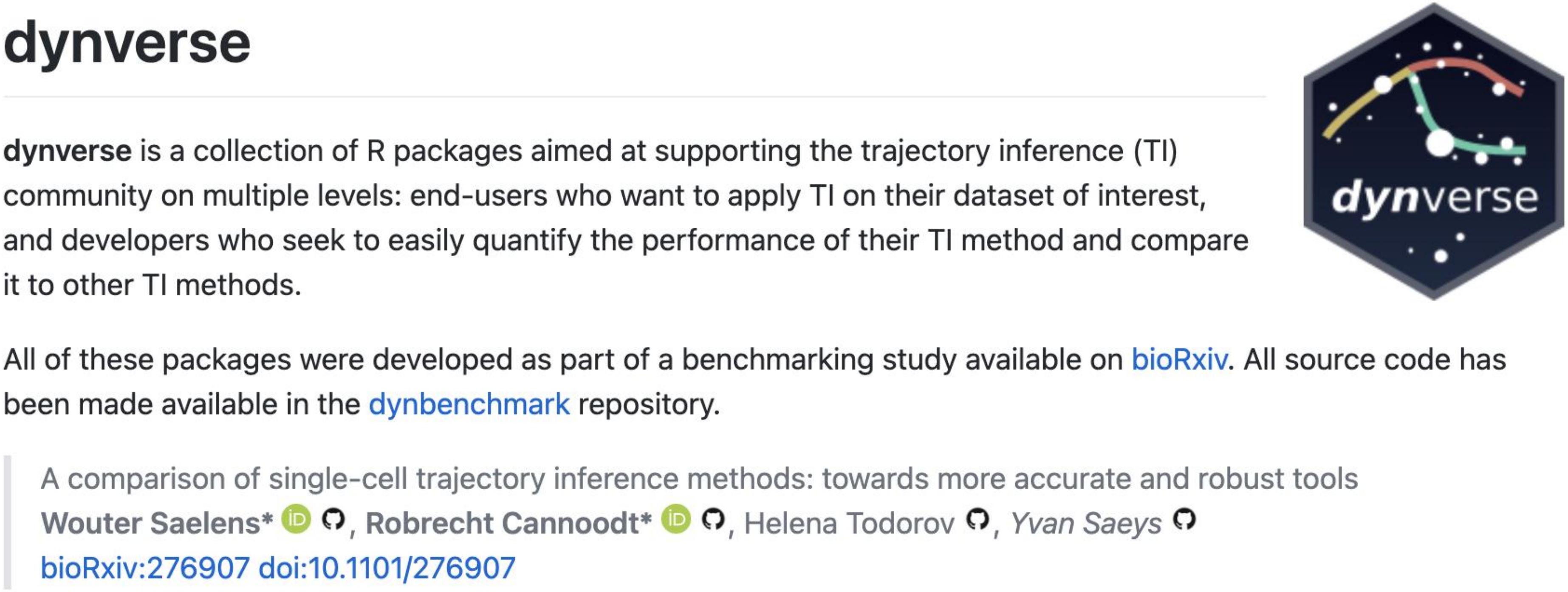
Vieth *et al*., Bioinformatics 2017

Splatter



Zappia *et al*., Genome Biology 2017

Dynverse



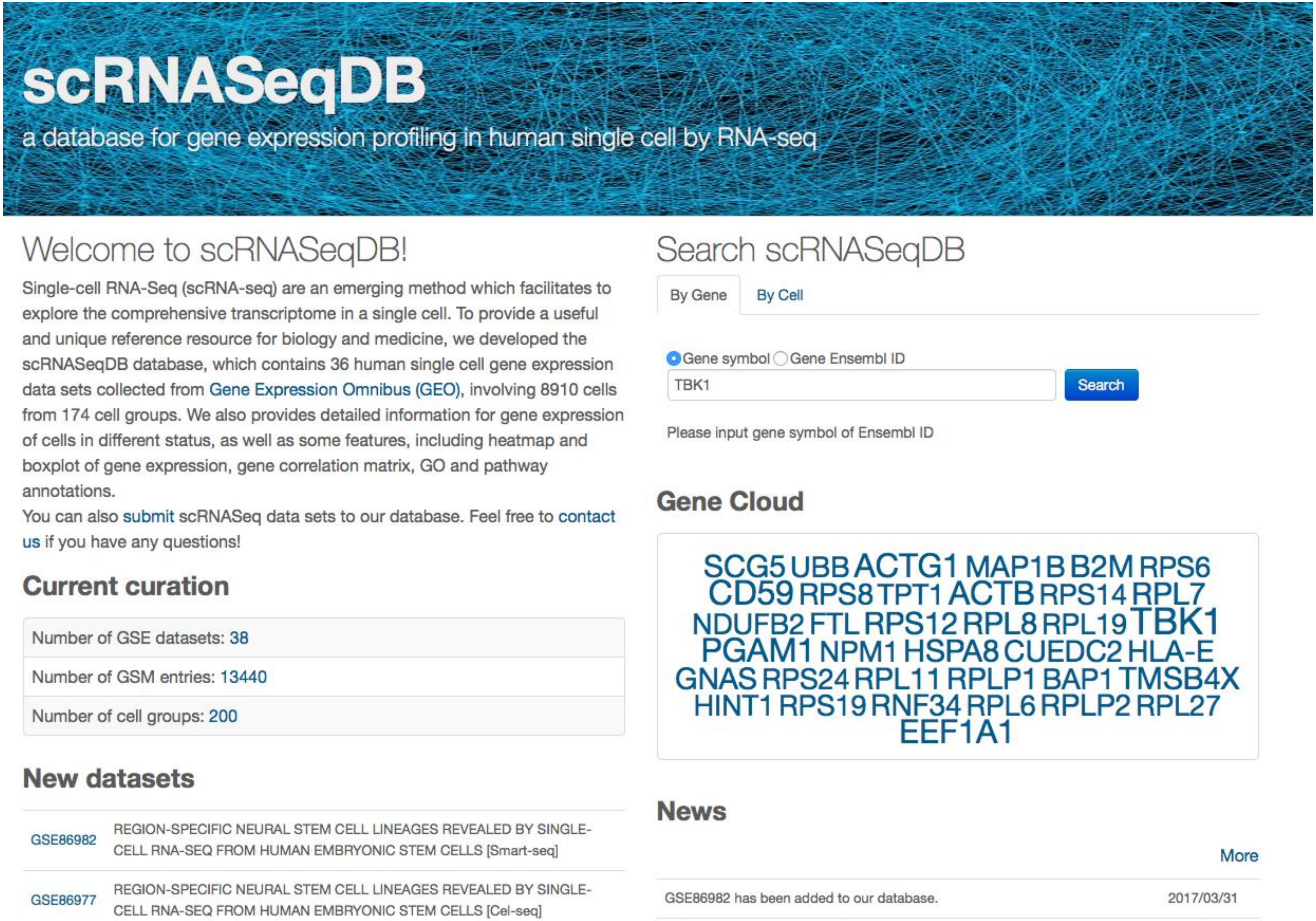
dynverse

dynverse is a collection of R packages aimed at supporting the trajectory inference (TI) community on multiple levels: end-users who want to apply TI on their dataset of interest, and developers who seek to easily quantify the performance of their TI method and compare it to other TI methods.

All of these packages were developed as part of a benchmarking study available on bioRxiv. All source code has been made available in the dynbenchmark repository.

> A comparison of single-cell trajectory inference methods: towards more accurate and robust tools
> **Wouter Saelens***, **Robrecht Cannoodt***, Helena Todorov, *Yvan Saeys*
> bioRxiv:276907 doi:10.1101/276907

https://github.com/dynverse/dynverse



scRNASeqDB

a database for gene expression profiling in human single cell by RNA-seq

Welcome to scRNASeqDB!

Single-cell RNA-Seq (scRNA-seq) are an emerging method which facilitates to explore the comprehensive transcriptome in a single cell. To provide a useful and unique reference resource for biology and medicine, we developed the scRNASeqDB database, which contains 36 human single cell gene expression data sets collected from Gene Expression Omnibus (GEO), involving 8910 cells from 174 cell groups. We also provides detailed information for gene expression of cells in different status, as well as some features, including heatmap and boxplot of gene expression, gene correlation matrix, GO and pathway annotations.

You can also submit scRNASeq data sets to our database. Feel free to contact us if you have any questions!

Current curation

Number of GSE datasets:	38
Number of GSM entries:	13440
Number of cell groups:	200

New datasets

GSE86982	REGION-SPECIFIC NEURAL STEM CELL LINEAGES REVEALED BY SINGLE-CELL RNA-SEQ FROM HUMAN EMBRYONIC STEM CELLS [Smart-seq]
GSE86977	REGION-SPECIFIC NEURAL STEM CELL LINEAGES REVEALED BY SINGLE-CELL RNA-SEQ FROM HUMAN EMBRYONIC STEM CELLS [Cel-seq]

Search scRNASeqDB

● Gene symbol ○ Gene Ensembl ID

TBK1

Please input gene symbol of Ensembl ID

Gene Cloud

SCG5 UBB ACTG1 MAP1B B2M RPS6
CD59 RPS8 TPT1 ACTB RPS14 RPL7
NDUFB2 FTL RPS12 RPL8 RPL19 TBK1
PGAM1 NPM1 HSPA8 CUEDC2 HLA-E
GNAS RPS24 RPL11 RPLP1 BAP1 TMSB4X
HINT1 RPS19 RNF34 RPL6 RPLP2 RPL27
EEF1A1

News

GSE86982 has been added to our database.	2017/03/31

https://bioinfo.uth.edu/scrnaseqdb/index.php?r=site/index

Questions?