Clinical and Research Genomics
Lecture 04-05
RNA-Sequencing, Epitranscriptomes, Single Cells, and Spatial Omics

Dr. Christopher E. Mason

-

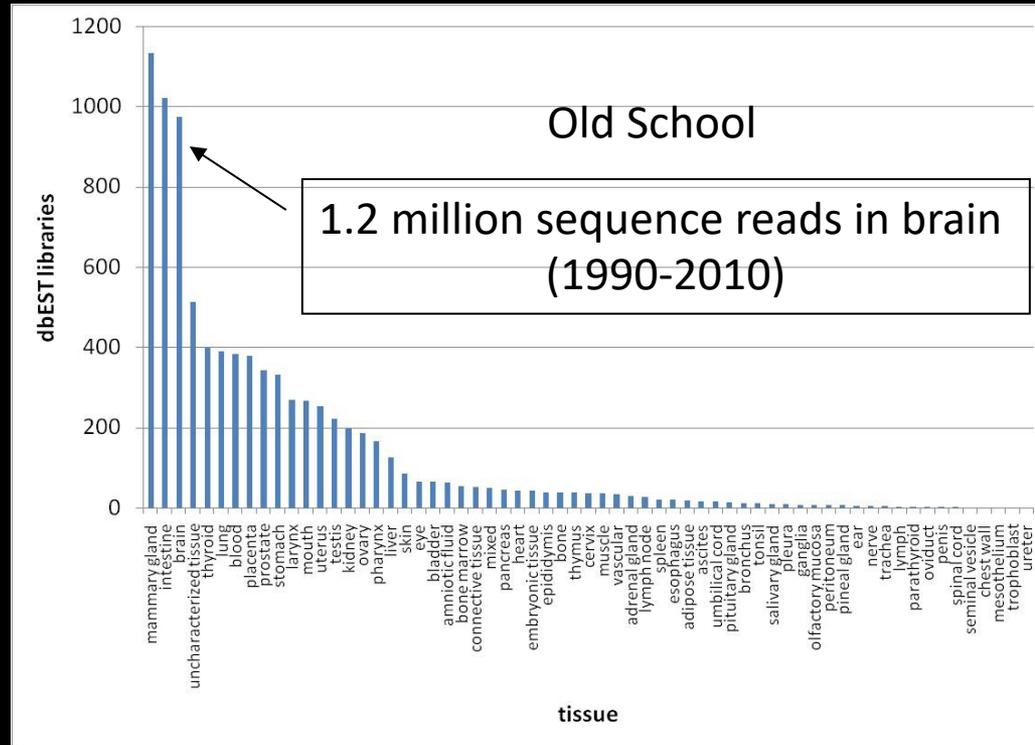March 16, 2021

(1)

Background

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{Protein}$$

James Watson, 1958

Epigenome

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein
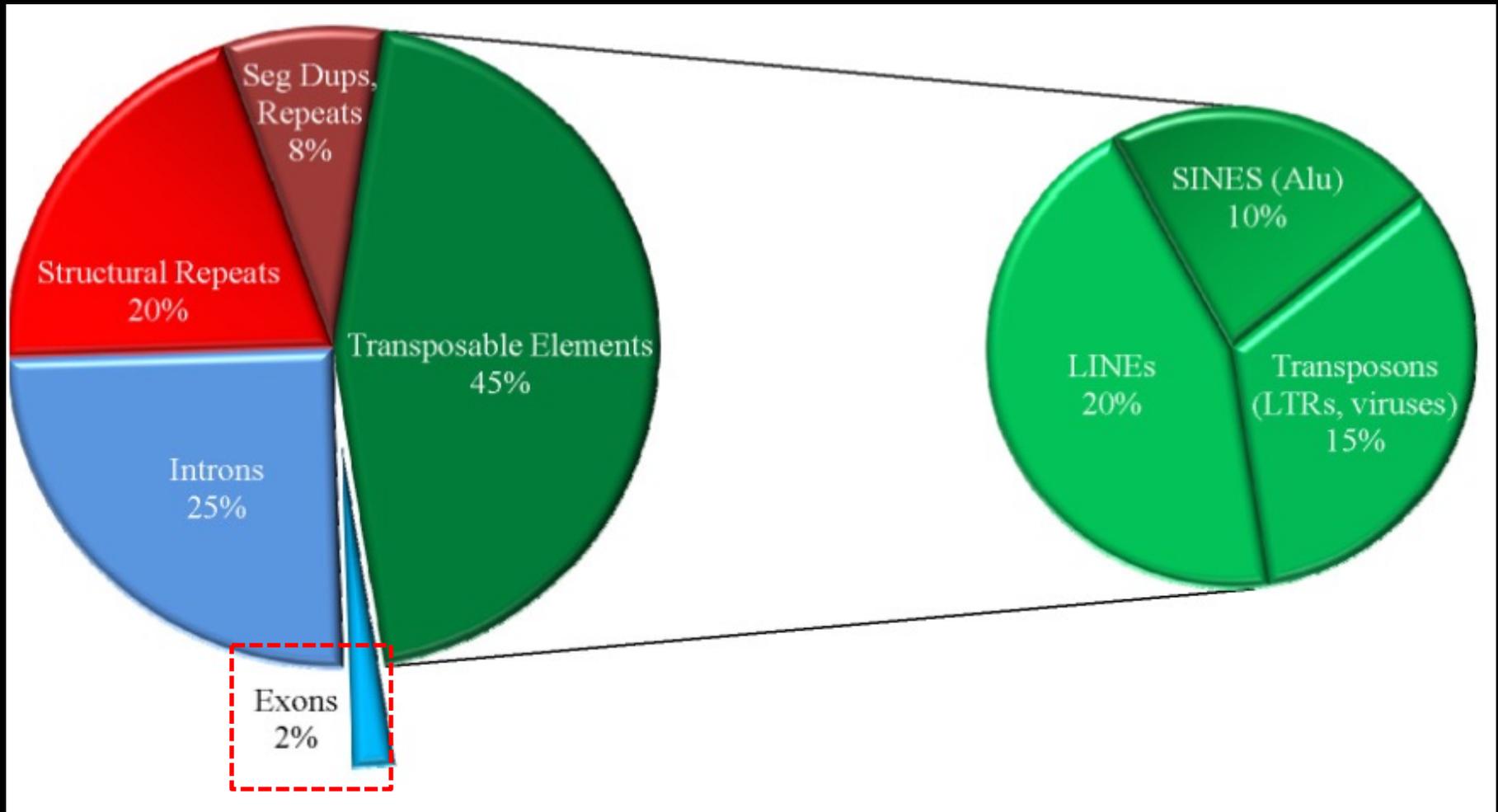
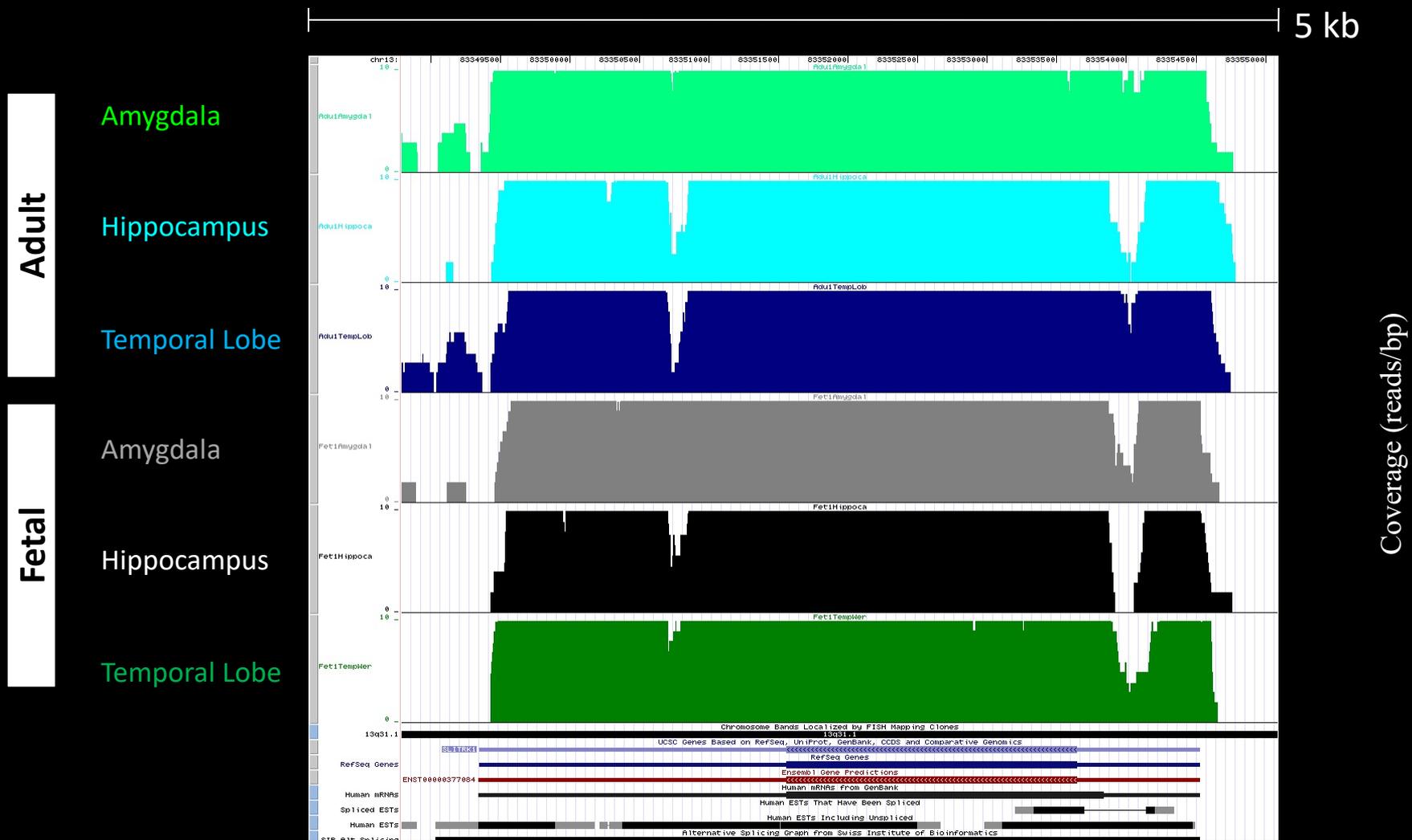# We can observe
many, many more molecules than before

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

The Annotation/Composition of the Human Genome

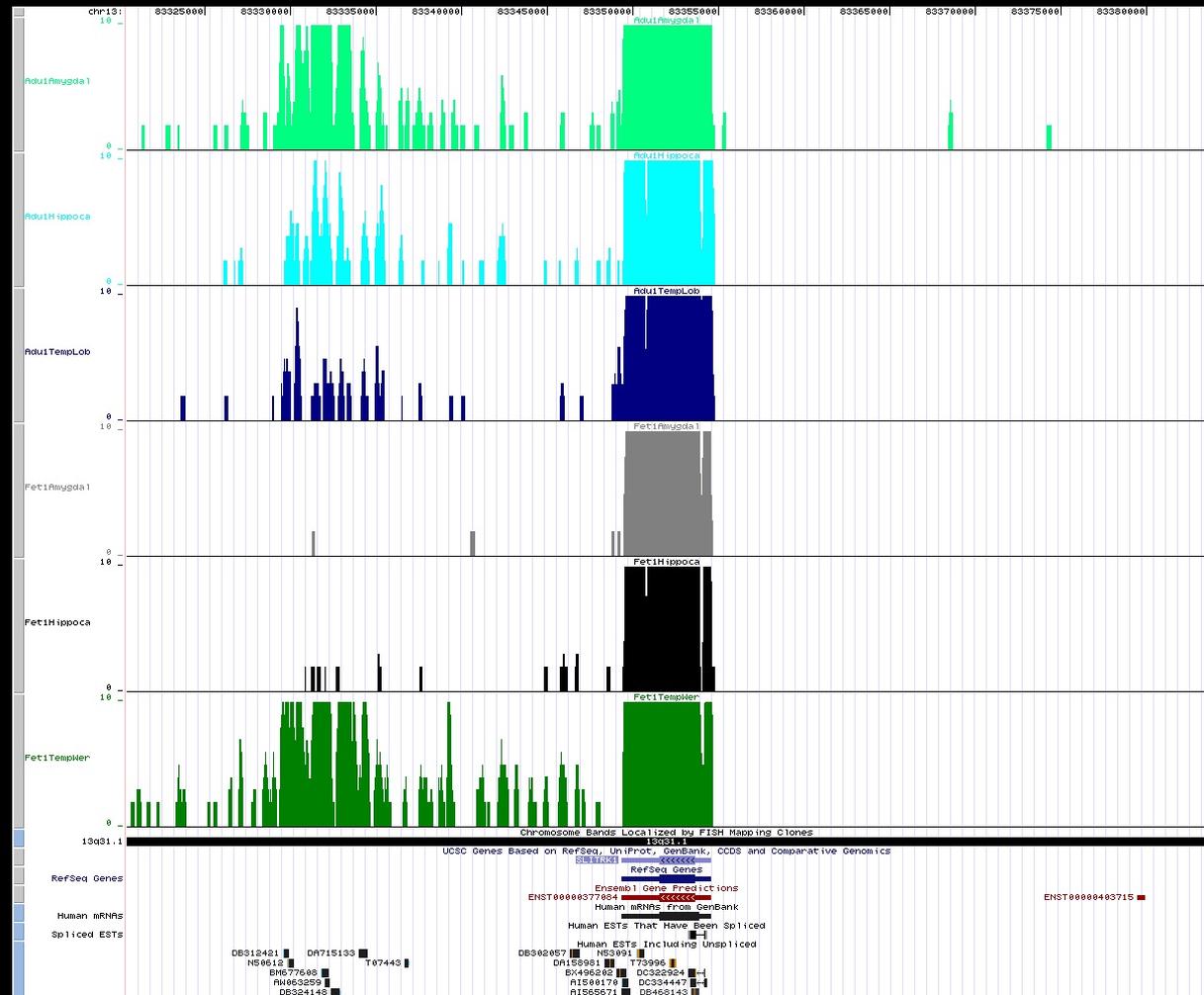# Validation of known Gene Boundaries

Find Longer Isoforms

Find New Genes
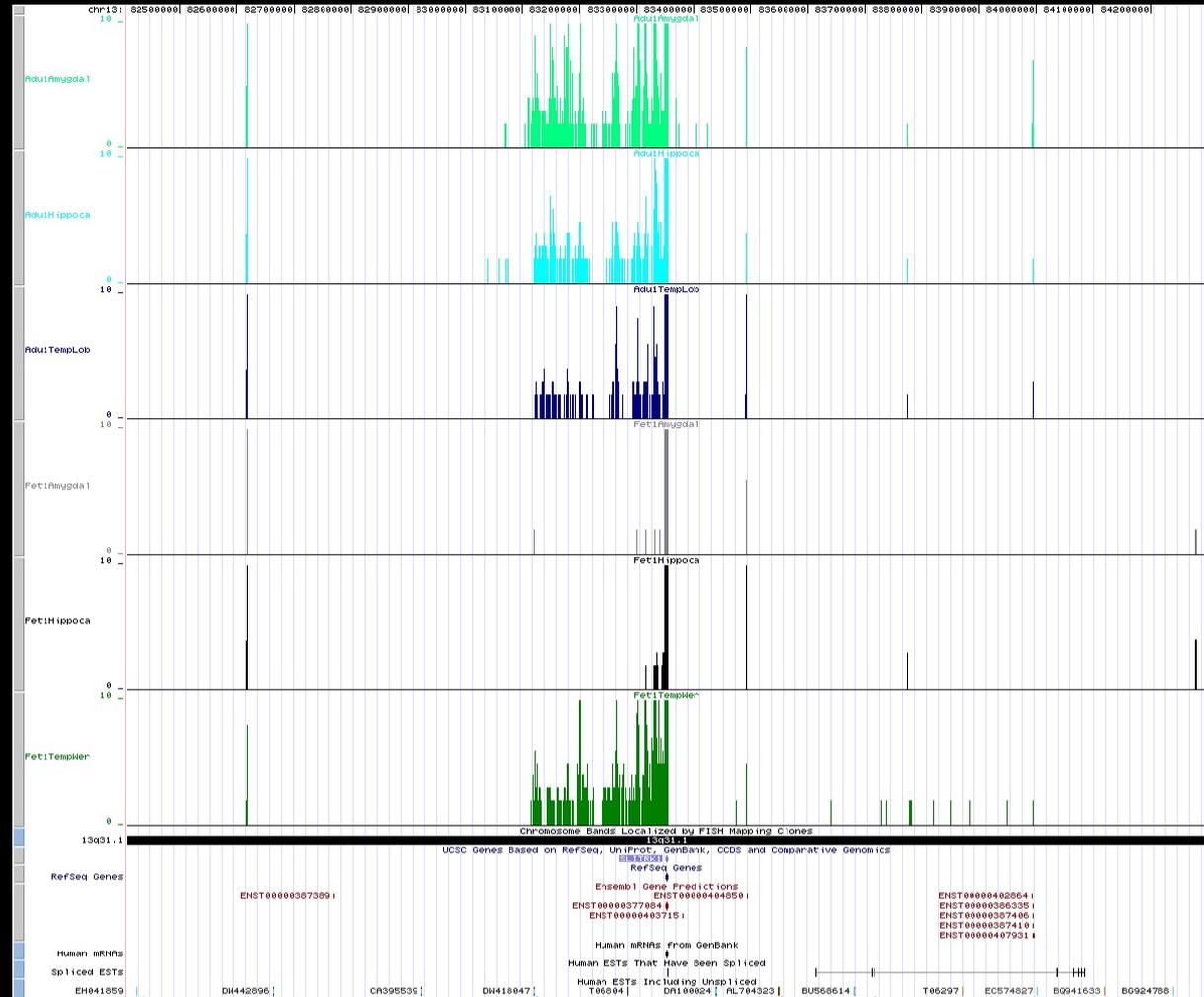
About Half of the Noncoding Genome is Transcriptionally Active

Mason, 2006

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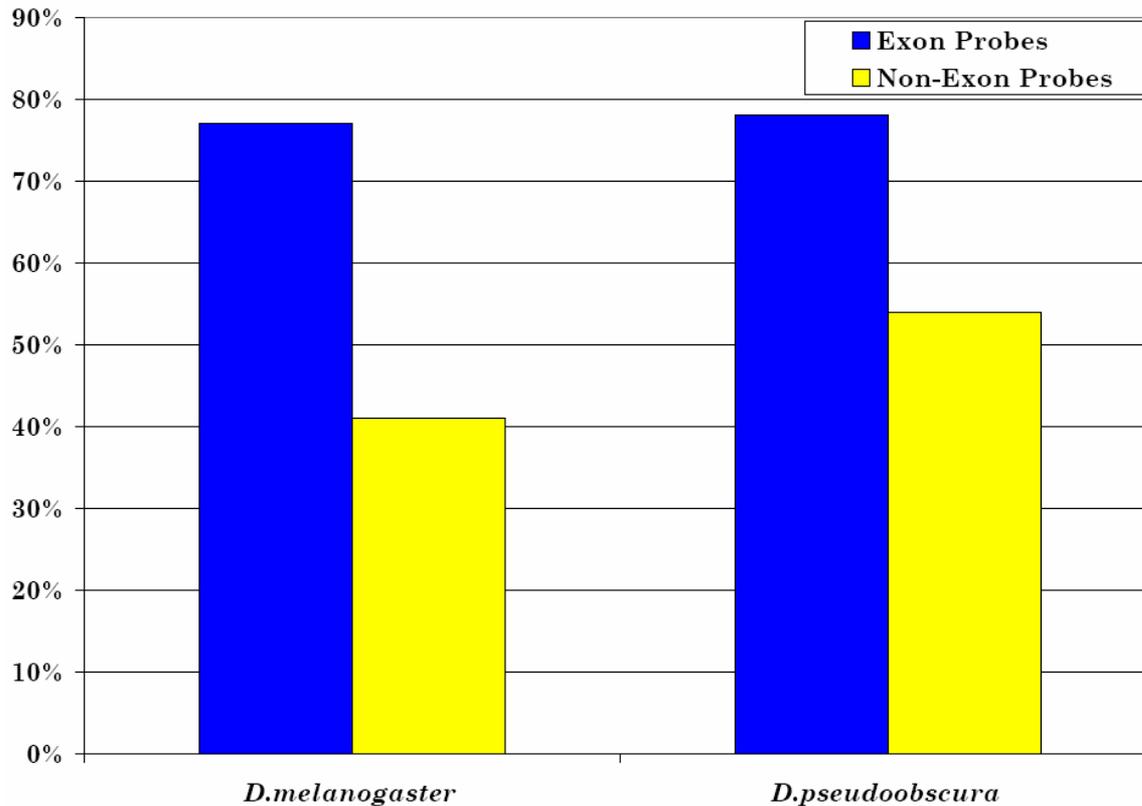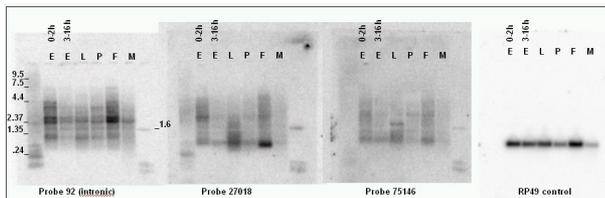
Humans: 47% (Schadt *et al.,* 2004*)*
Arabadopsis: 51% (Yamada *et al.*, 2003)

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
$2^n - 1$		$\dfrac{n(n-1)}{2}$

Exon 1   Exon 2   Exon 3   Exon4

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

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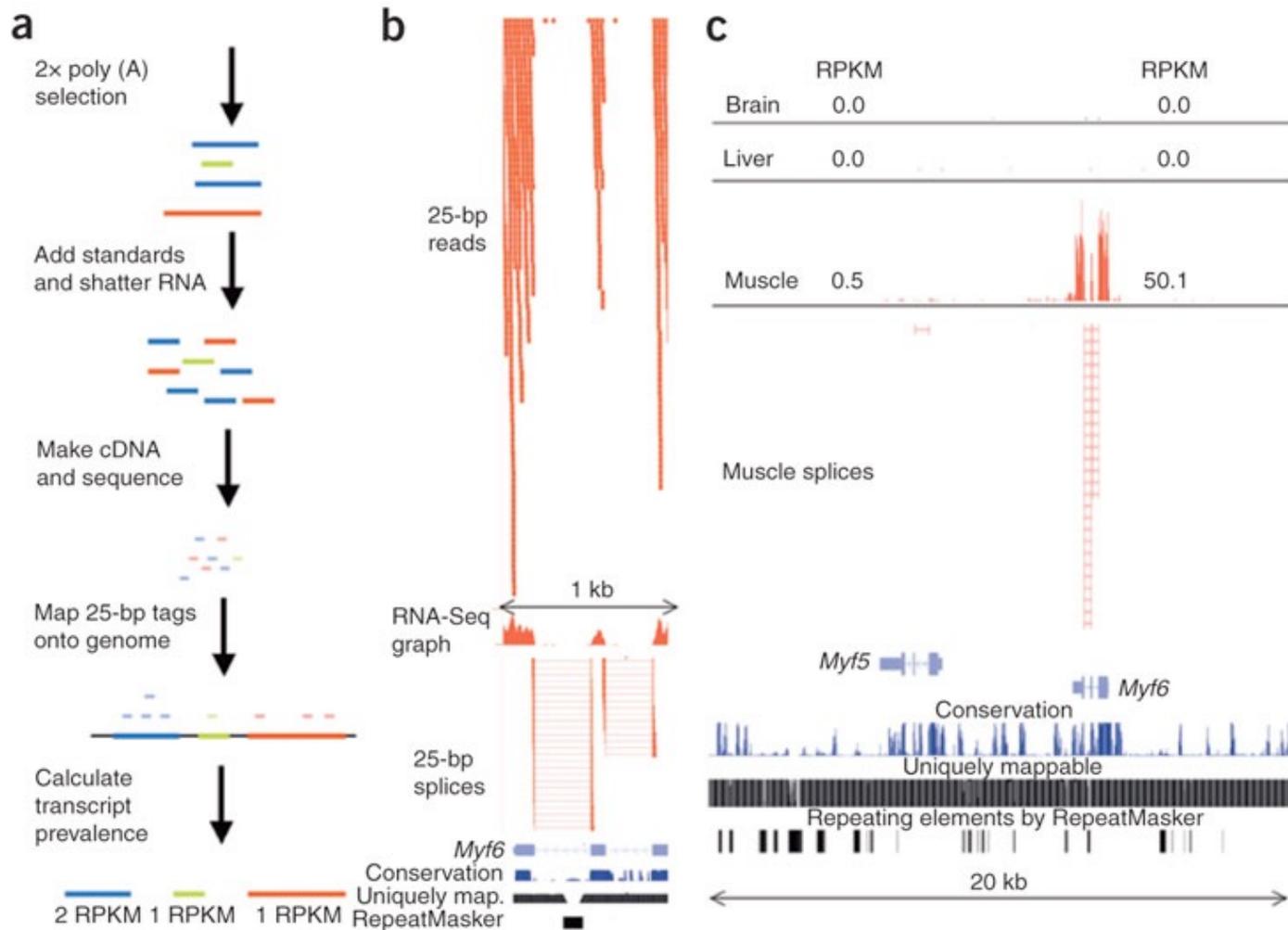*8×10^{80} atoms in the universe*
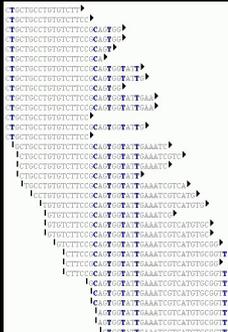*(1^{59} atoms/star, 1^{11} stars/galaxy, 1^{10} galaxies)*

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]
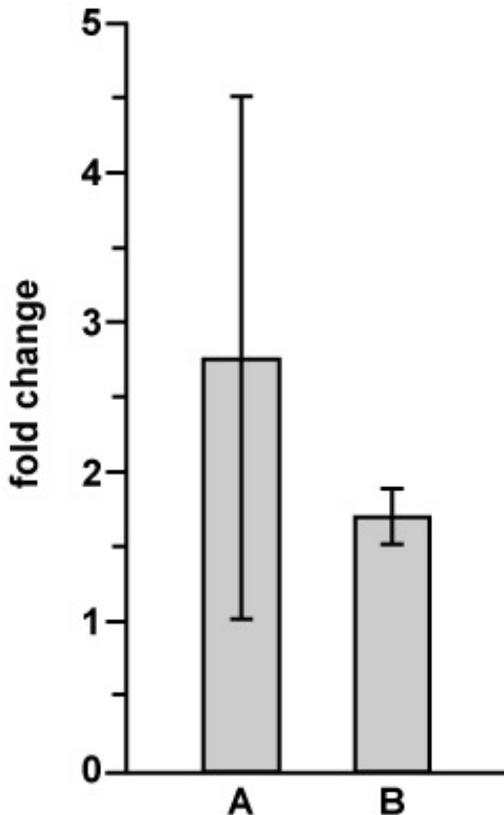

2008

Can RNA-Seq replace microarrays?

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

# 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

Comparing Solexa and Affymetrix

Spearman correlation = 0.72
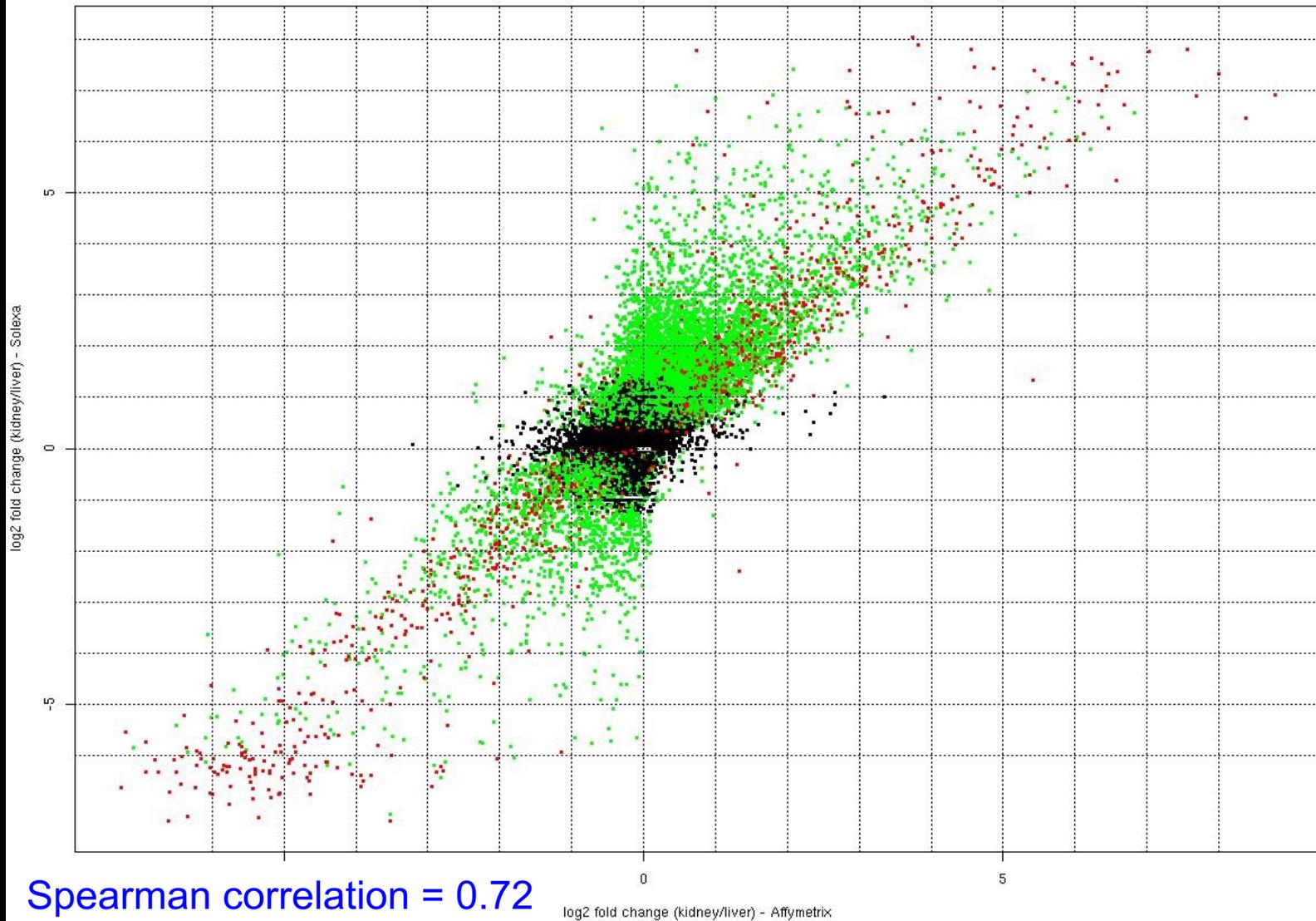
13,072 Differentially Expressed (DE) Genes

(37%)
DE in
Solexa
4,959

(50%)
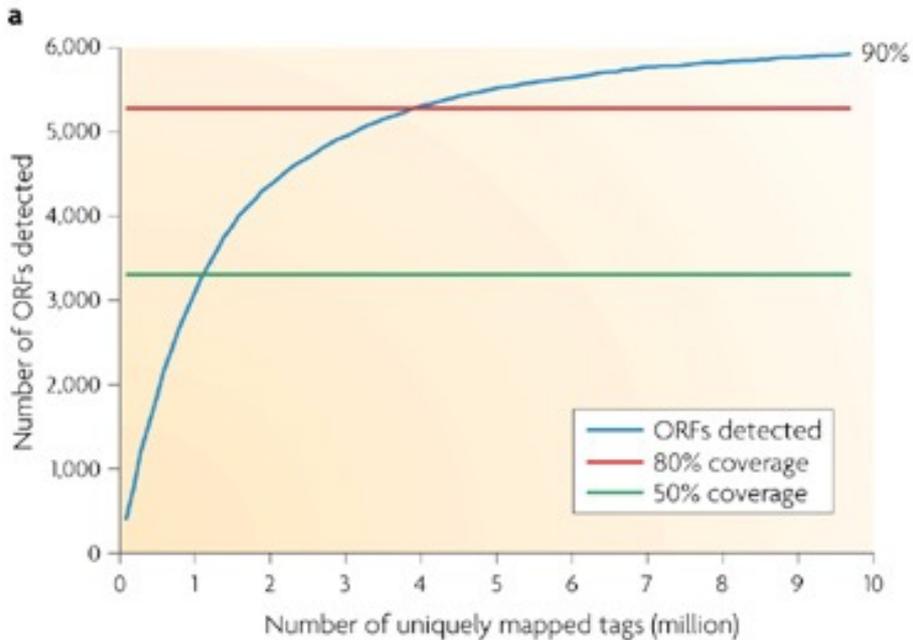Both
6,534

(12%)
DE in
Affy
1,579
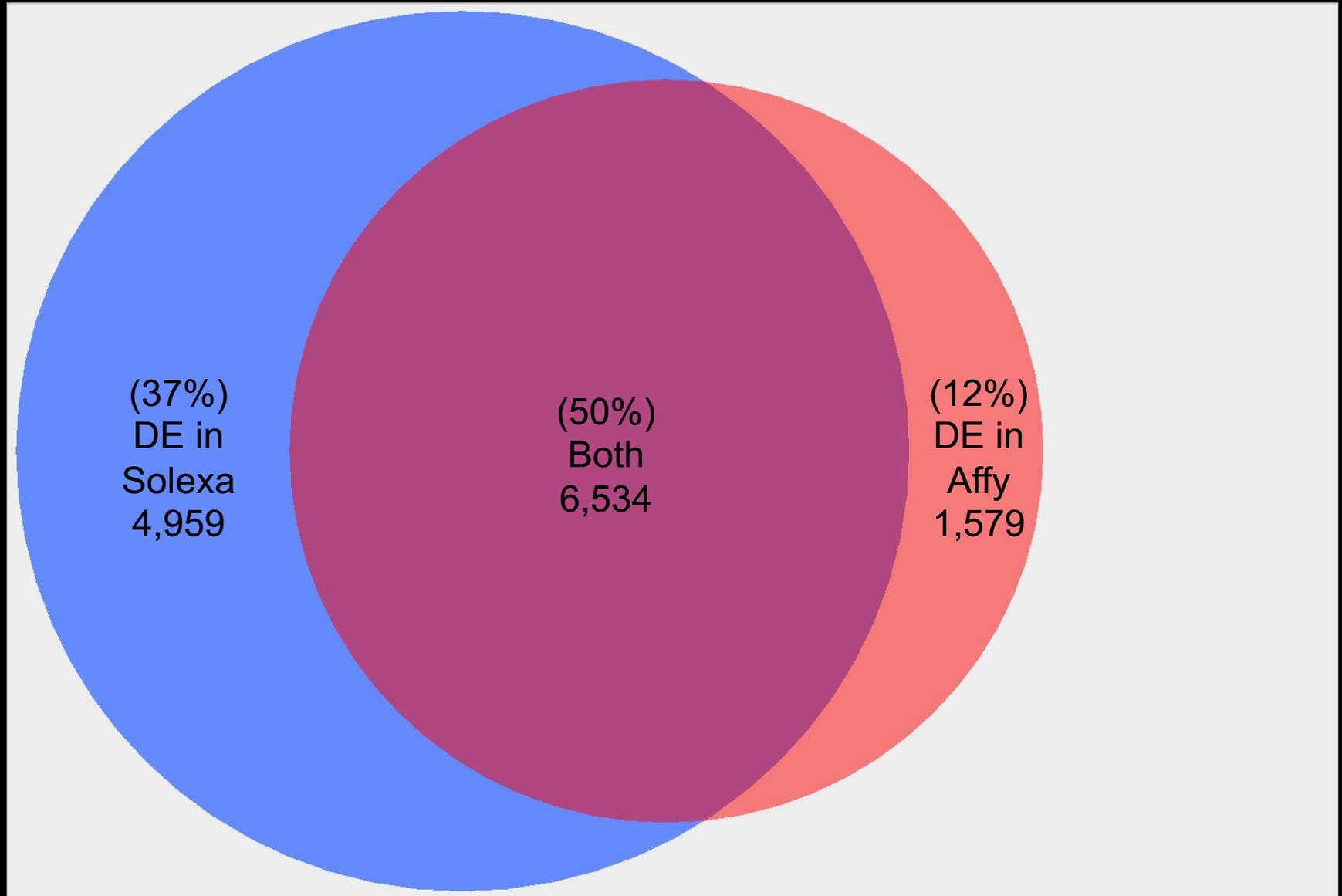
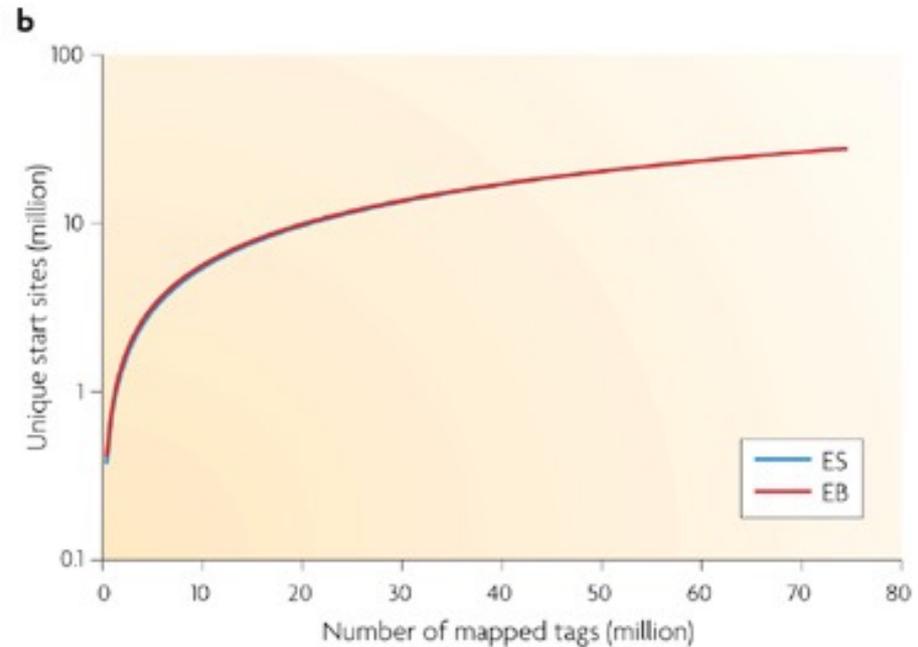# Coverage Requirements: How many lanes/plates/wells?

<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

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

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

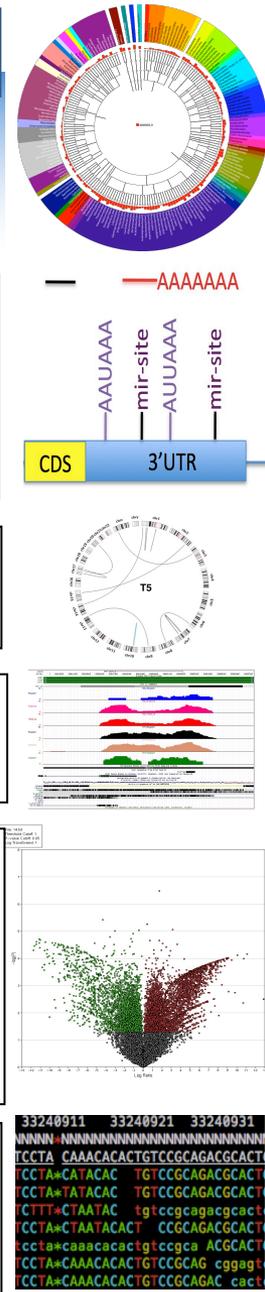| Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

Accurate gene quantification requires greater depth than gene discovery

Toung et al, 2011
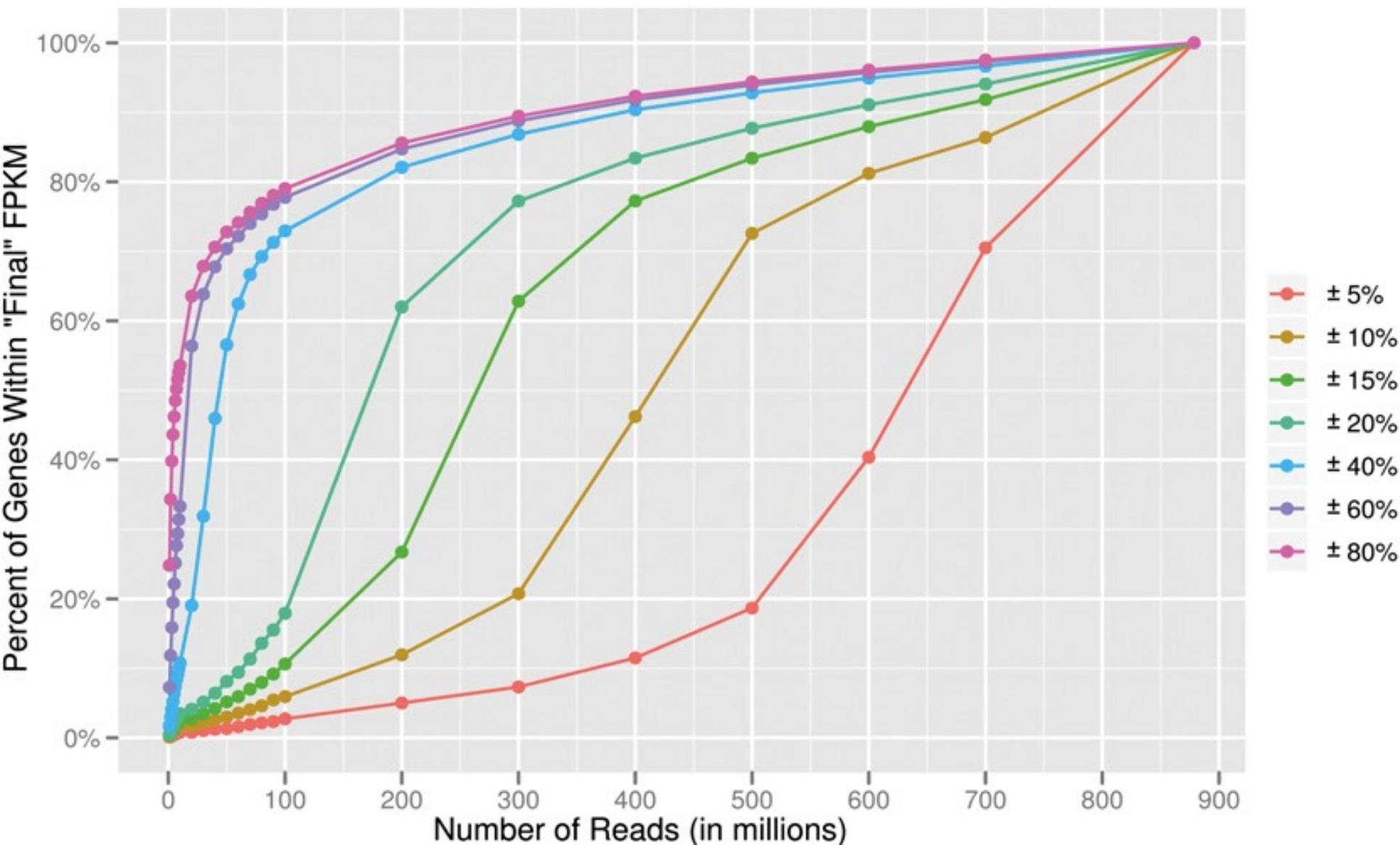
(3)

Tools & Standards

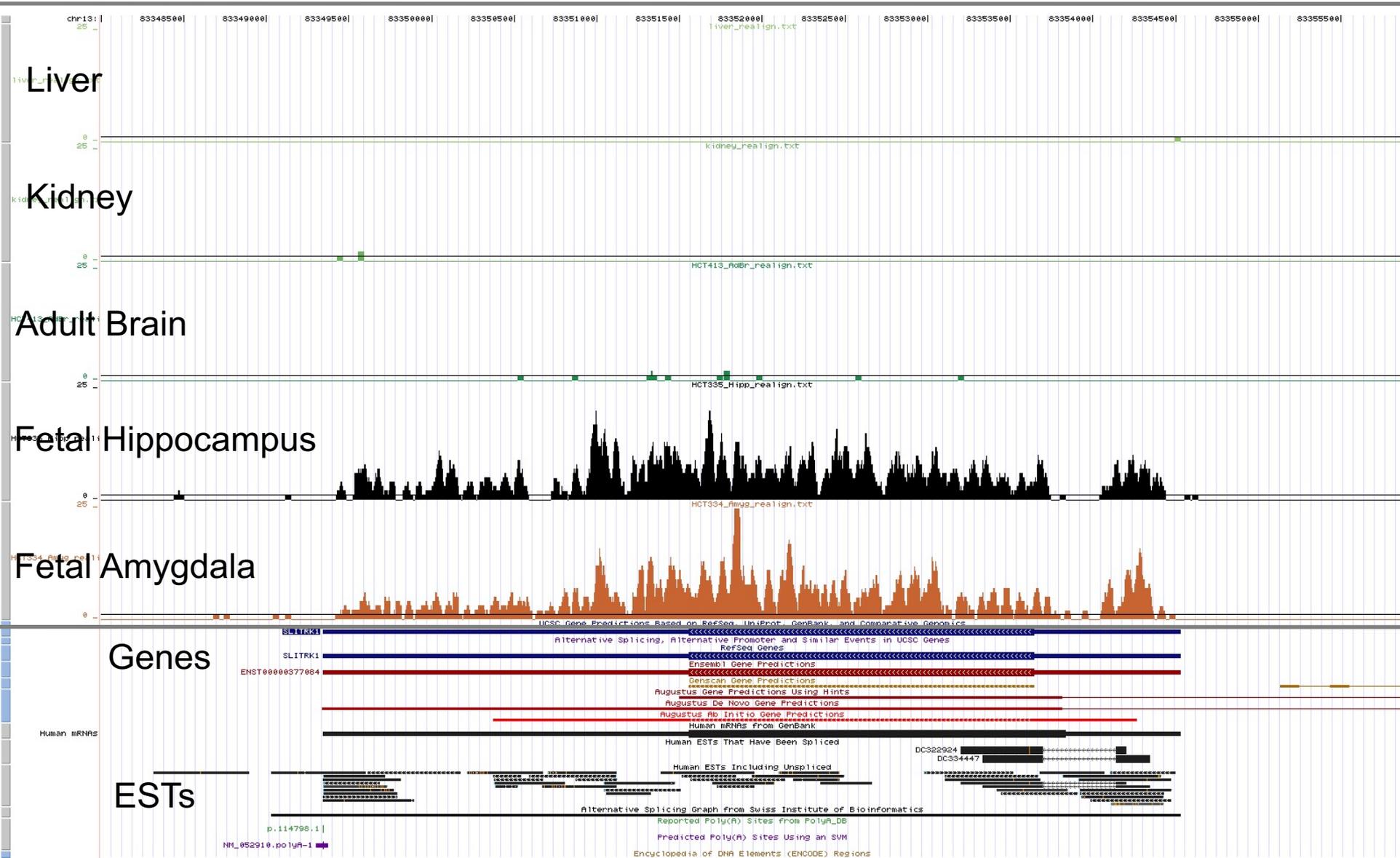# RNA-Seq and all its flavors

create excitement
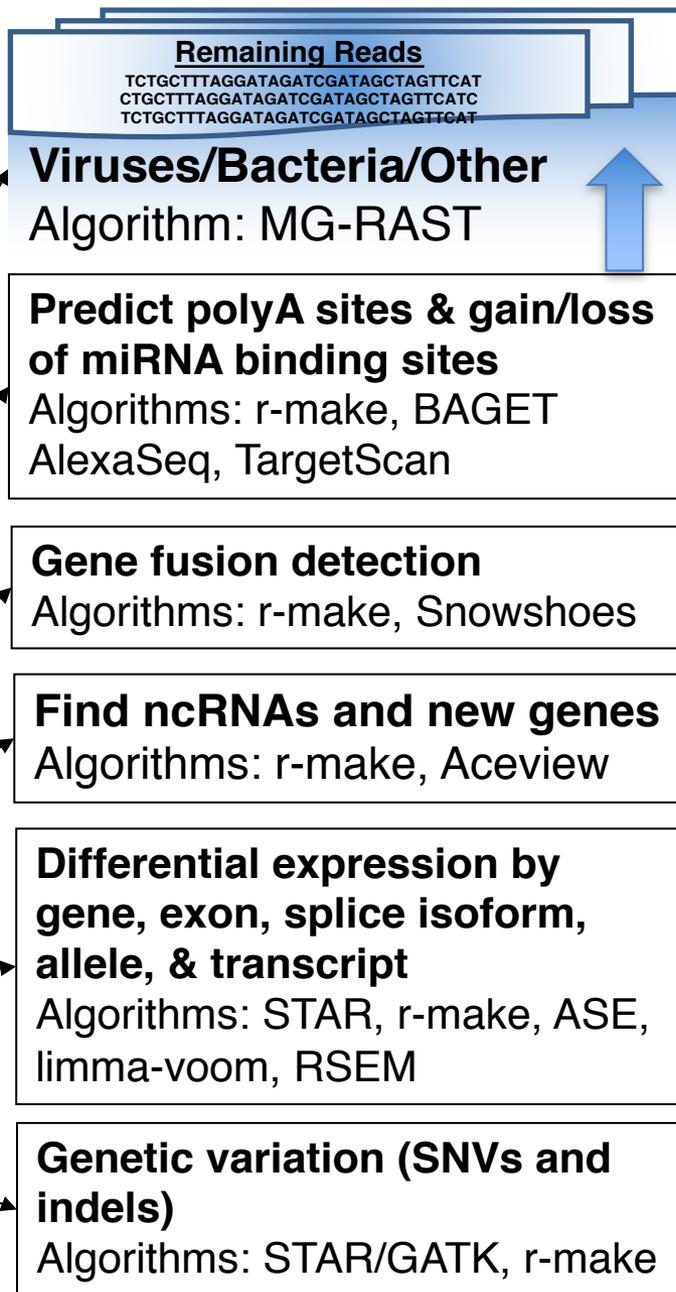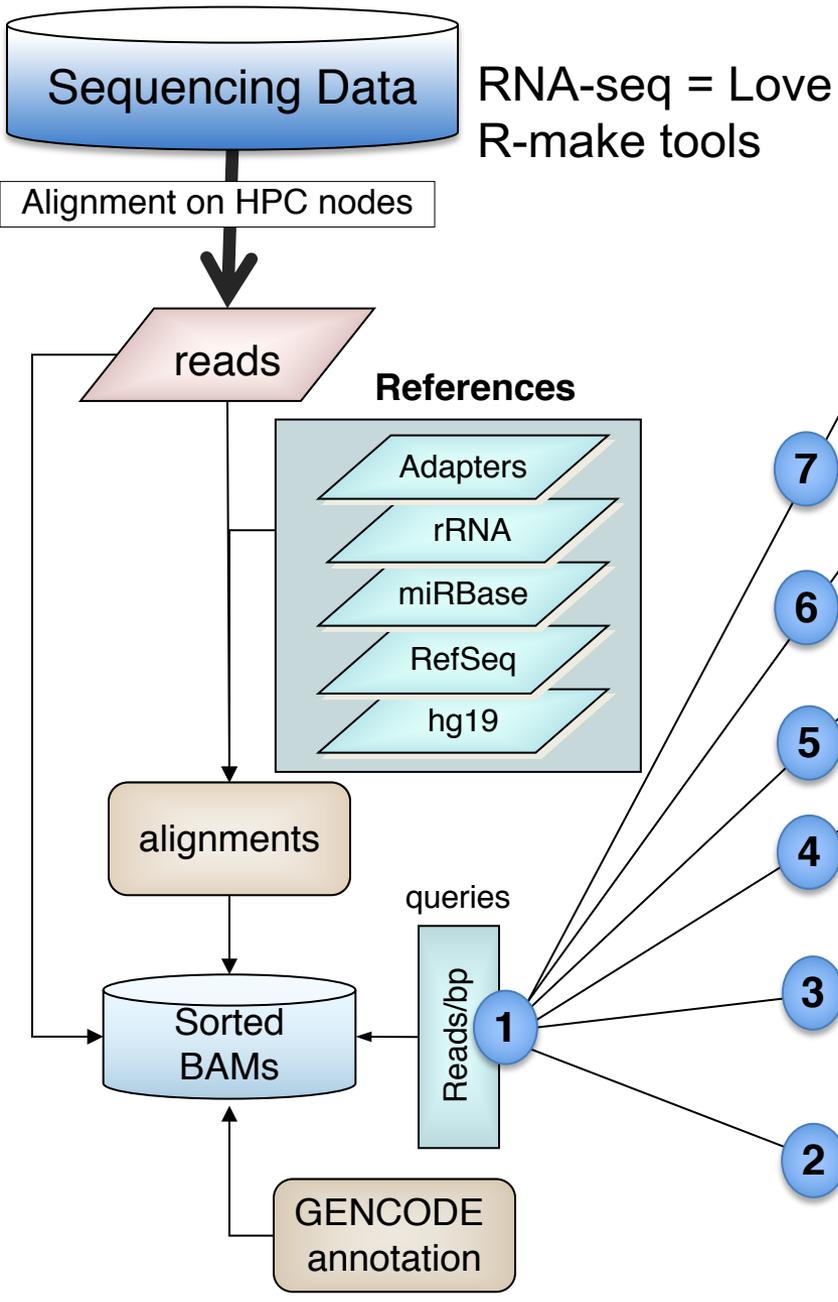
But!

There is some noise

What is the source of the wiggles?

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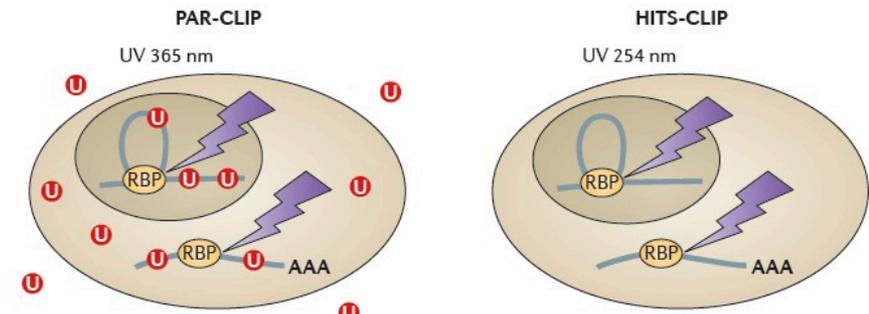
HITS-CLIP: genome-wide means of mapping protein–RNA binding sites in vivo.

PAR-CLIP: identifying the binding sites of cellular RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) in tissue culture cells.
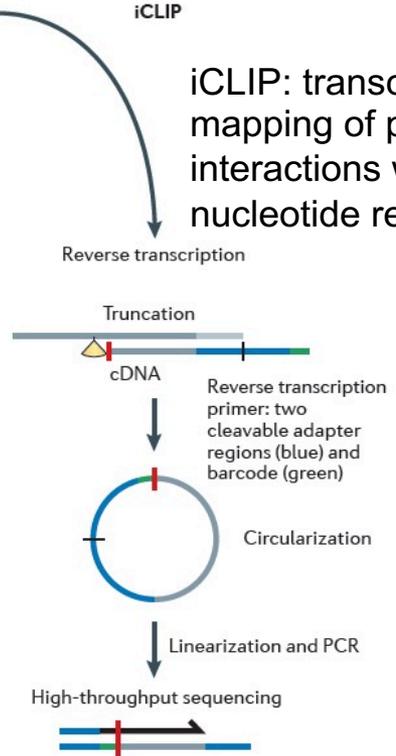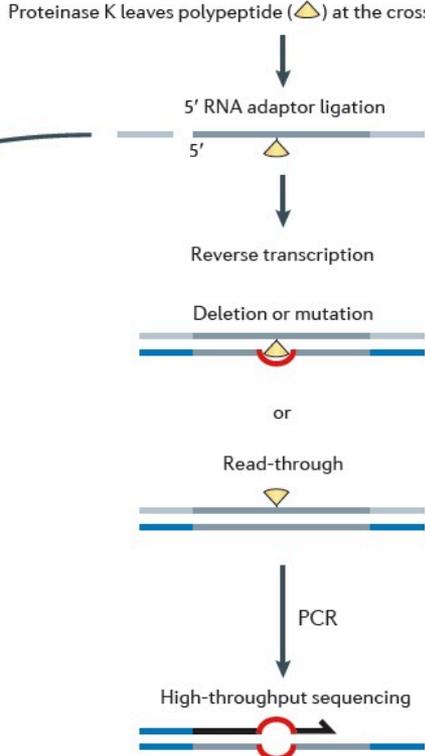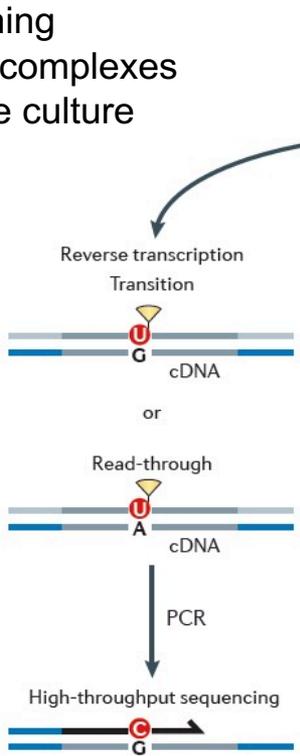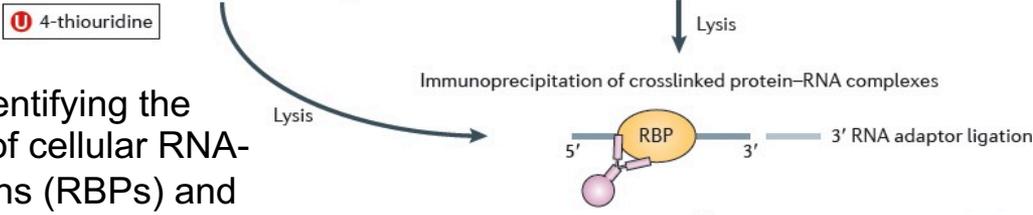
iCLIP: transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution.

Konig 2012

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, "Methods for RNA Isolation, Characterization and Sequencing (RNA-Seq)."

RNAs can have structure/function all their own

- mFOLD/sFOLD

- RNAMotifScan

- RNAfold

Halobacterium halobium SRP RNA
(SRPDB, March 10, 2000)
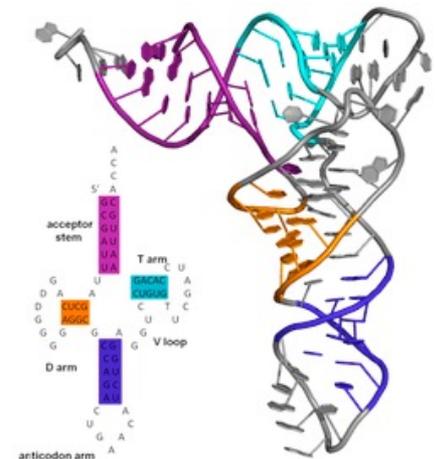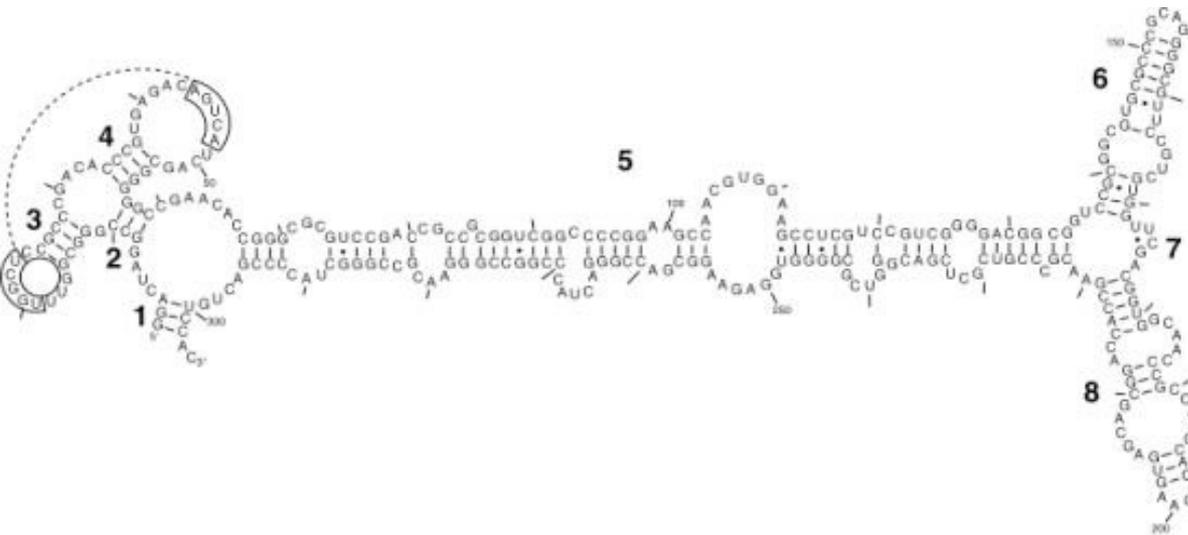
And - which one do we use?
Technologies Bifurcate into two main realms:

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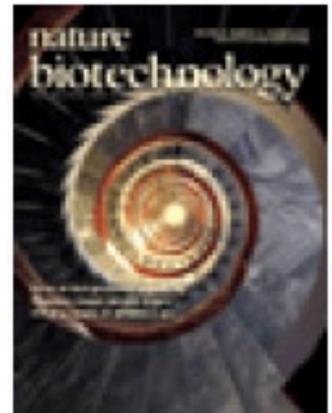
"… a related endeavor that would help better benchmark the different next-generation sequencing technologies would be to carry out an initiative similar to the Microarray Quality Control [MAQC] consortium where different platforms would be compared against one another as well as against DNA microarrays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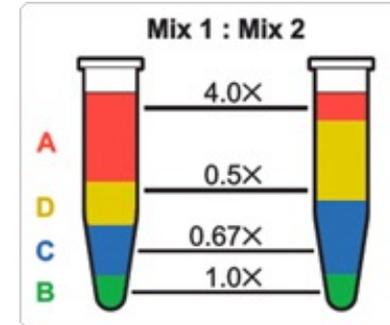
There is some hope from at least five places:

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)

*"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
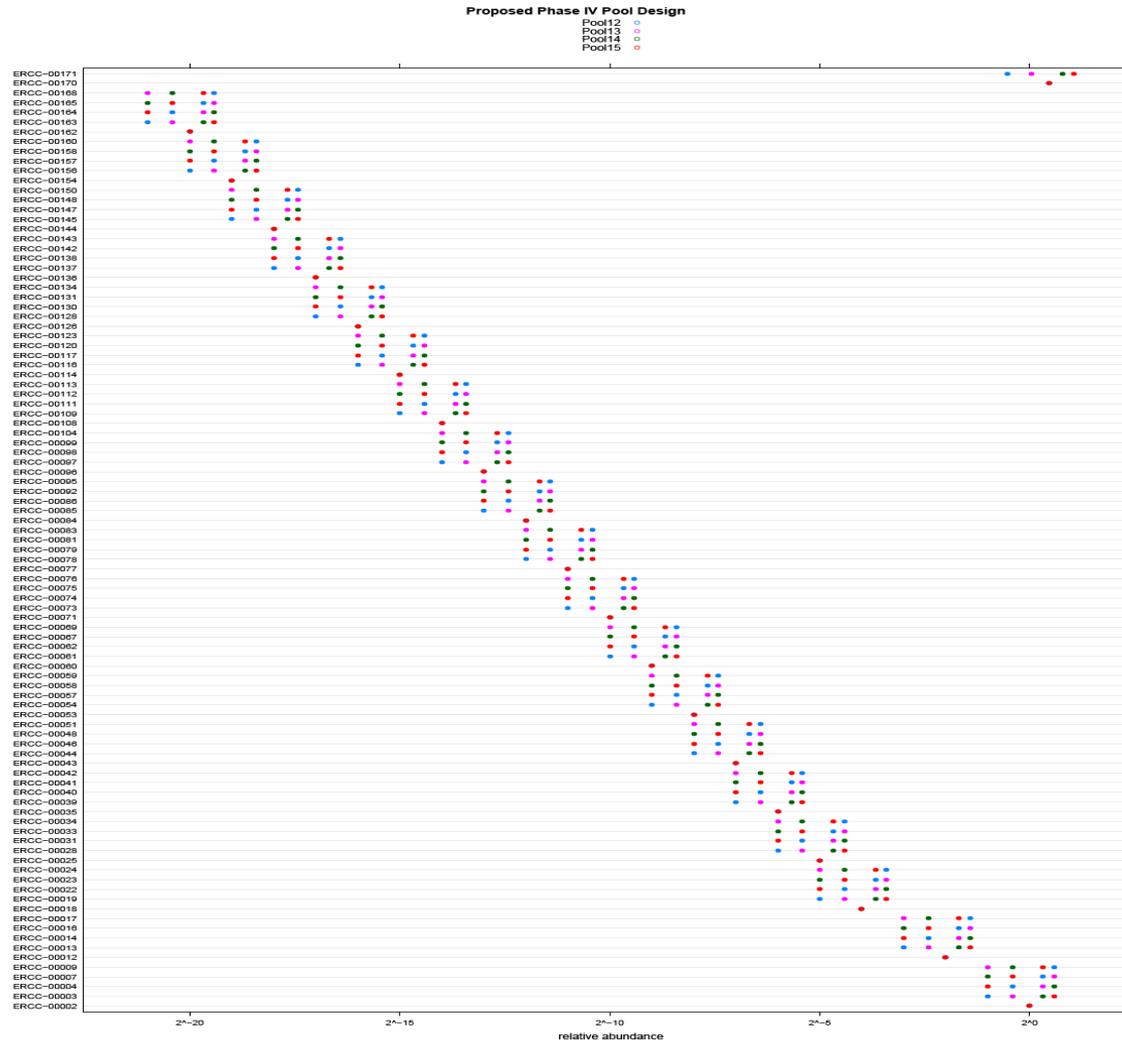
Samples of the MAQC, SEQC and ABRF-NGS Study

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

25

Ambion Human Brain Reference RNA (HBRR)

(B)

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

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

50 µg
200 µg
2.5 mg

Ambion

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

26

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

Association of Biomolecular Research Facilities (ABRF) Next Generation Sequencing (NGS) Study:

Phase I: RNA-Seq and degraded RNA-Seq (2011-2013)
Phase II: DNA-Seq and Hard-to-sequence regions (2013-2015)
Phase III: Asteroid and Martian Surface Sequencing (2015-2050?)

1. Cross Platform: HiSeq2000, 454, PGM, Proton, PacBio
2. Cross-Protocol: ribo-, directional, non-directional, degraded RNA
3. Cross-Site: 3 sites for each platform, replicates at each site

Improve the detection and quantification of molecular signatures for use in basic, applied, and clinical research.

Specifically, the ABRF-NGS group will test and characterize the myriad, rapidly evolving NGS tecehniques for RNA-Seq and DNA-

VOLUME 32 NUMBER 9 SEPTEMBER 2014
www.nature.com/naturebiotechnology

nature biotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

Focus on RNA sequencing quality control (SEQC)
ABRF evaluation of RNA-seq
Genome editing in hexaploid wheat
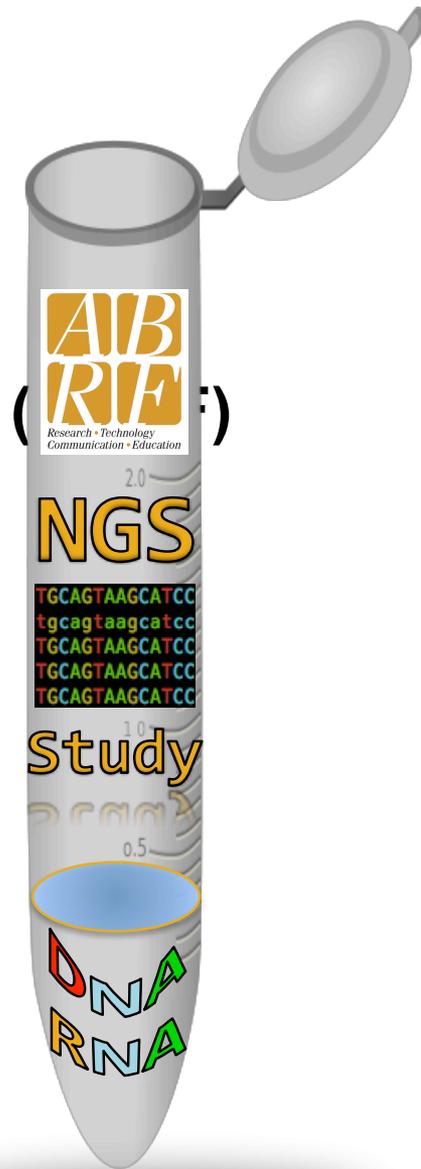
Current Results
Phase I: RNA Standards

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.

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

Special issue printed and hosted site

ABRF-NGS study

Gene coverage distributions reveal platform- and prep-specific effects
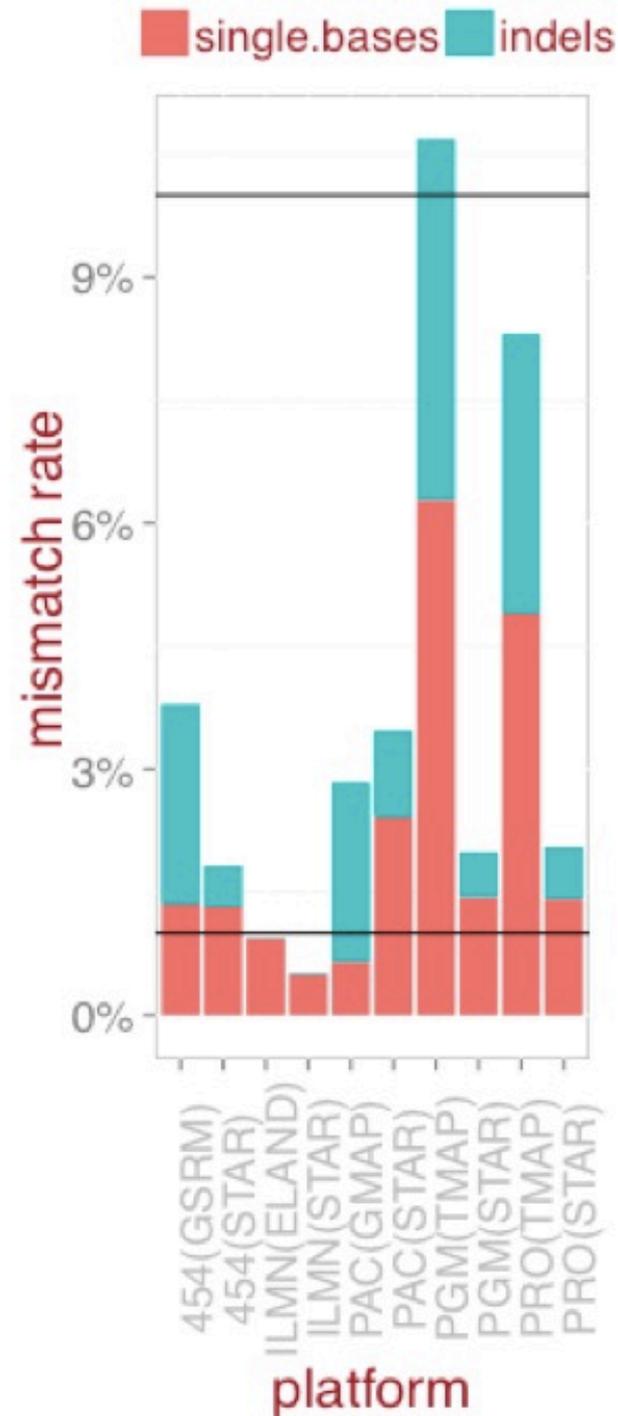
Error models highly variable among platforms

454: Roche 454 GS FLX+

ILMN: Illumina HiSeq 2000/2500

PAC: Pacific Biosciences RS I
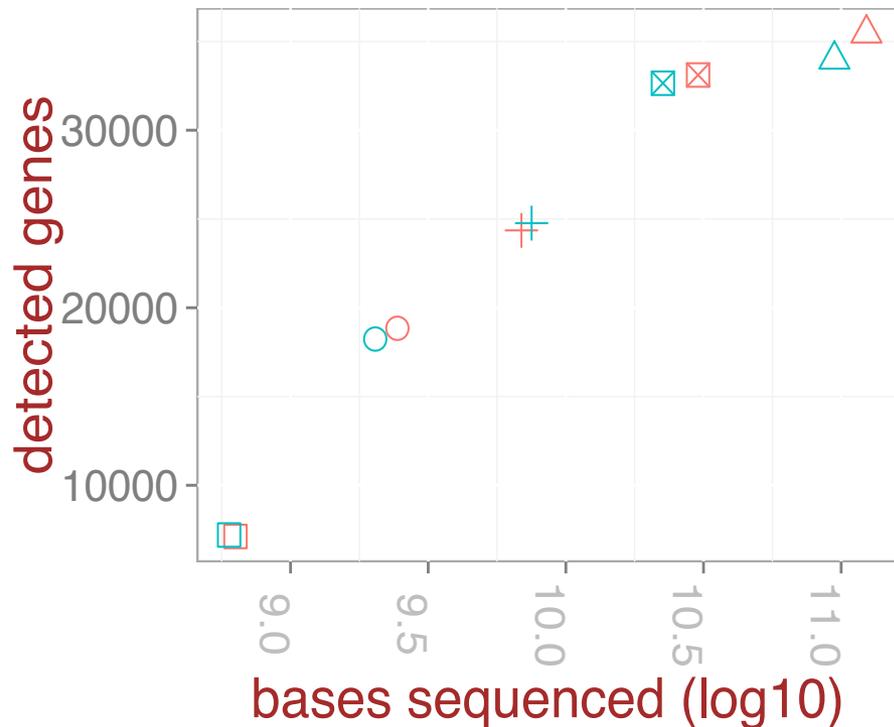
PGM: Ion Personal Genome Machine

PRO: Ion Torrent Proton

platform

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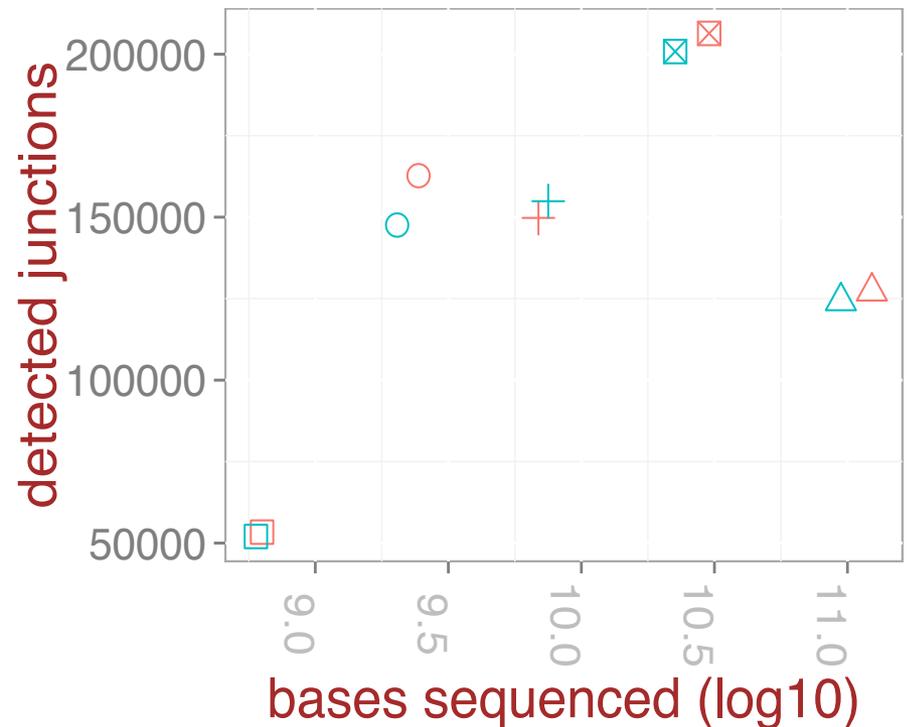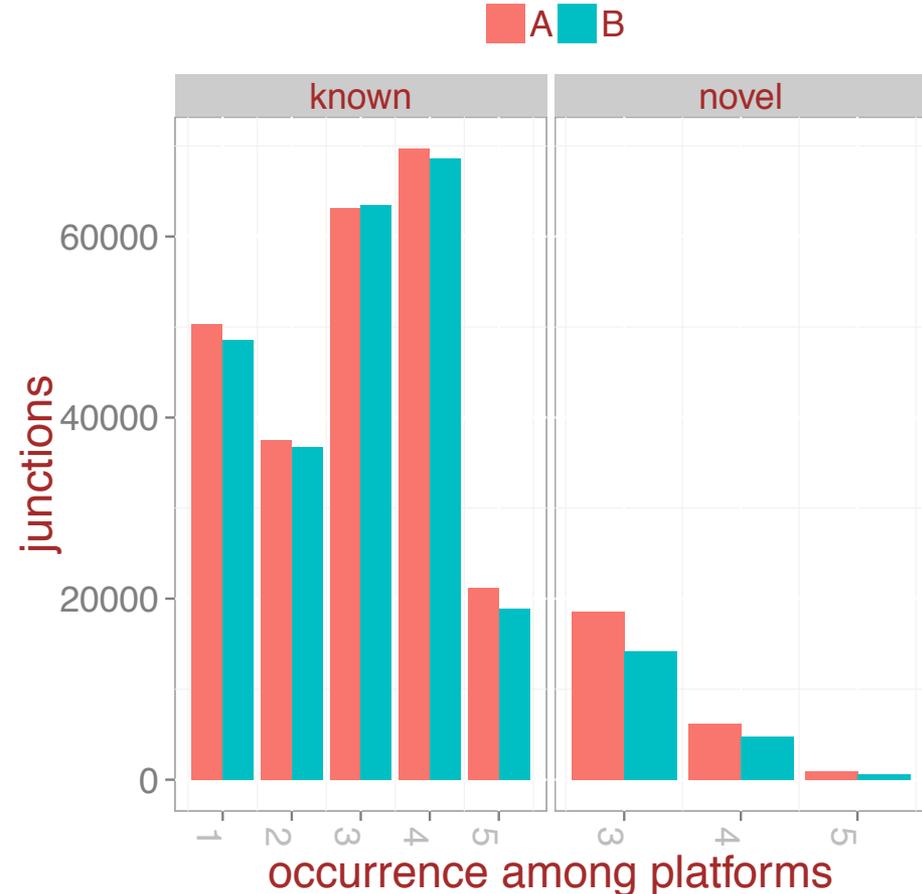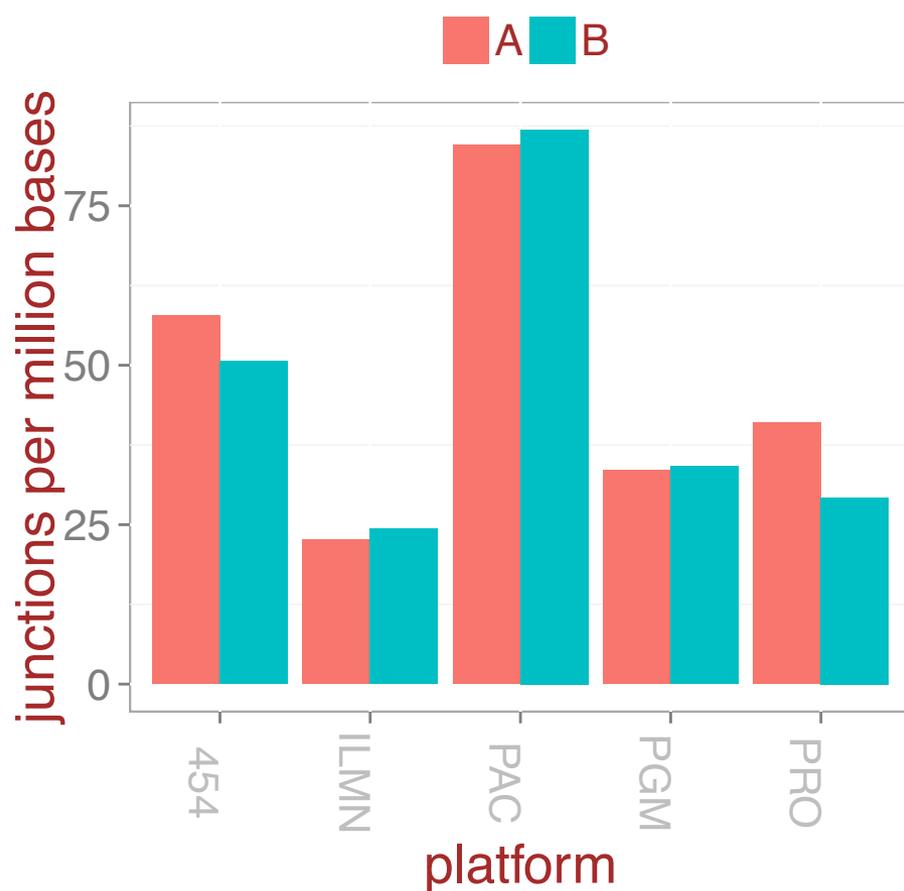
Note: Only use a subset of reads among platforms
to normalize the scale

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

Sequencing depth is important to discover low abundance transcripts
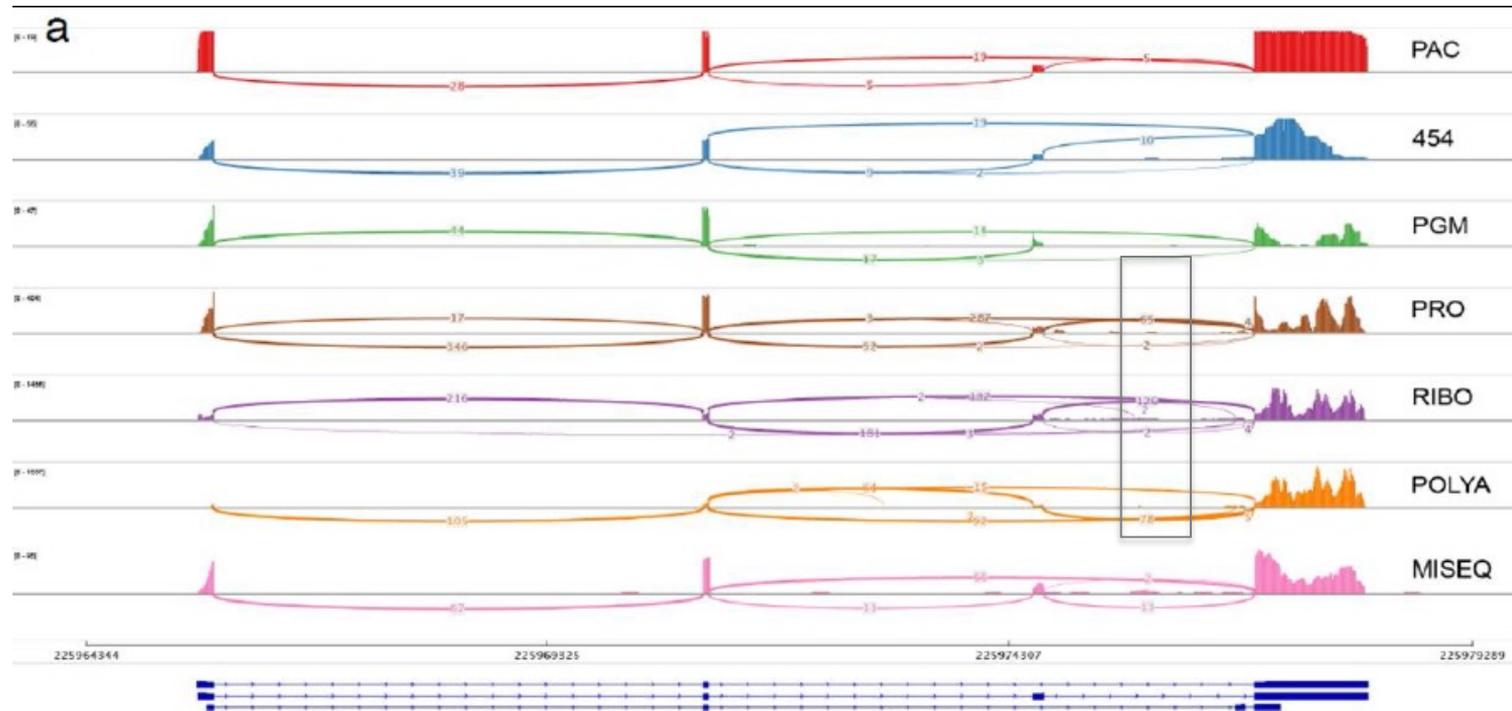
Inter-platform differential gene expression show 88-97% agreement
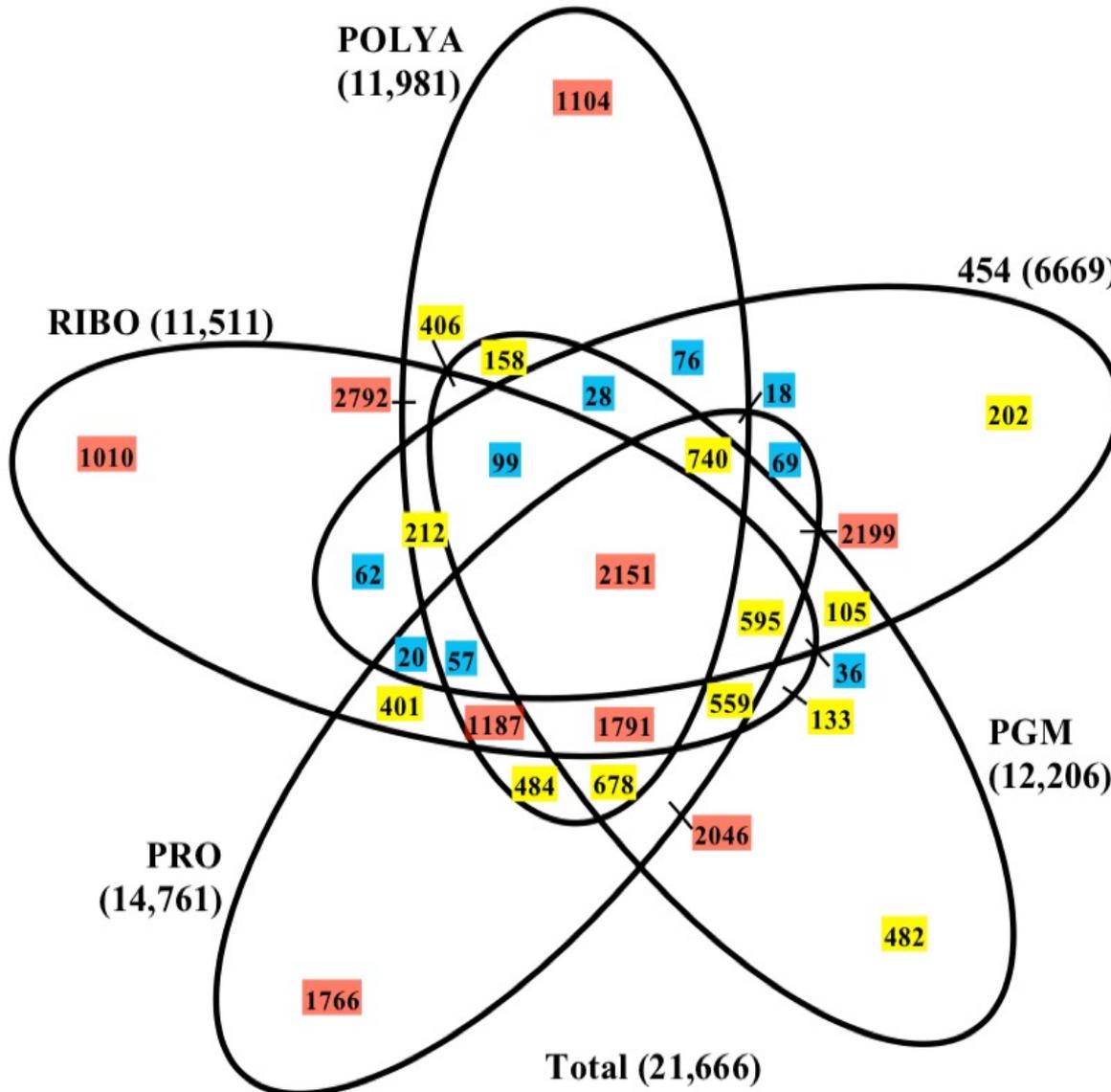
Shared sets of greater than 1000 genes are indicated in red,
100-999 yellow,
<100 blue.

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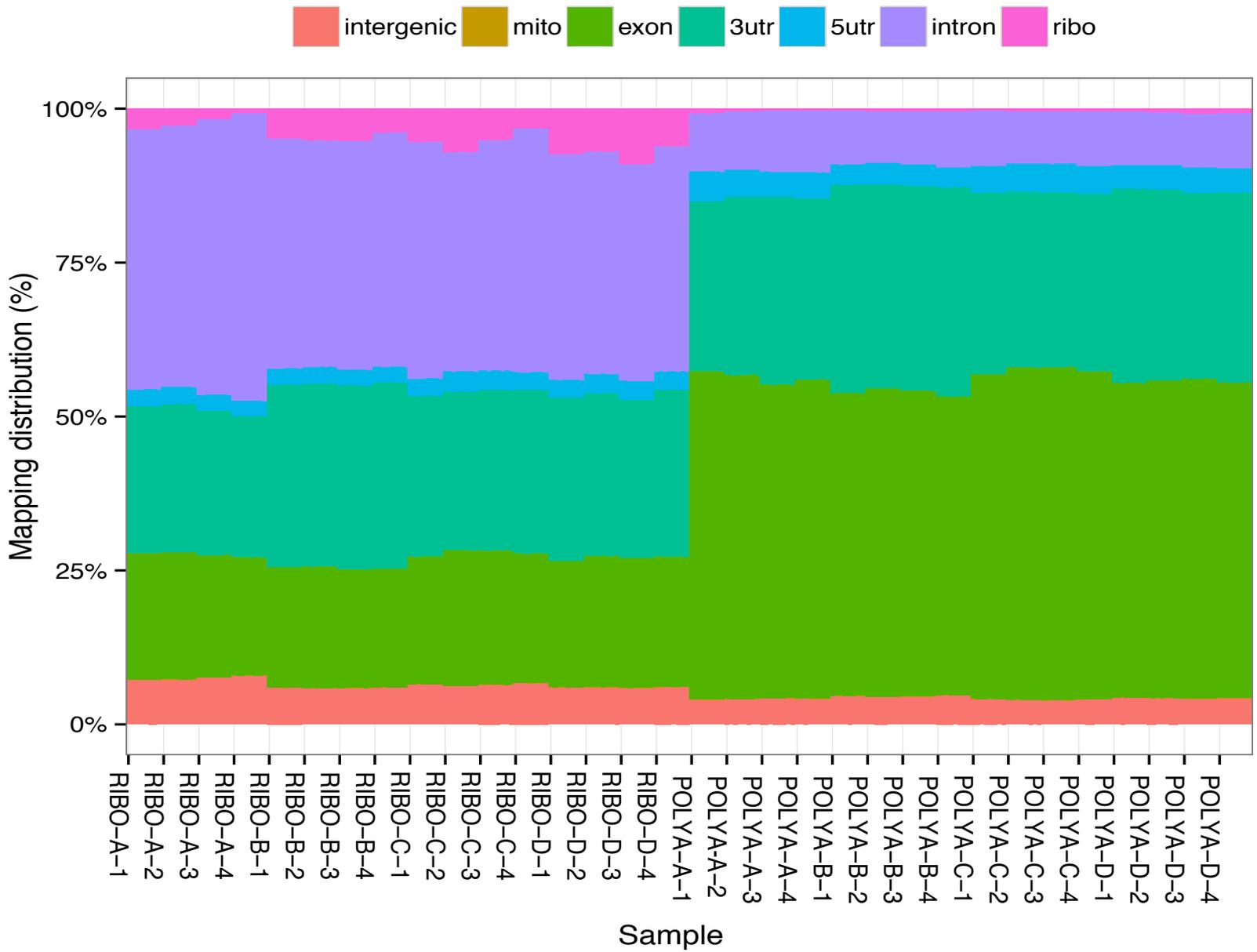
Supplemental Table 3 - Top 25 Genes with Highest Enrichment from Ribo-Depletion Preparation

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
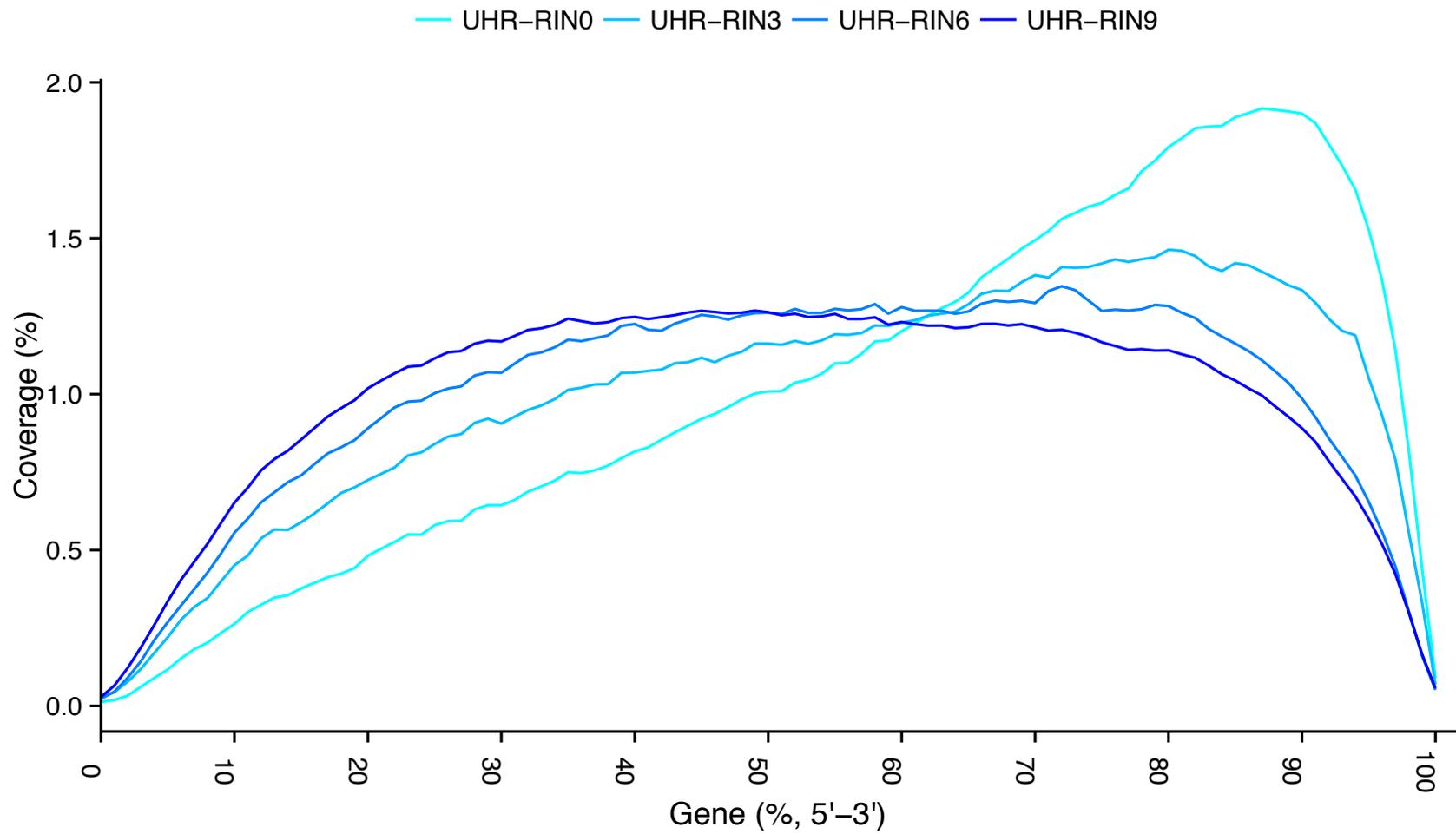
Entropy is usually a source of fear

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)

"**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!
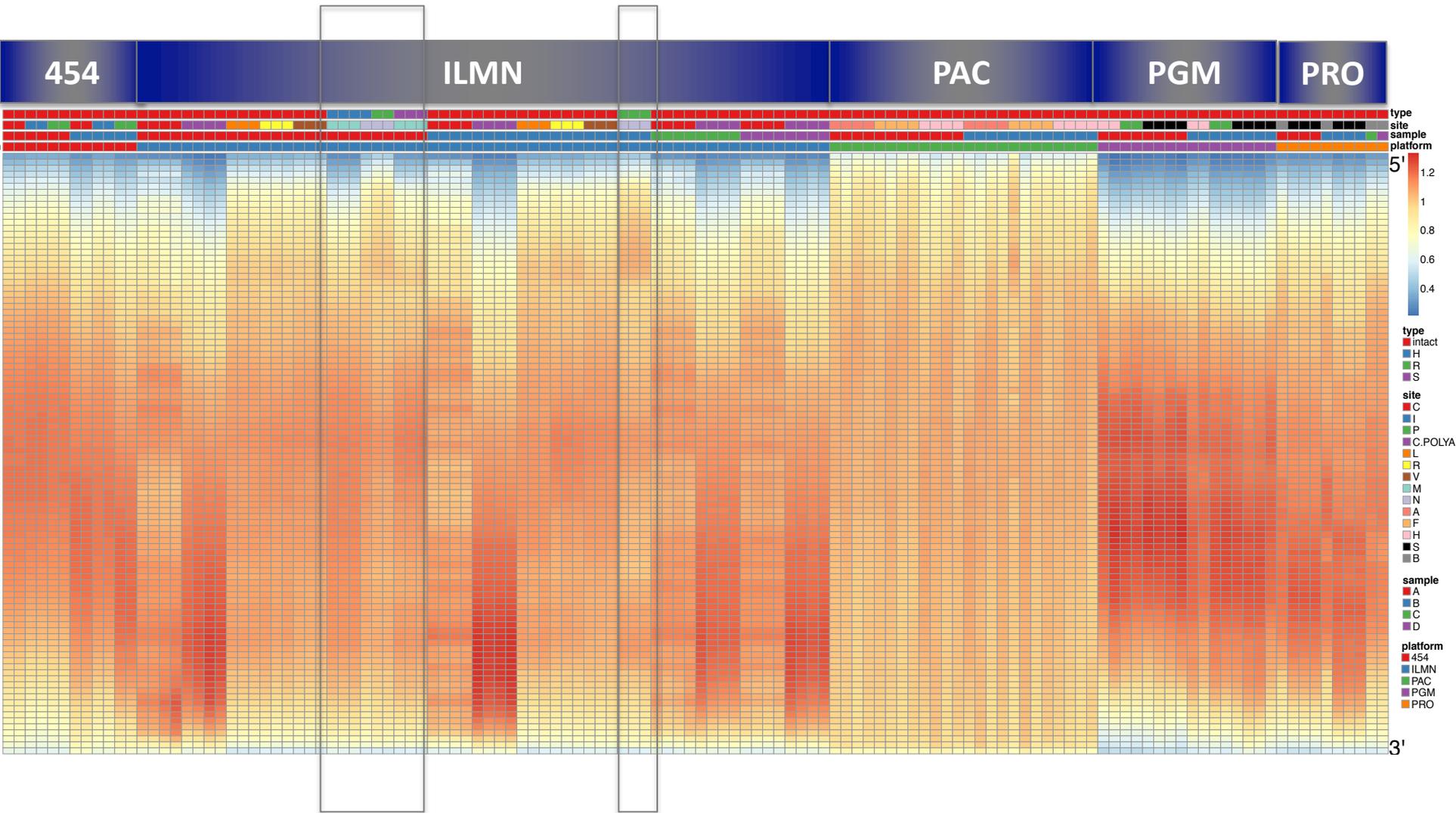
Degraded RNA highly correlates with intact RNA gene expression
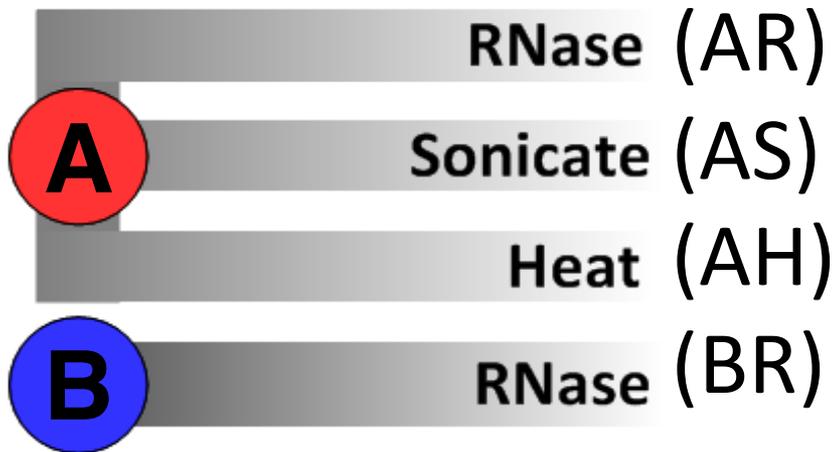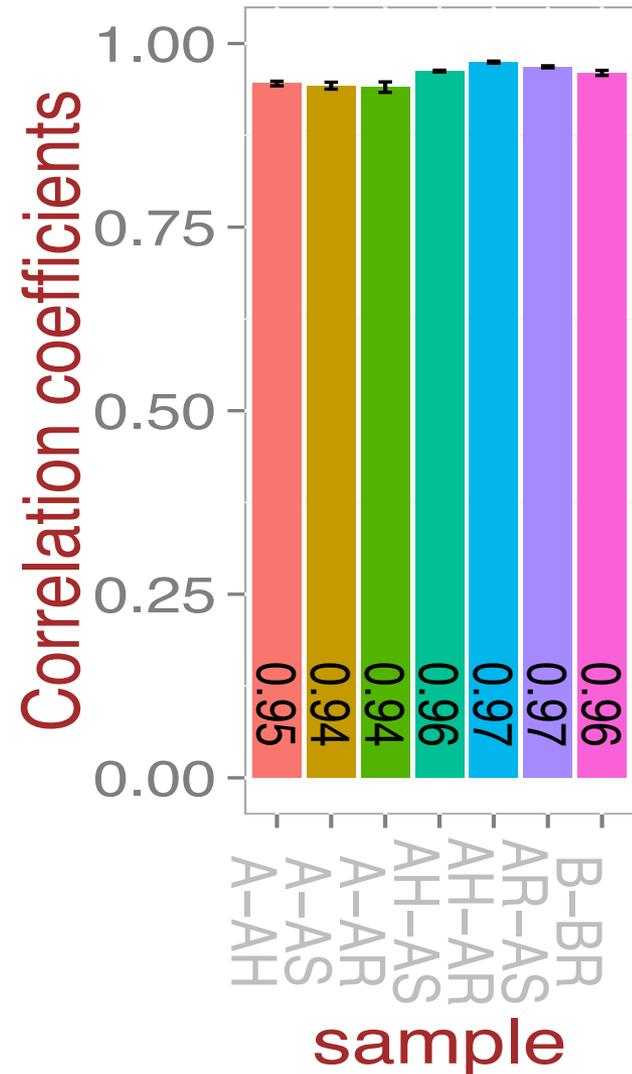
Ameliorating Inter-site Variation

SEQC
samples

Each site has
4 replicates

ILM1   ILM2   ILM3   ILM4   ILM5   ILM6

HiSeq 2000
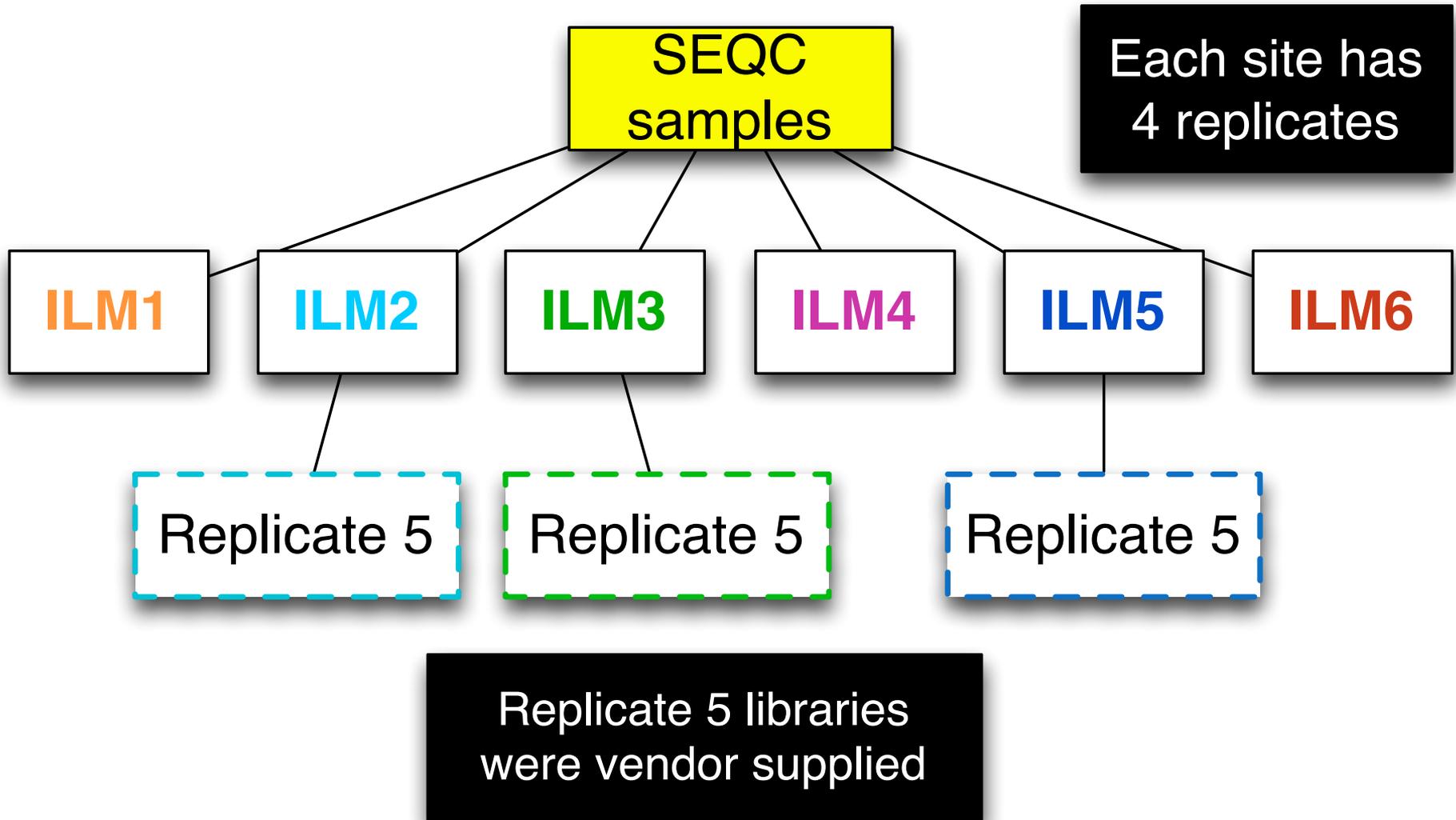
Differential expression calls – AvsB

significant @ p < 1%

Paweł Łabaj

Differential expression calls – CvsD

significant @ *p* < 1%

Paweł Łabaj

Differential expression calls – AvsA

significant @ p < 1%

Microarray AvsA   **RNA−Seq AvsA**
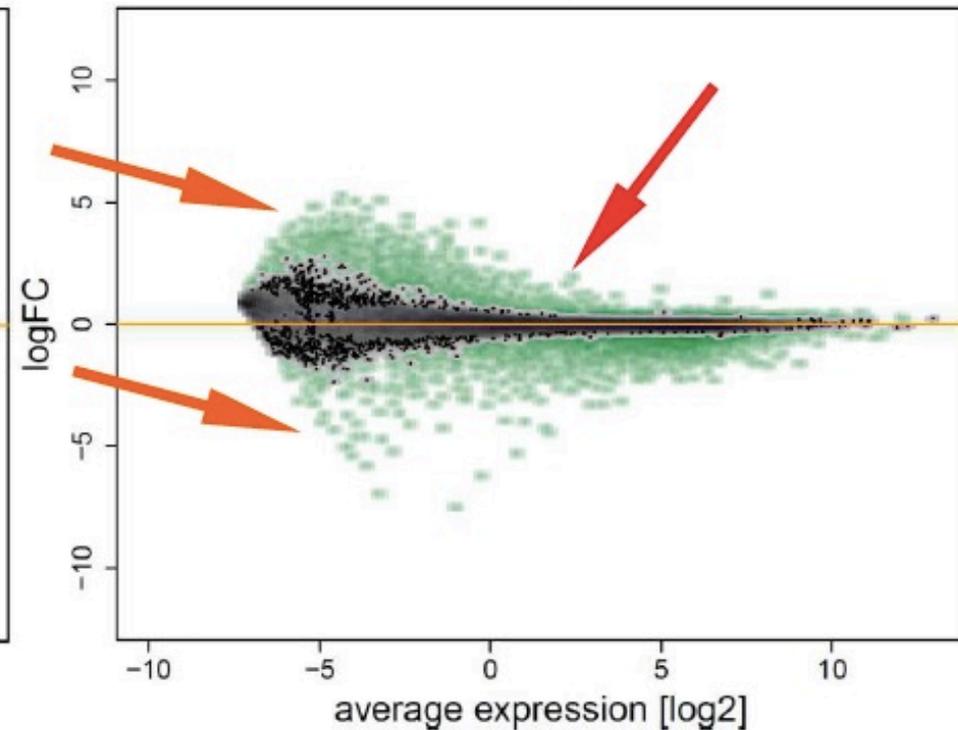
RNA-Seq and microarrays (MAQC-I): (Nature Biotech, 2006)

− 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
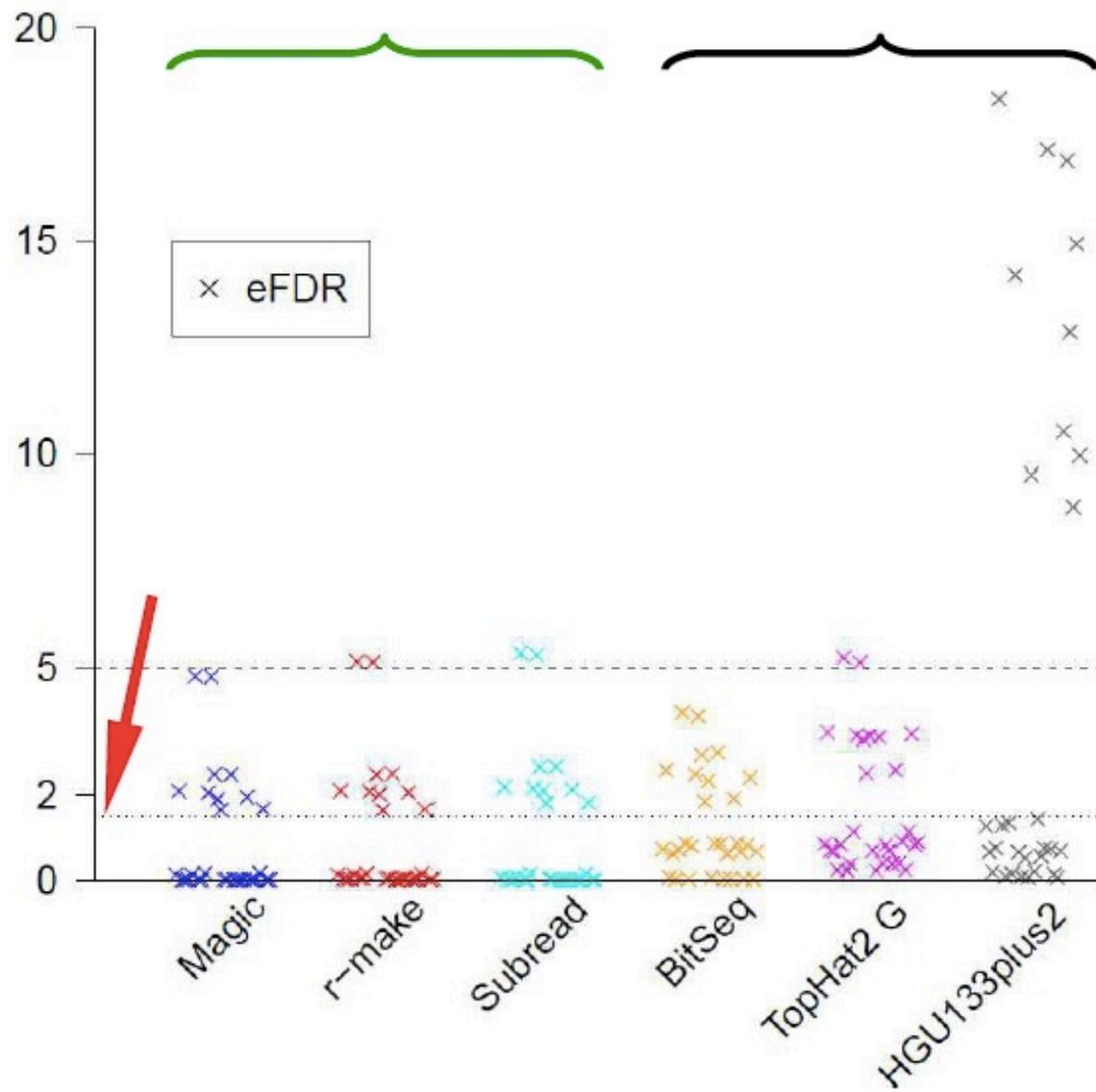
Systematic variation removal - False Positives

GC-content bias correction:

- EDASeq

- cqn

Factor analysis based on ERCCs:

- RUV2 (ERCC)

Latent variables:

- sva

- PEER

Paweł Łabaj
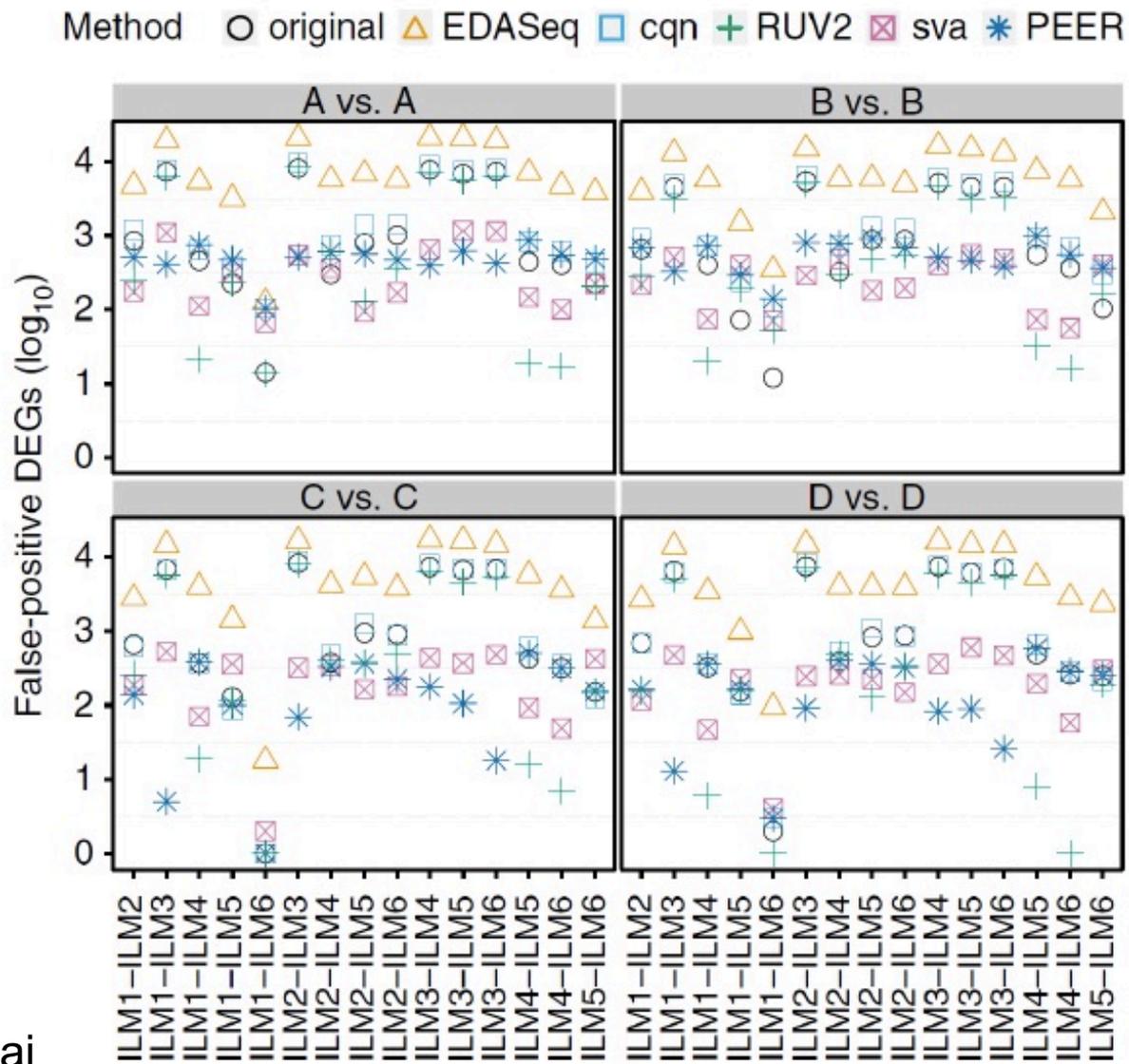
Identification of the underlying sources of variation – GC content

Outlier site: ILM 3

5th replicate → **not** affected

Paweł Łabaj

Identification of the underlying sources of variation – base error rate

C   ● ILM1 ● ILM2 ● ILM3 ● ILM4 ● ILM5 ● ILM6

○ 1 △ 2 □ 3 + 4 ⊠ 5

Outlier site:
ILM 3

5th replicate
→ *affected*

Paweł Łabaj
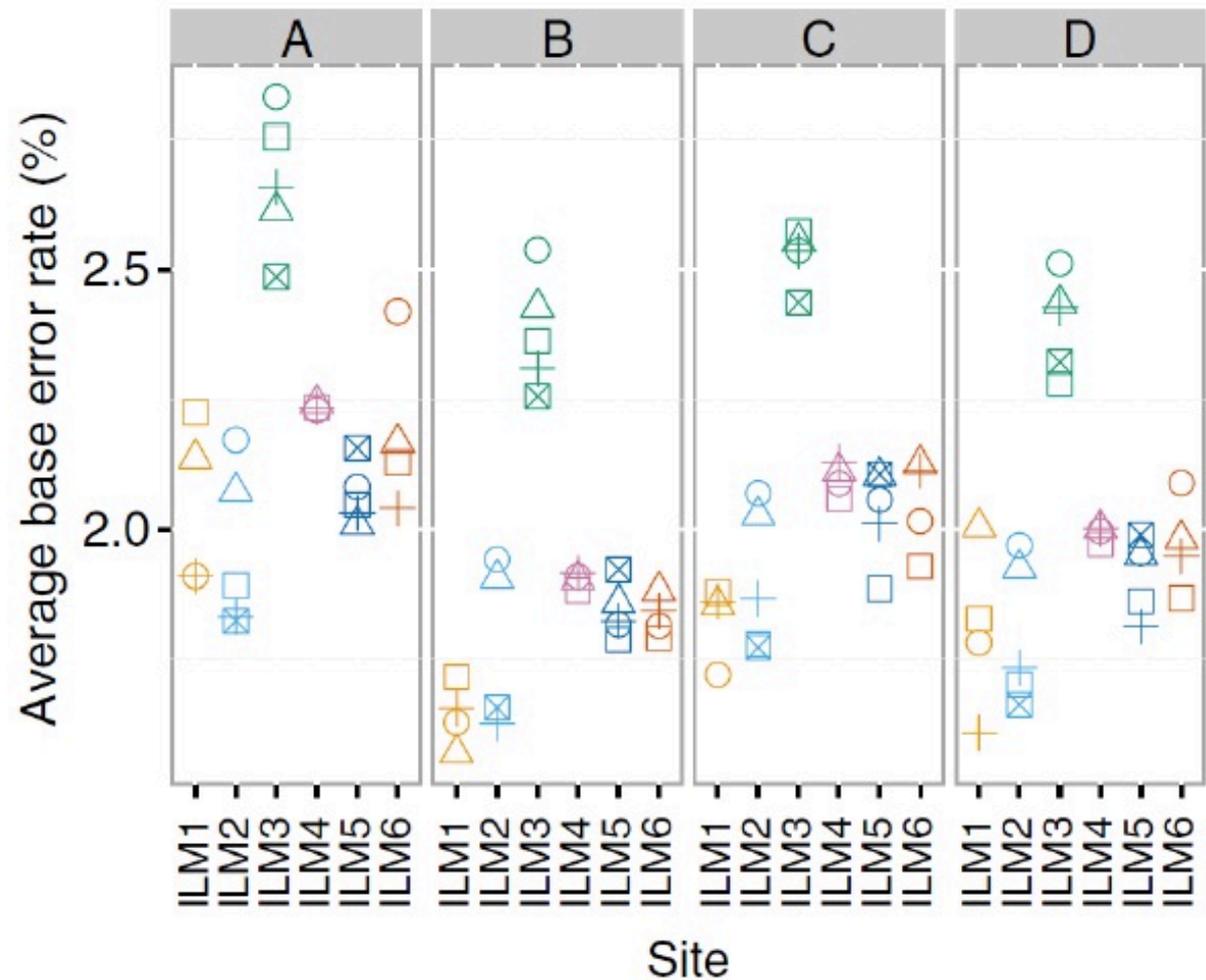
Identification of the underlying sources of variation – gene body
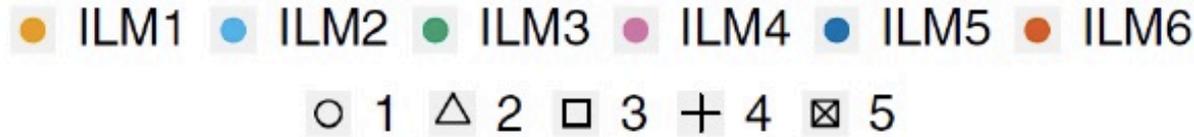
Outlier site:
ILM 3

5th replicate
→ **not** affected

Paweł Łabaj
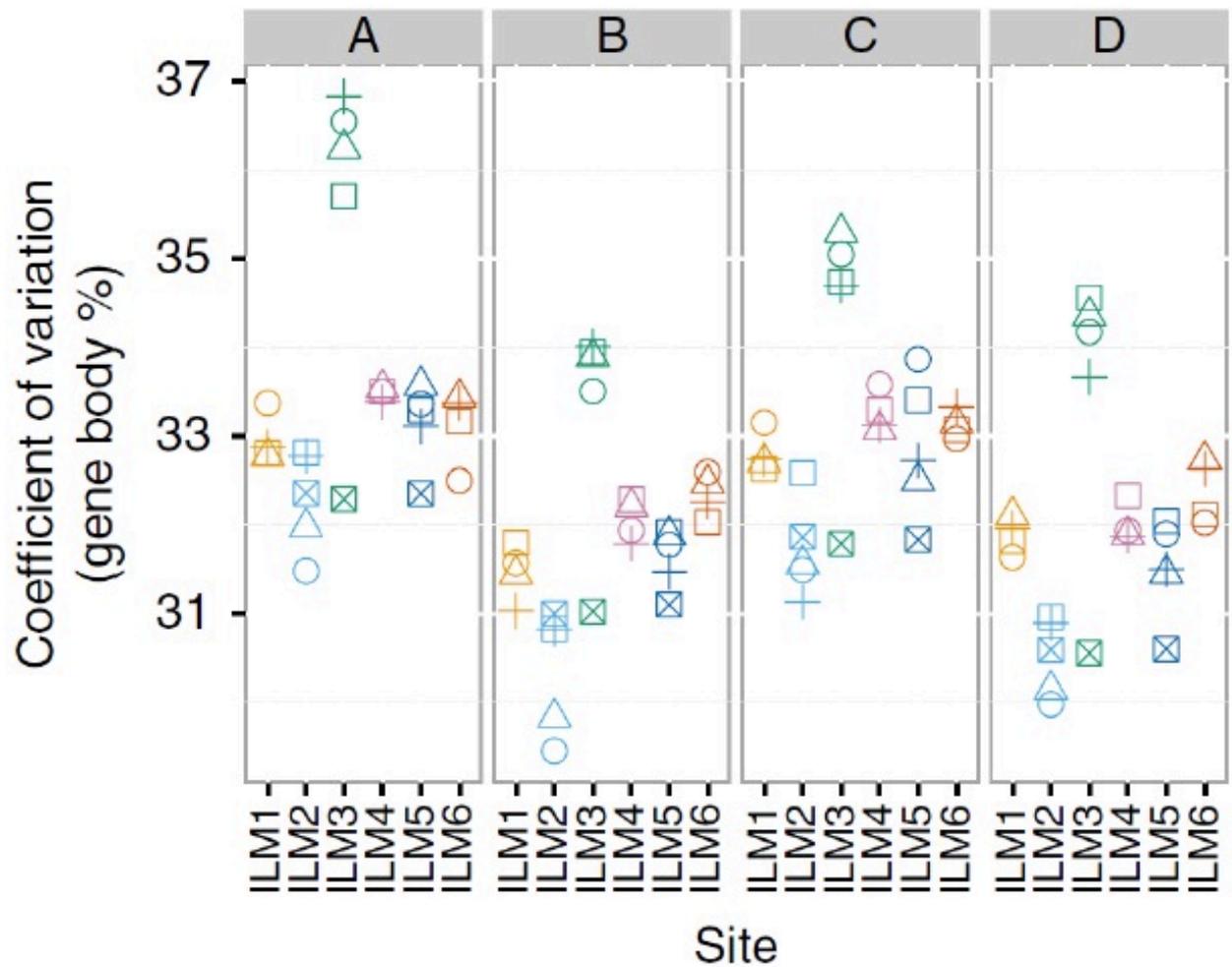
Identification of the underlying sources of variation – nucleotide composition
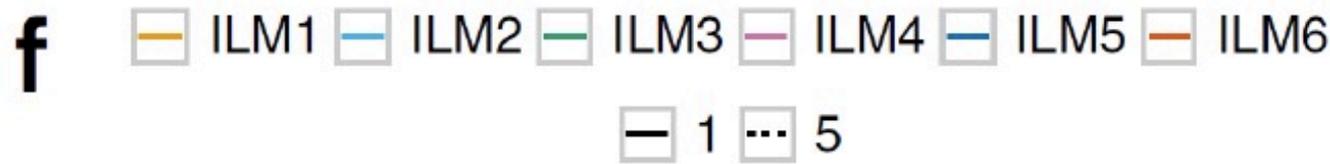
Outlier sites:
ILM 2 and ILM 3

5th replicate
→ **not** affected

Paweł Łabaj
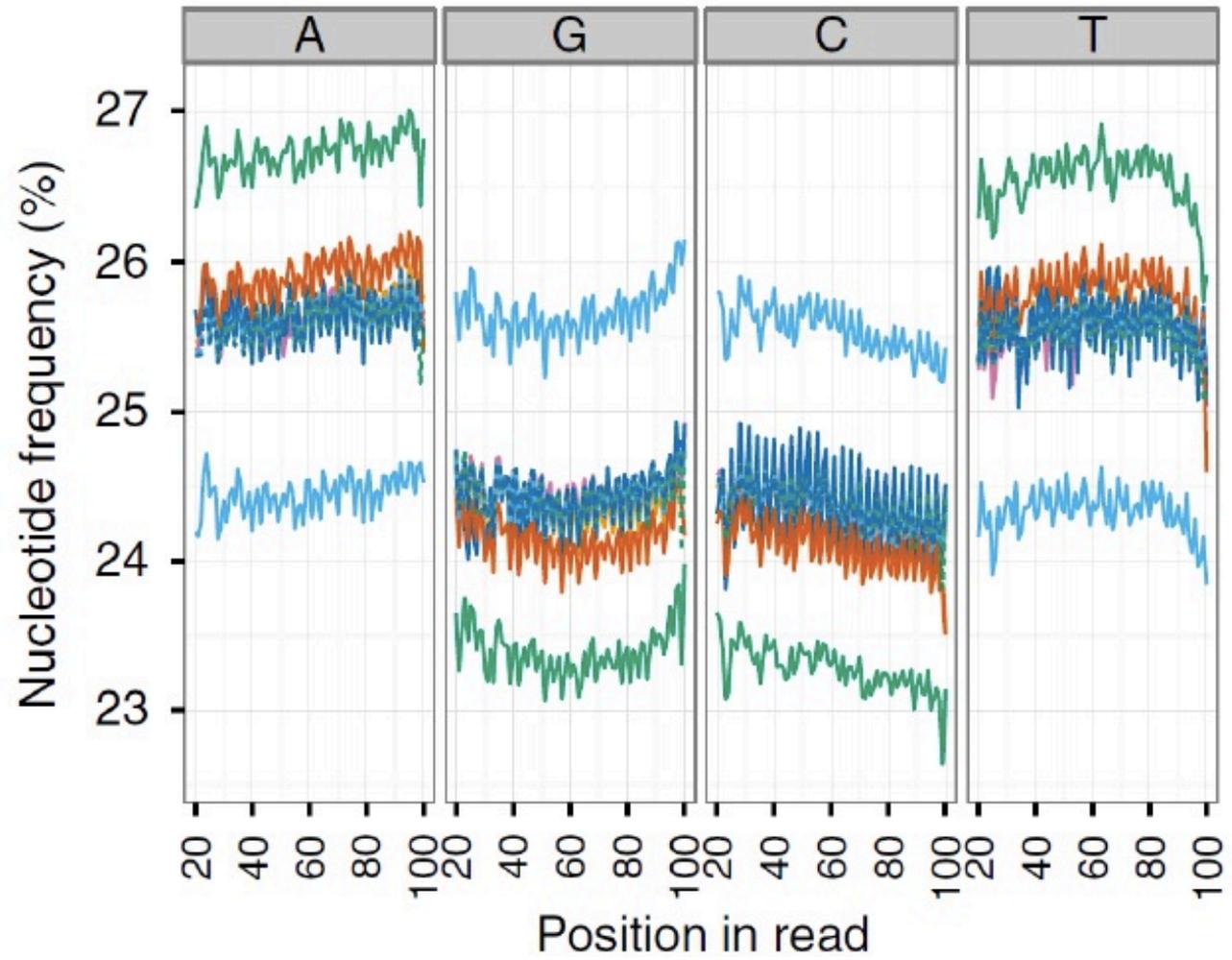
Determine sequencing variation sources

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

New genes still coming
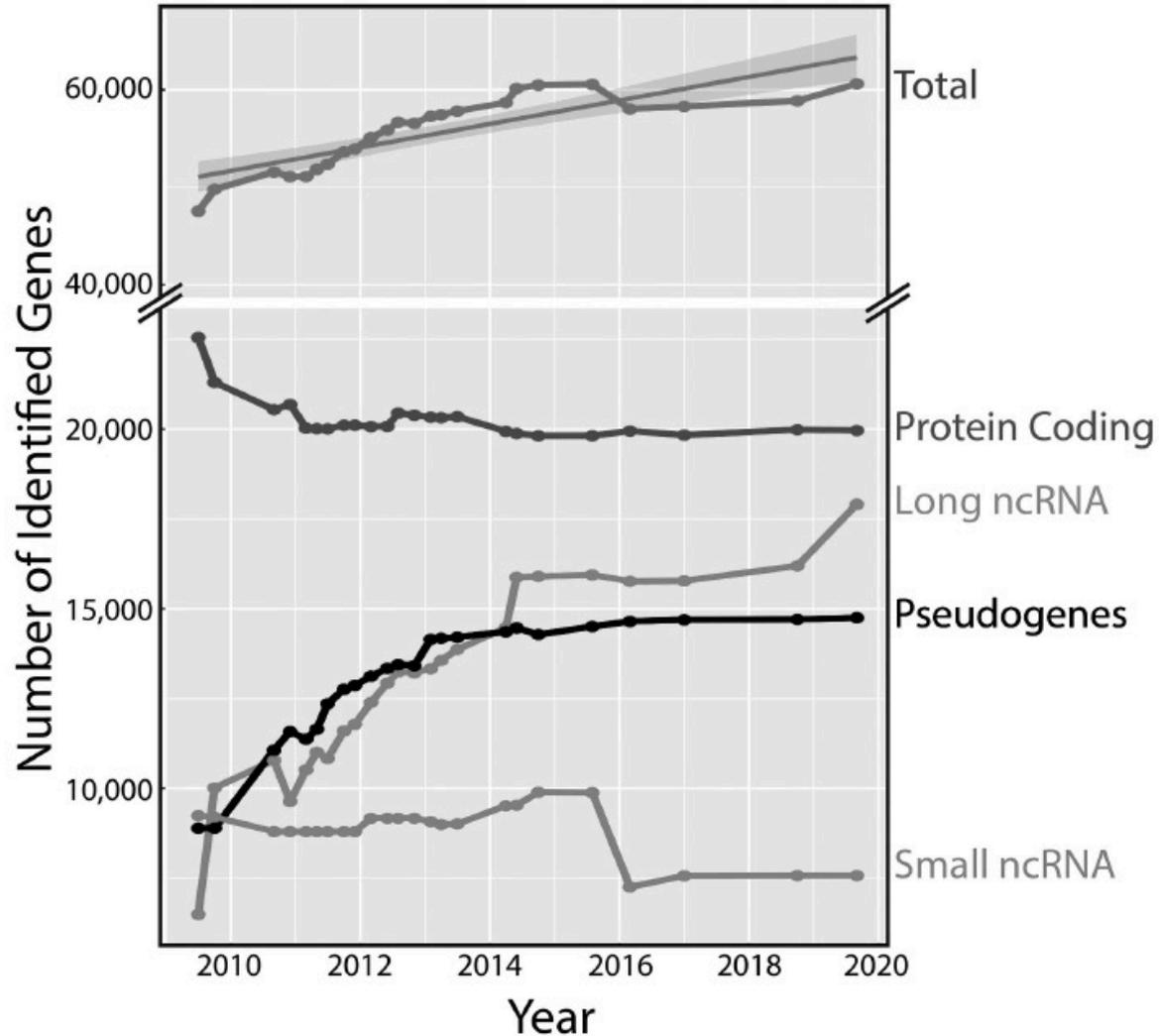
Annotations are a shifting sand, but so is the genome

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		

Version 15 (August 2012 freeze, GRCh37) -Ensembl 70

General stats

Total No of Genes	56680	Total No of Transcripts	195433
Protein-coding genes	20447	Protein-coding transcripts	82336
Long non-coding RNA genes	13249	- full length protein-coding:	57664
Small non-coding RNA genes	9173	- partial length protein-coding:	24672
Pseudogenes	13447	Nonsense mediated decay transcripts	12882
- processed pseudogenes:	9881	Long non-coding RNA loci transcripts	22531
- unprocessed pseudogenes:	2891		
- unitary pseudogenes:	160		
- polymorphic pseudogenes:	27		
- pseudogenes:	292		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	83177
- protein coding segments:	364	Genes that have more than one distinct translations	14807
- pseudogenes:	196		

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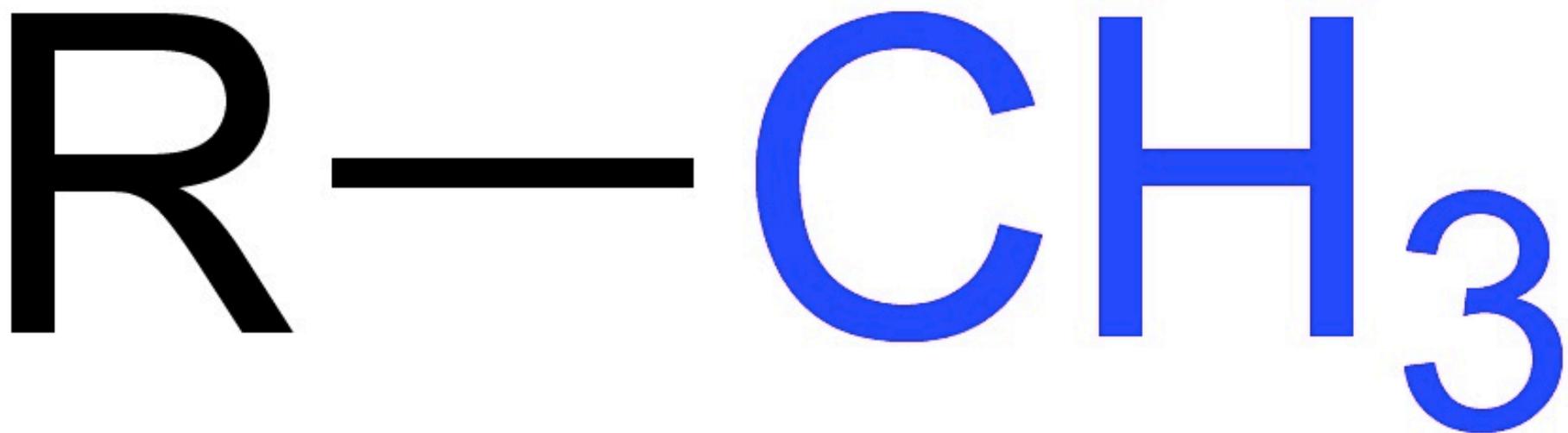
Slides from the GRC's presentation at ASHG 2012 are available on the new Workshops page.

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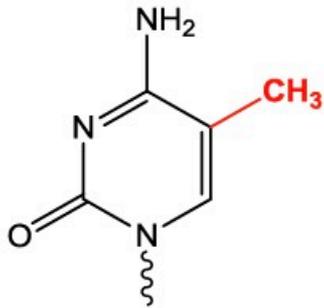
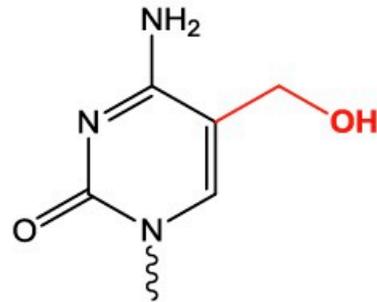See our blog for more information on why we think this is important.

(5)

Epitranscritpome

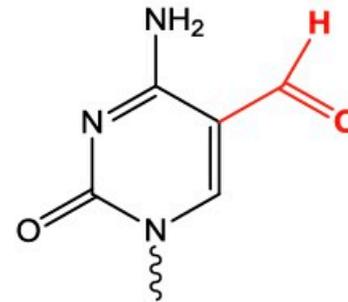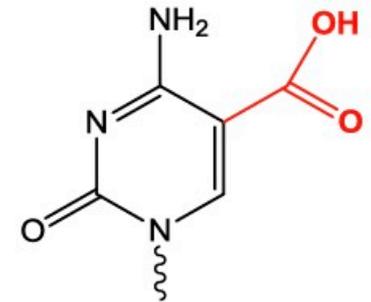$$R - CH_3$$

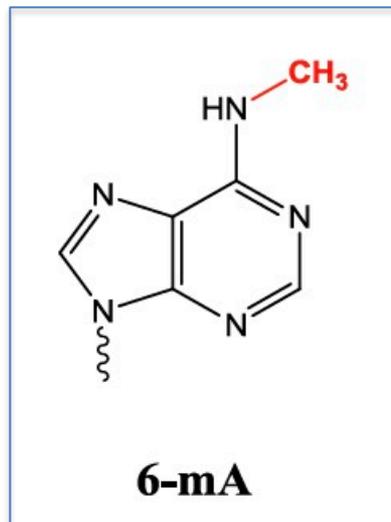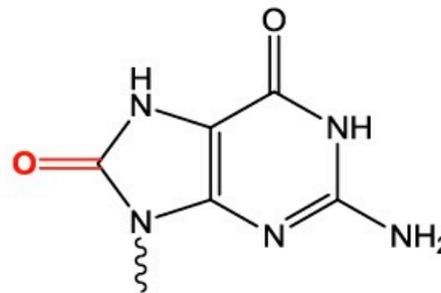# The four-base genome is just the beginning

5-mC

5-hmC

5-fC

5-caC
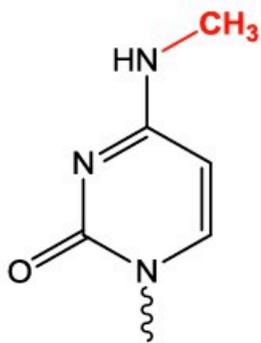
4-mC

6-mA
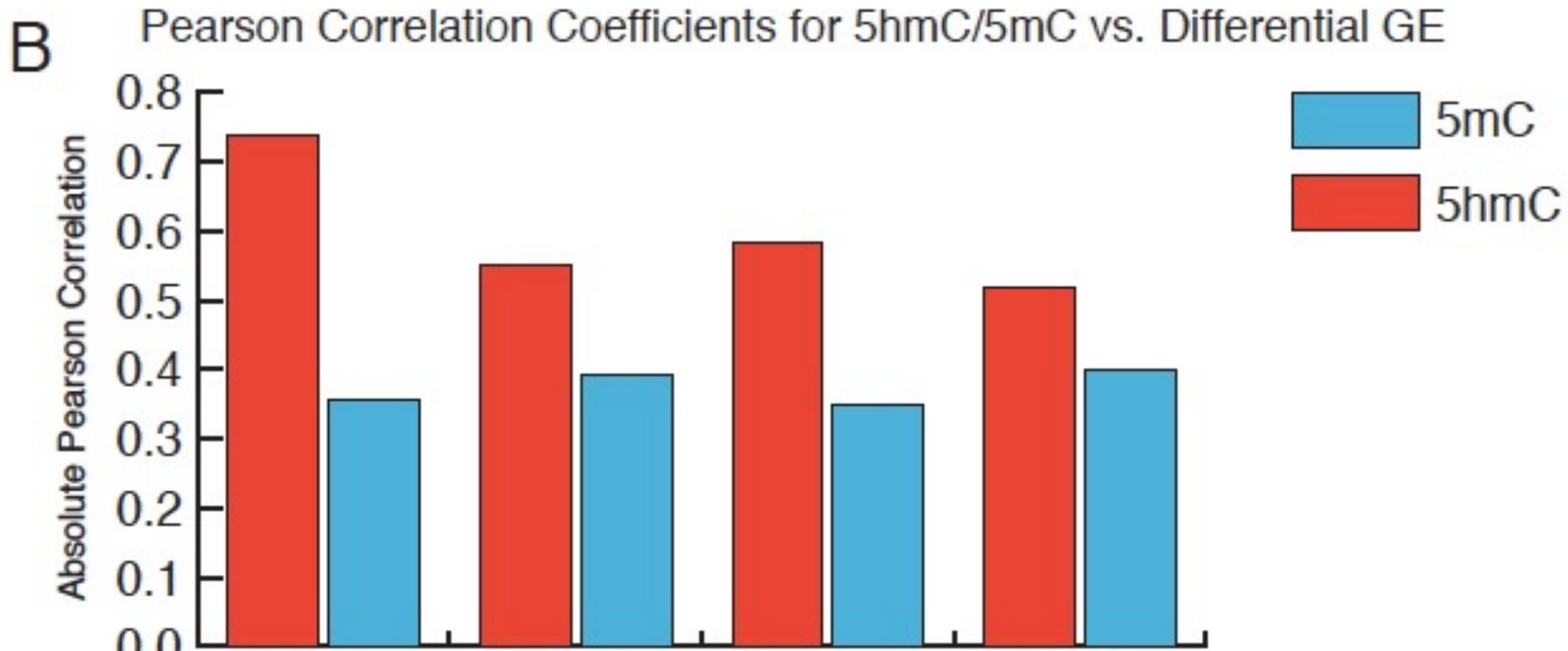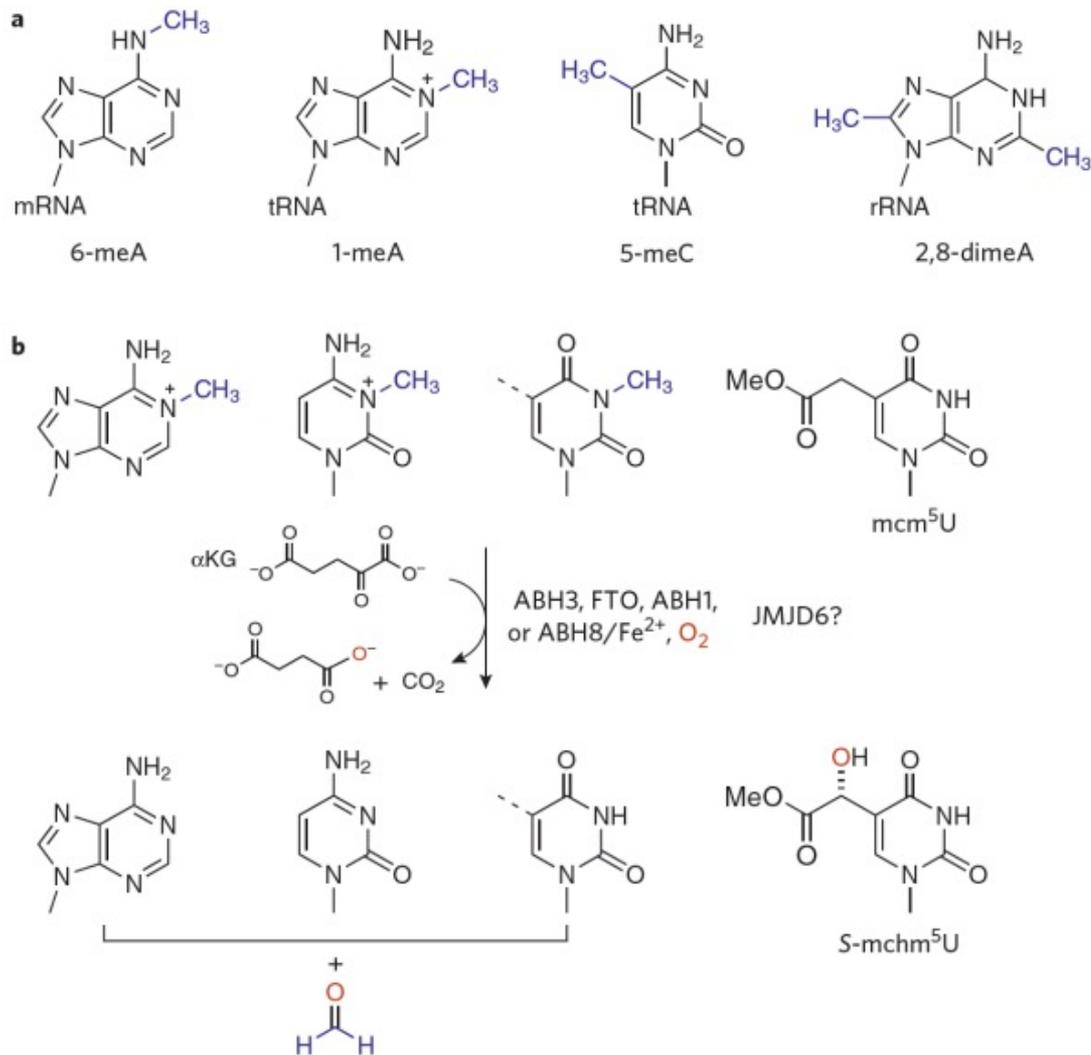
8-oxoG
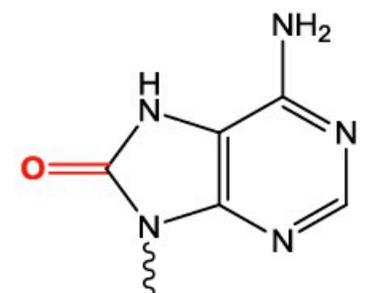
8-oxoA

# hmC predicts gene expresssion level

Pearson Correlation Coefficients for 5hmC/5mC vs. Differential GE

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 (m⁶A) in RNA, and is more prominent in brain & adults

Meyer et al., Cell, 2012
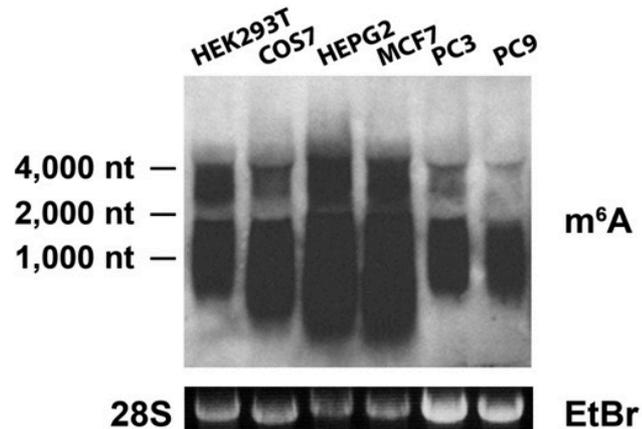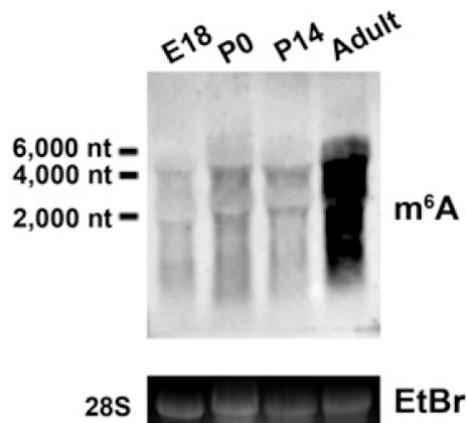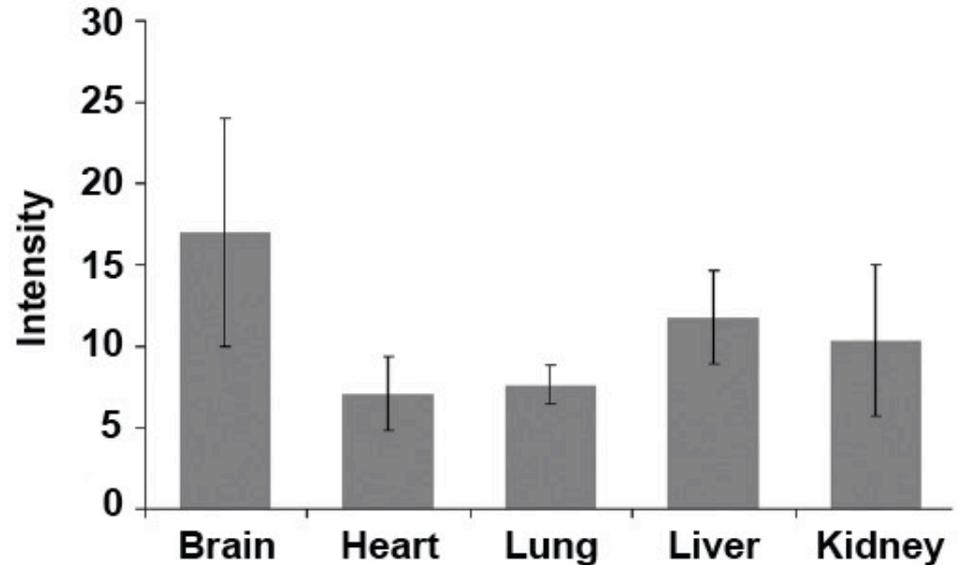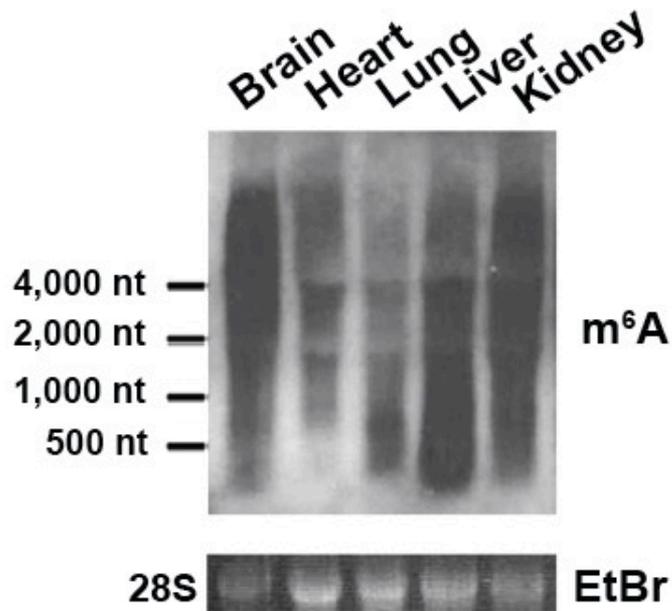
A new method: MeRIP-Seq

Meyer et al., Cell, 2012
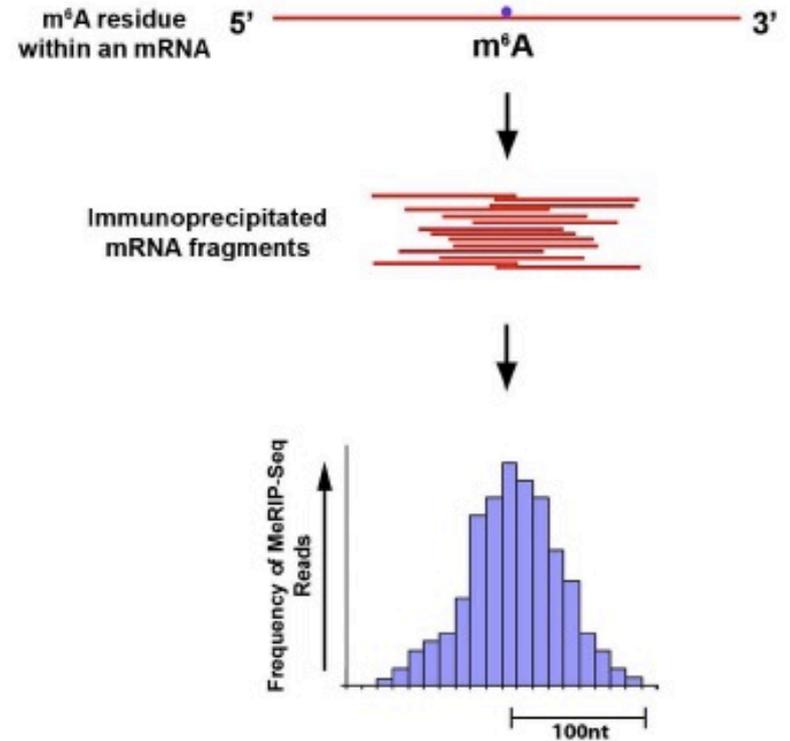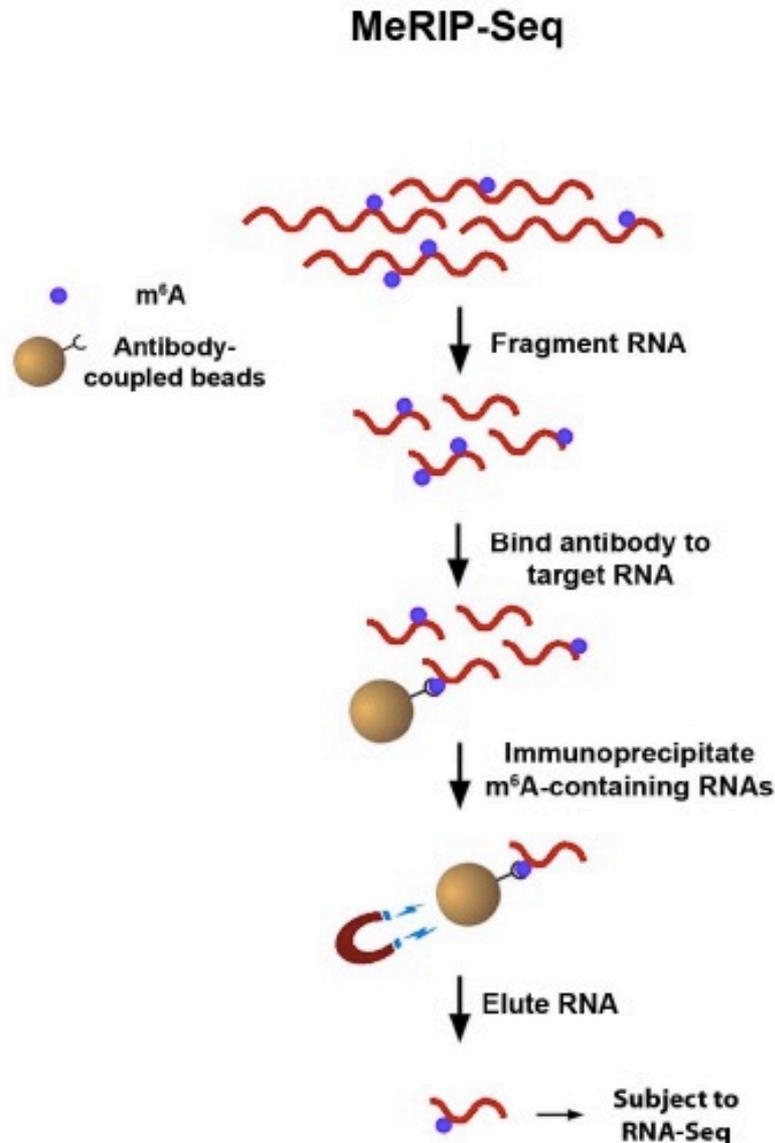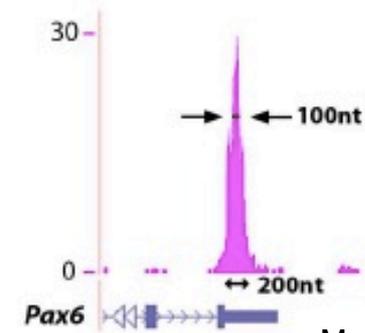
Peaks consistent across replicates

Meyer et al., Cell, 201
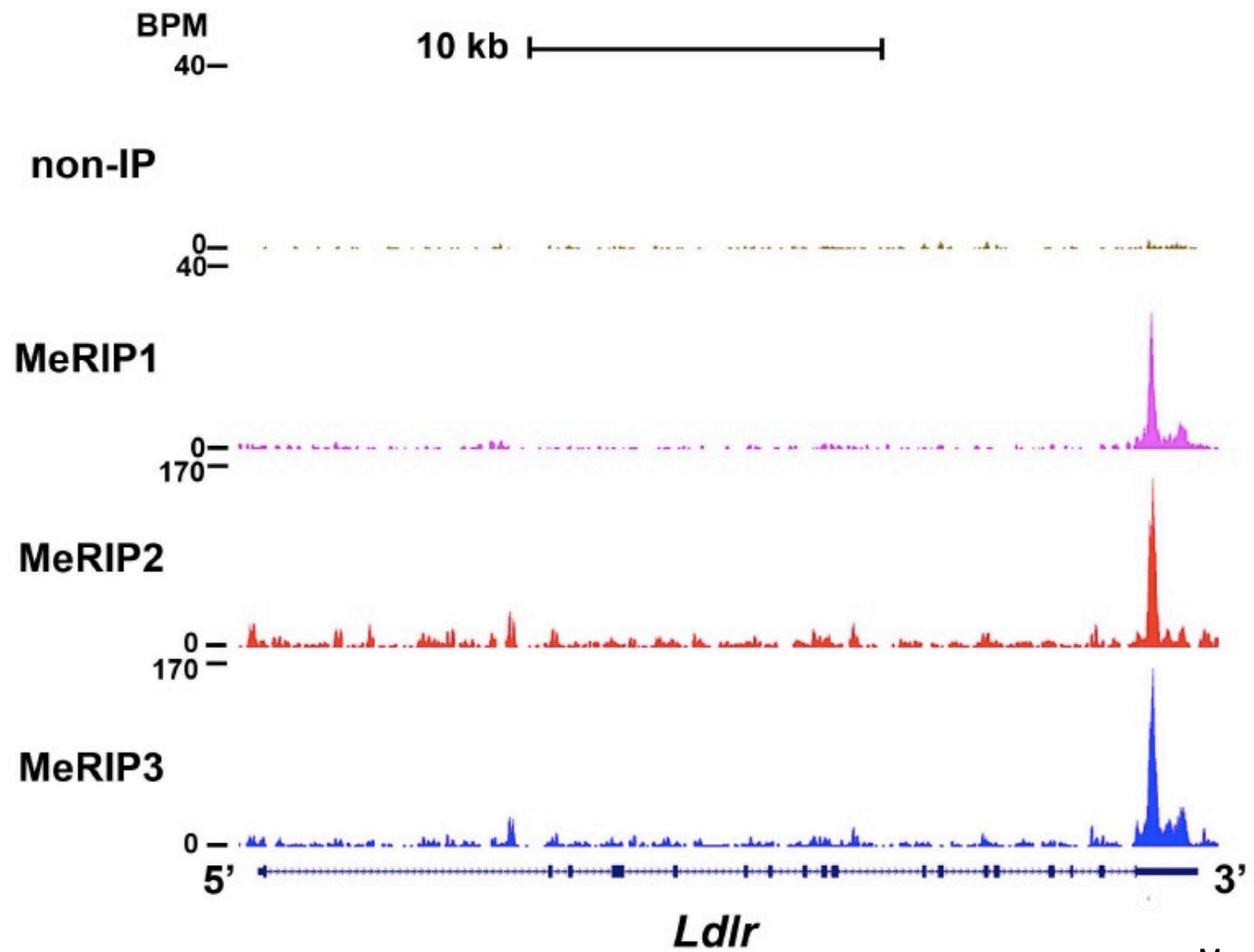
Conservations of signal and sites in >10,000 orthologous genes
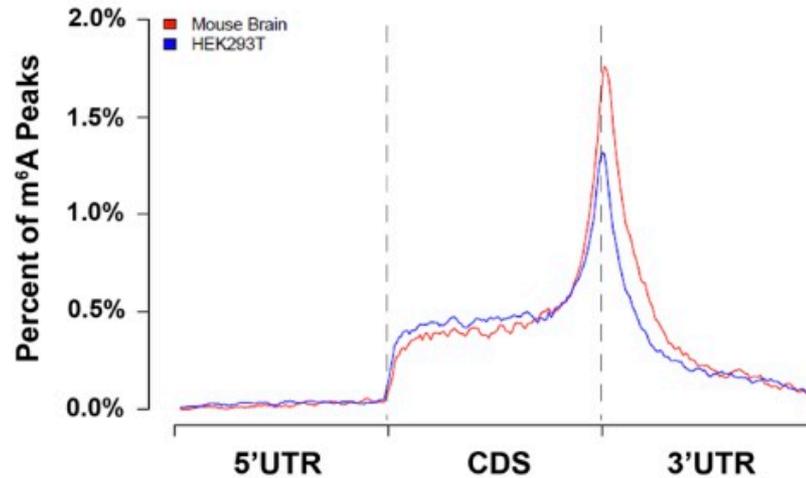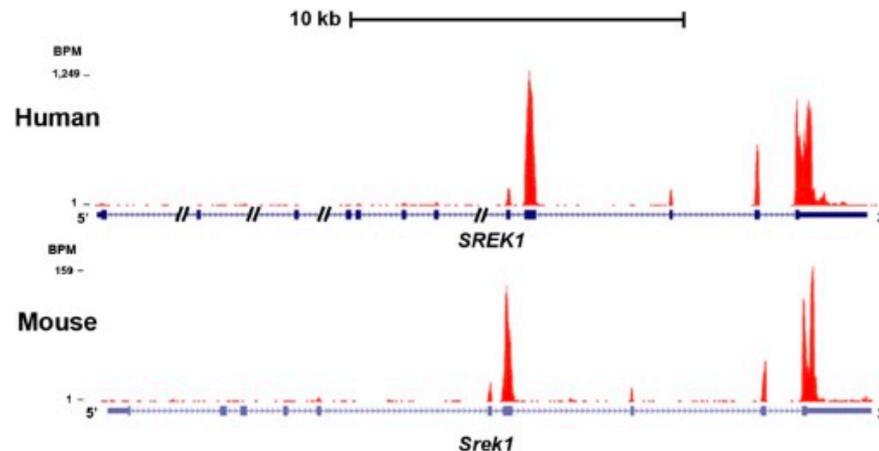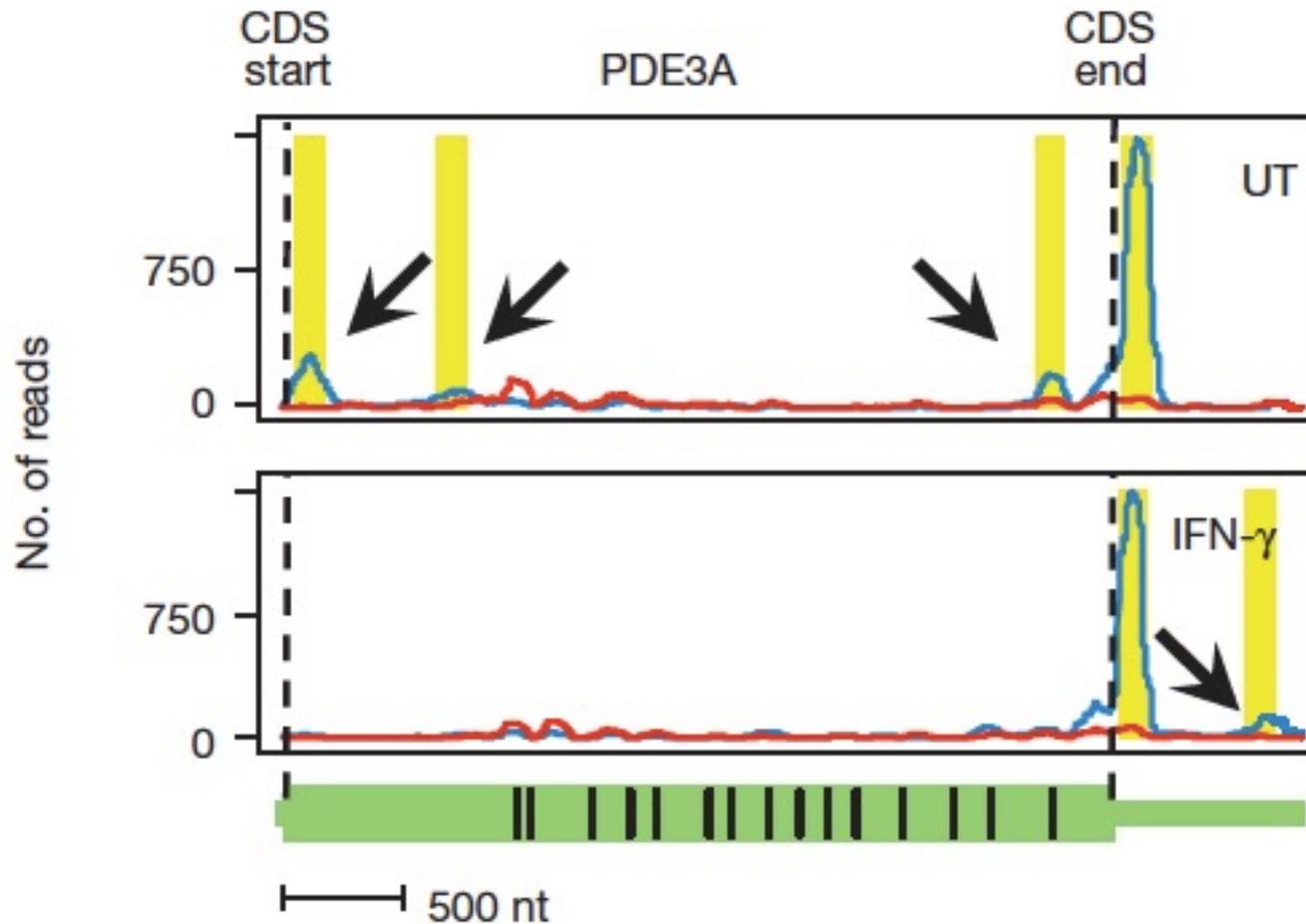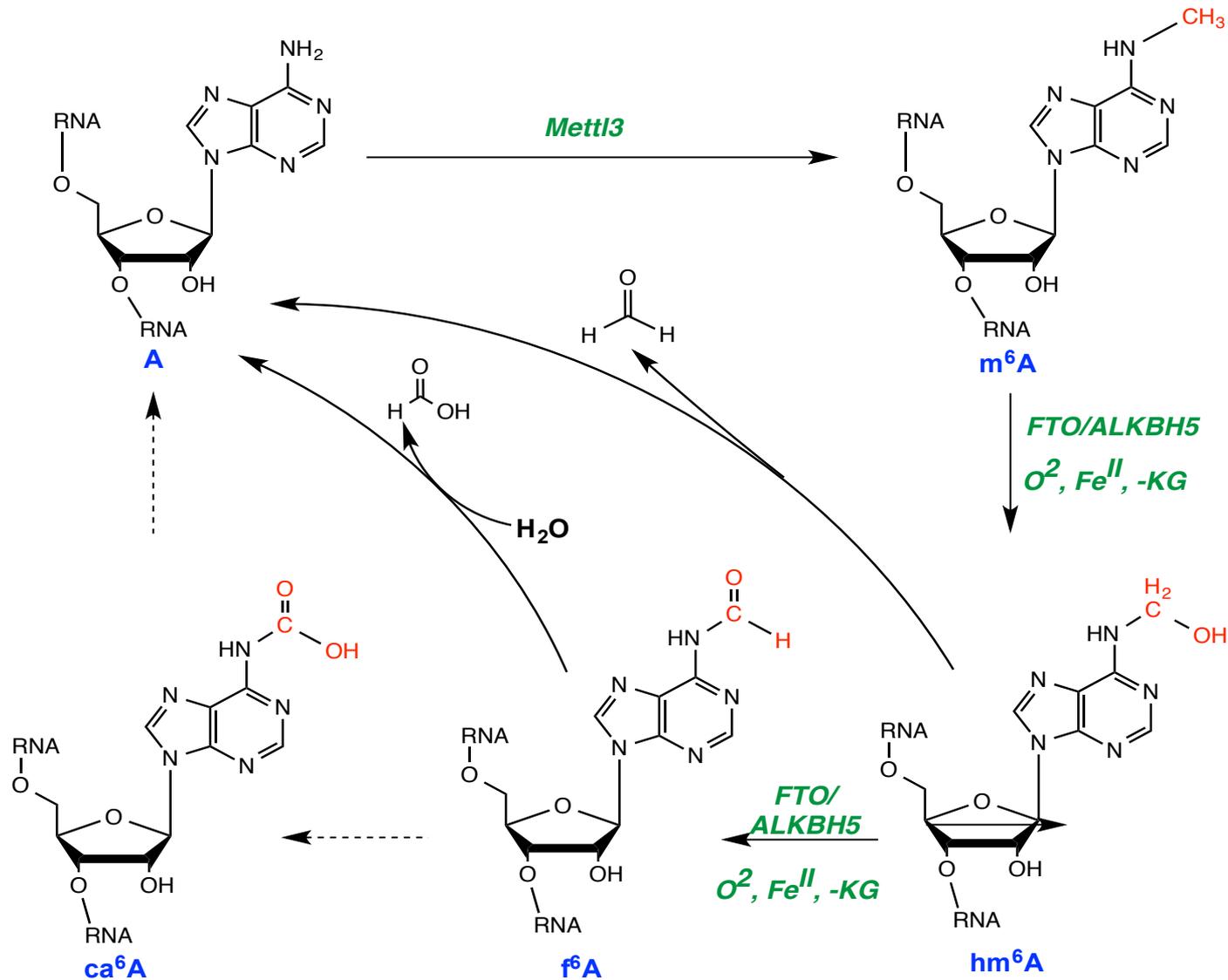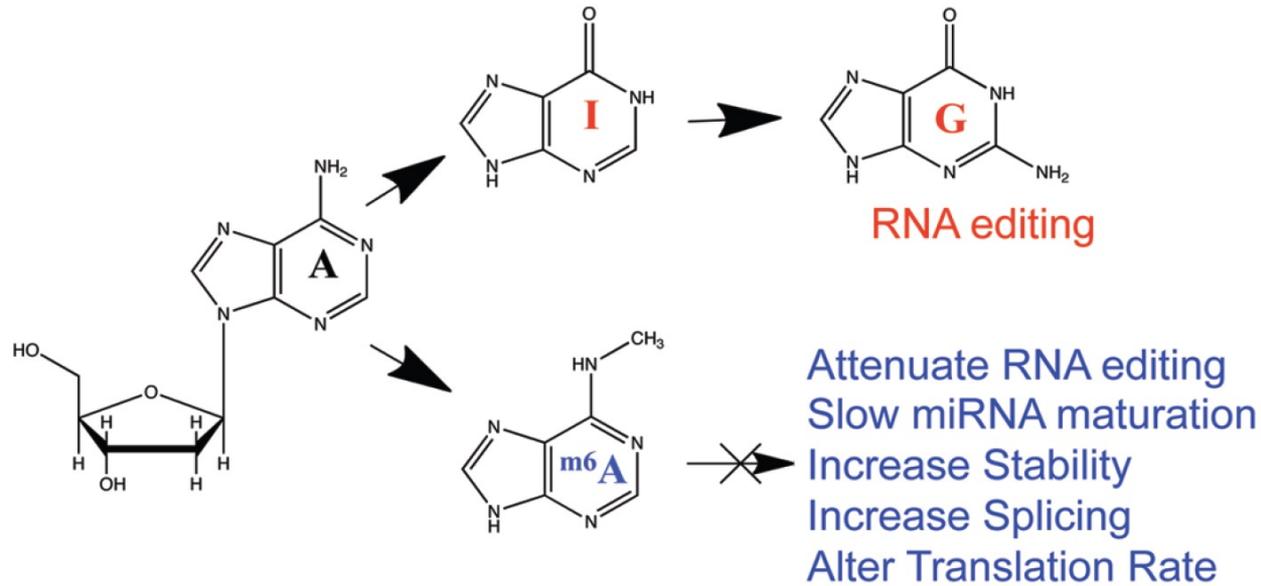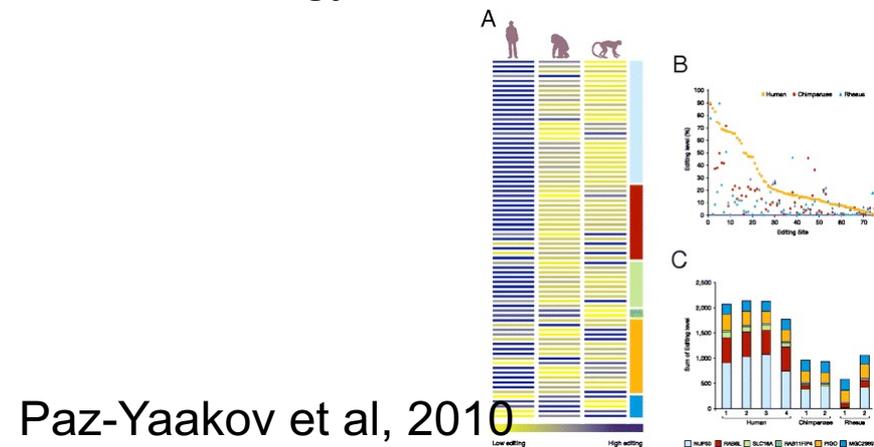
m⁶A levels may also change splicing patterns in genes

RNA modifications give a new layer of cellular regulation

Many putative roles for m⁶A 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, N6-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

| Abbreviation	Chemical name
m^{1}acp^{3}Ψ	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}Ψ	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
Ψm	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}Ψ	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
τm^{5}s^{2}U	5-taurinomethyl-2-thiouridine
τm^{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^{2,2}G	N^2,7-dimethylguanosine
m^{2,2}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
Ψ	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

Table 1 - List of Base Modifications Covered by Claims

Molecular Cell

Article

Cell PRESS

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 m⁶A defects perturb germline development
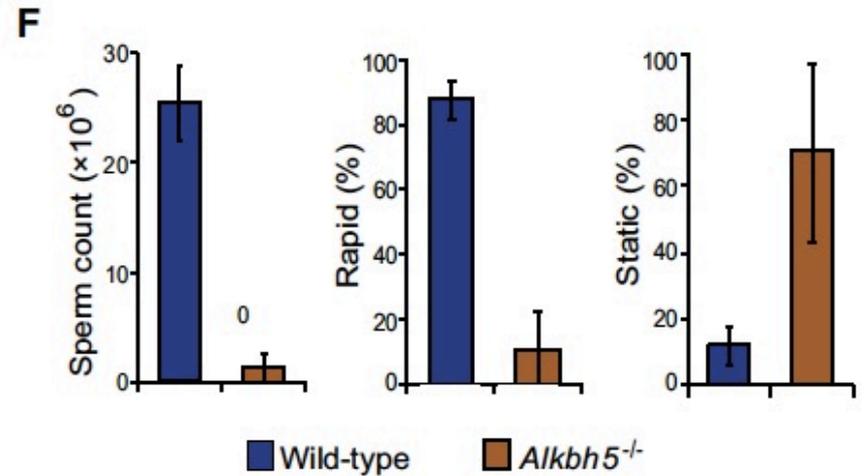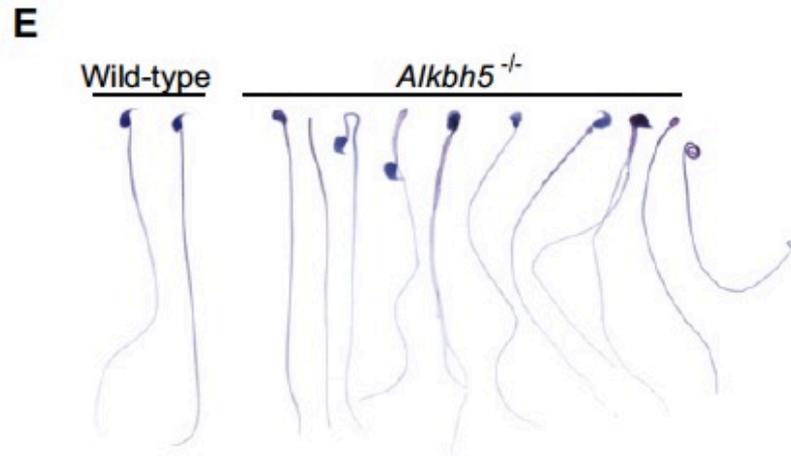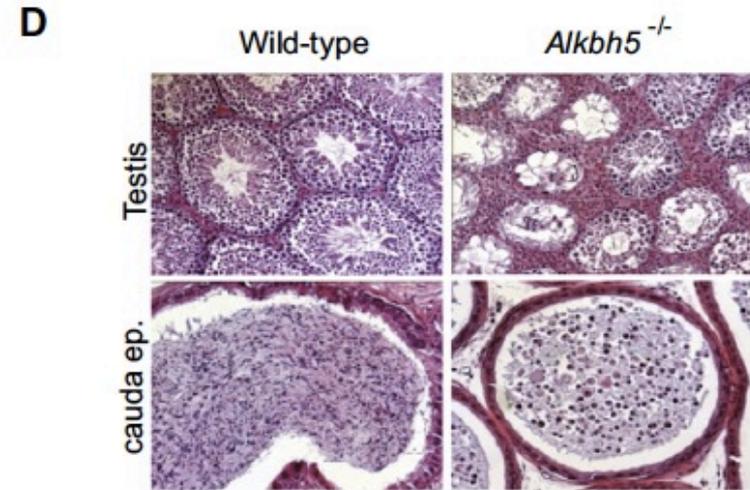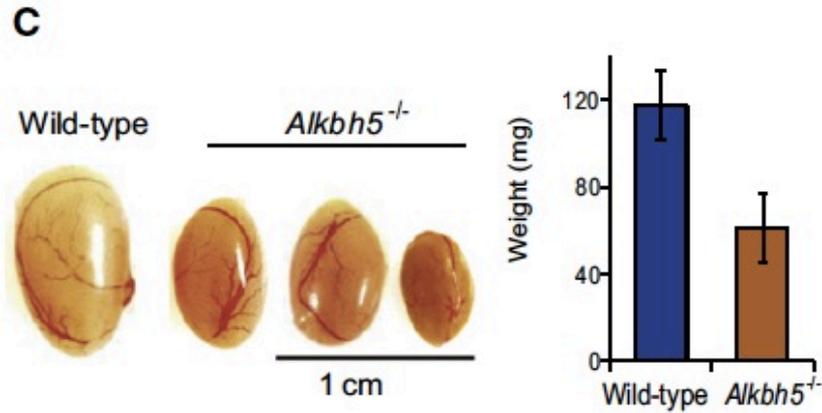
Dysregulated m⁶A affects many epigenetic modifiers
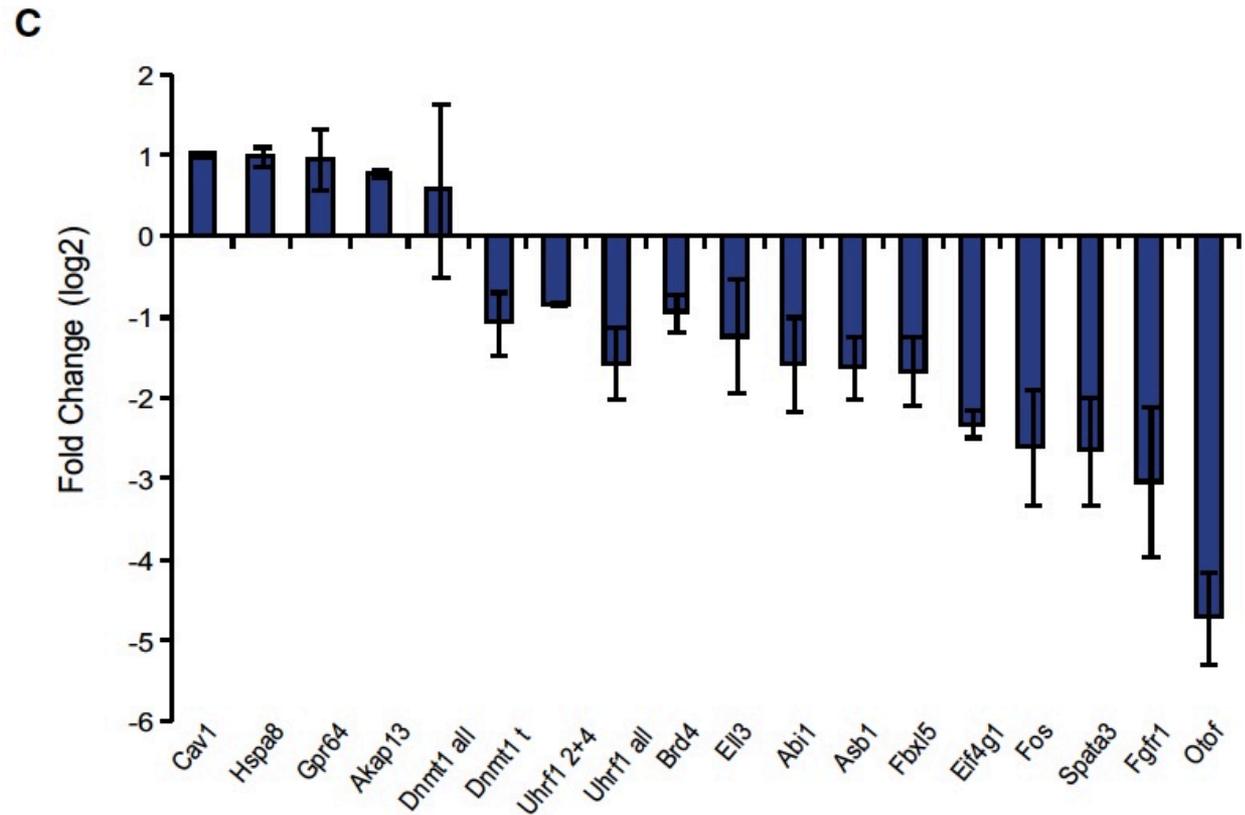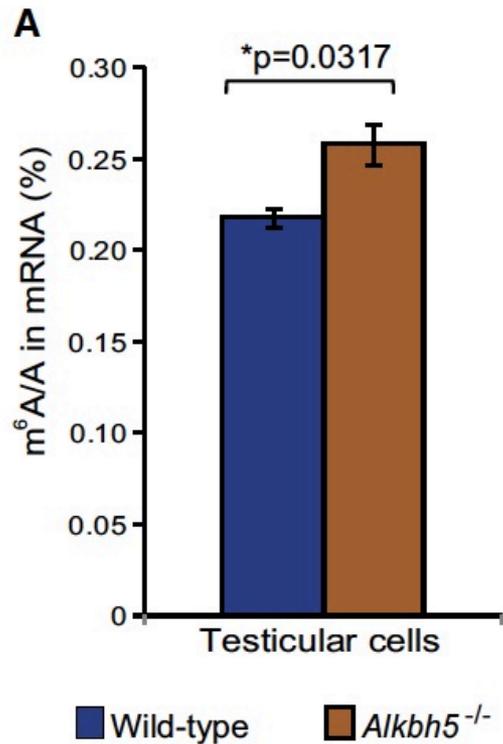
Information also pass between generations in RNA
Evidence of a Trans-generational Anti-viral RNAi response

Rechavi et al, 2011

from Saletore *et al.*, Genome Biology, 2012
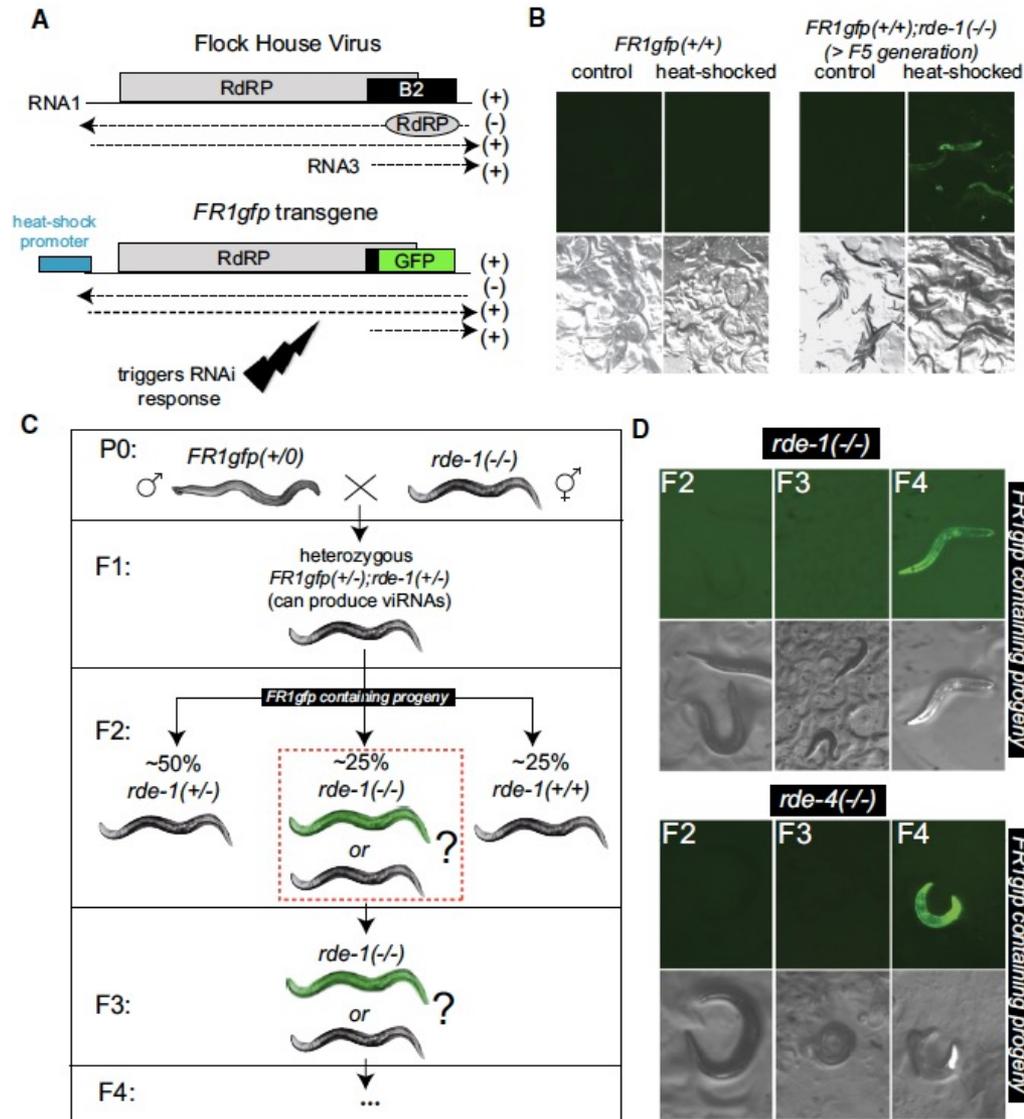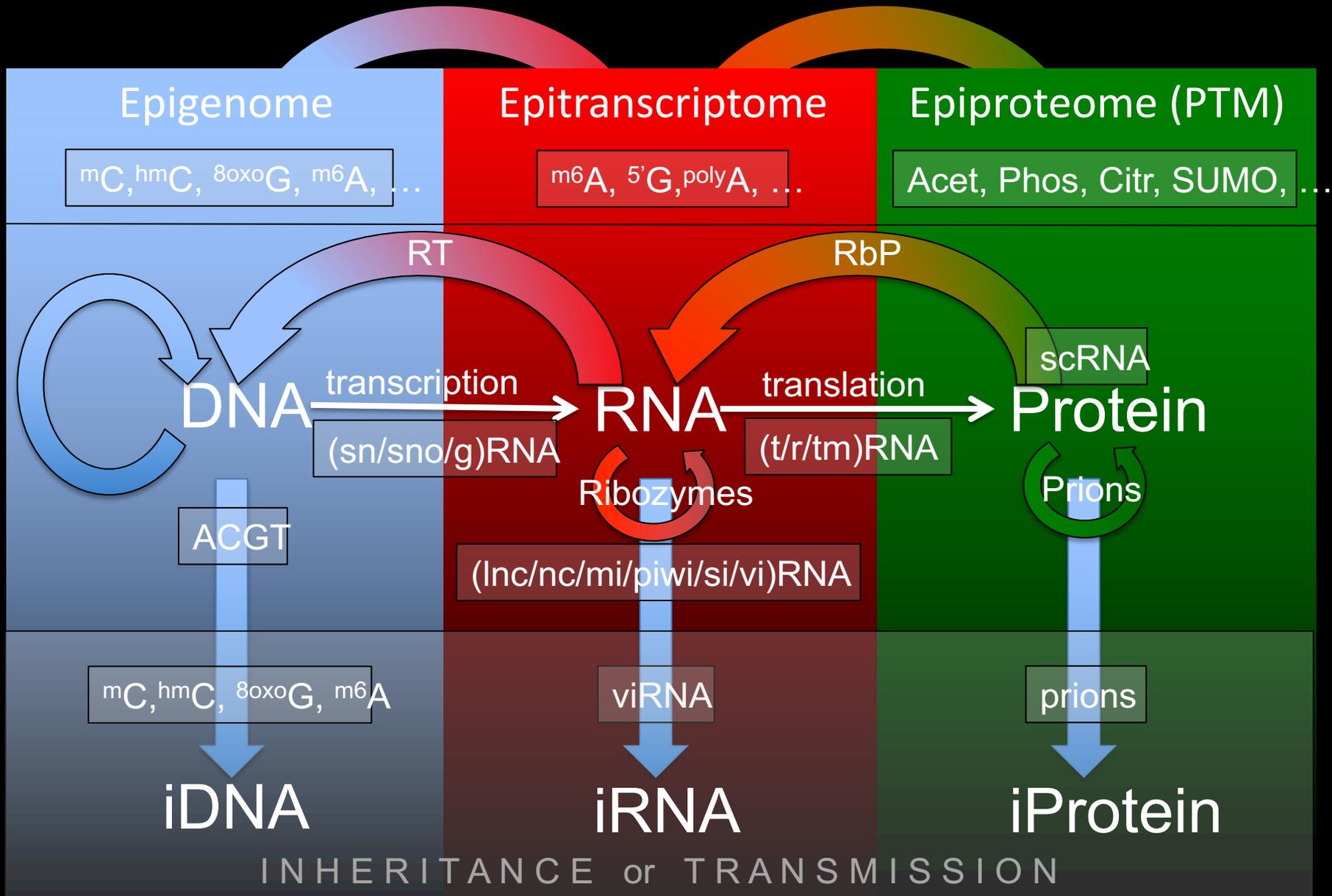
Used to be very hard to look at individual cells

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

CrossMark
click for updates

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 ⌃

This Issue

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

◄ PREV ARTICLE   NEXT ARTICLE ►

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

| Enrich | Load and Capture | Wash, Stain and Image | Lyse, RT, and Amplify | Prepare Library | Sequence | Analyze |

C₁™ Single-Cell Auto Prep System

Any Illumina® System

Drop-Seq

http://mccarrolllab.com/dropseq/

http://www.cell.com/abstract/S0092-8674%2815%2900549-8
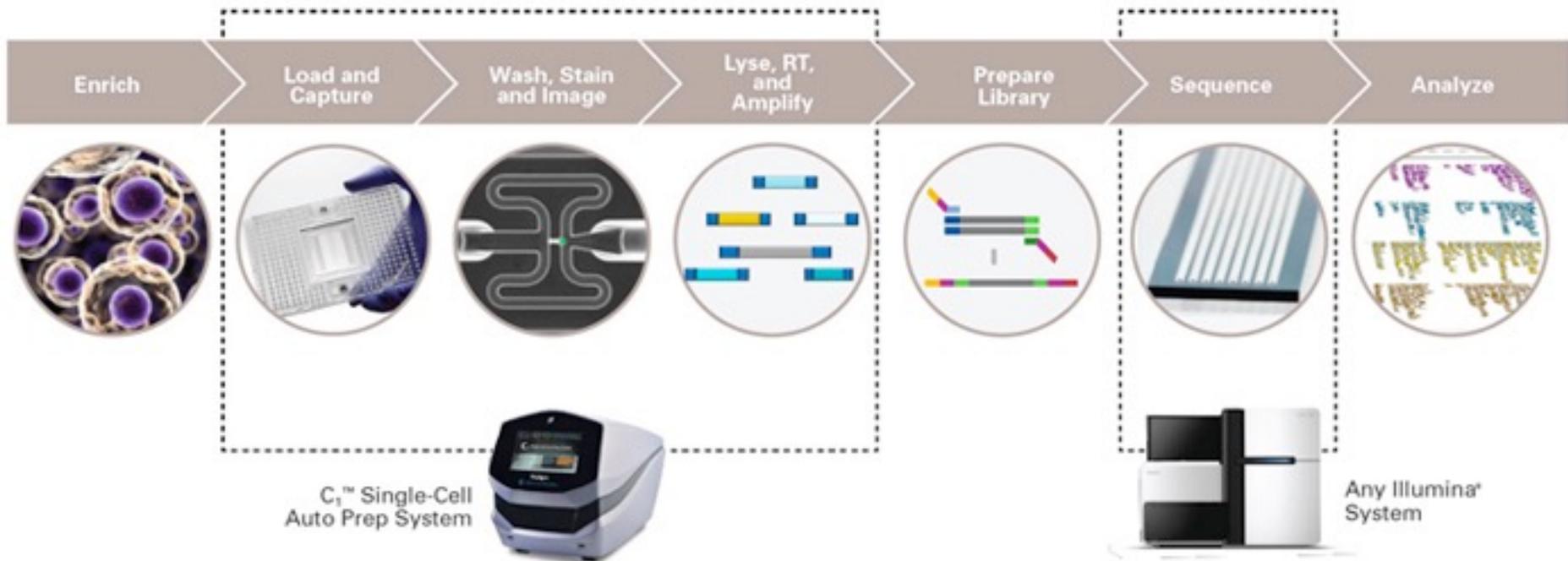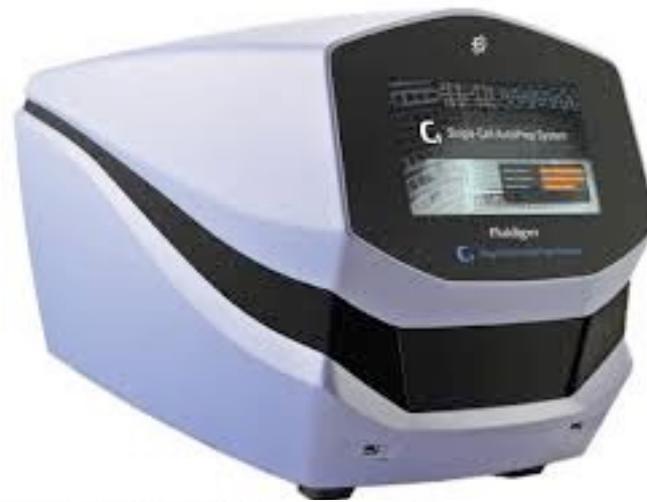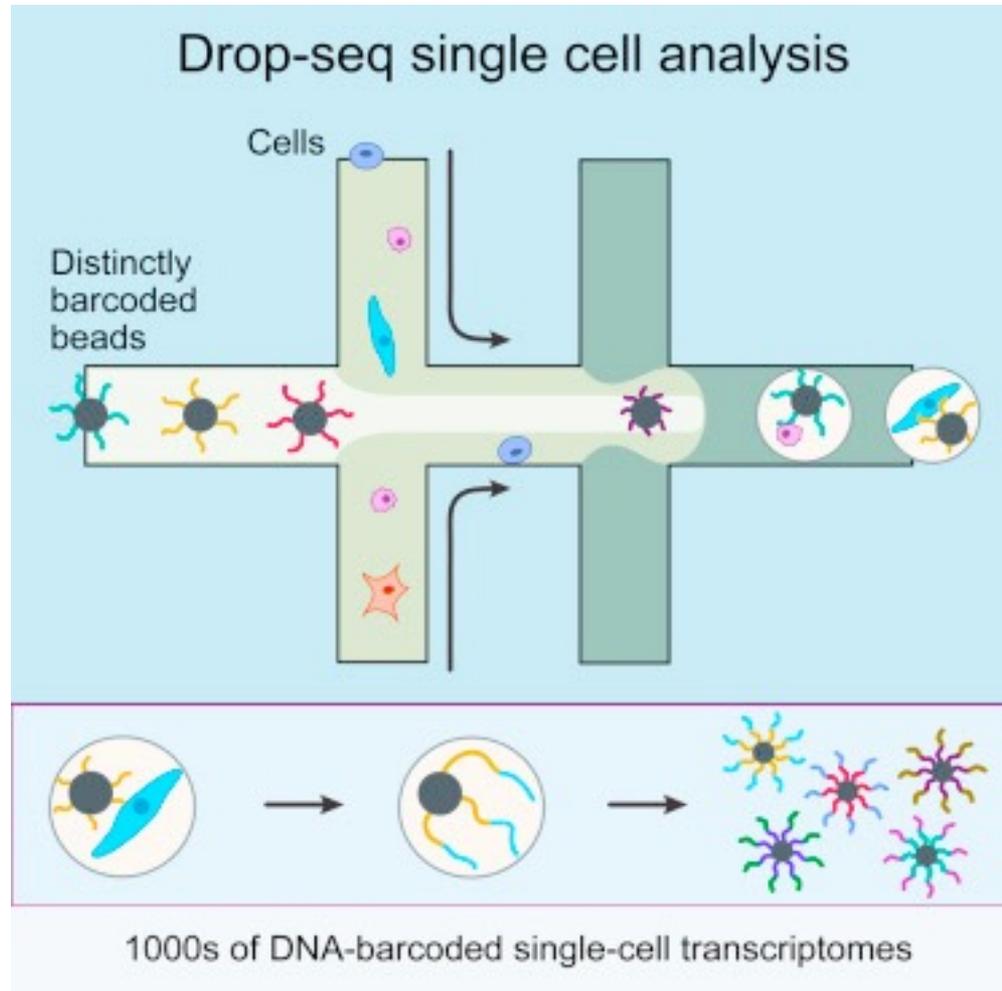
WaferGen iCell8
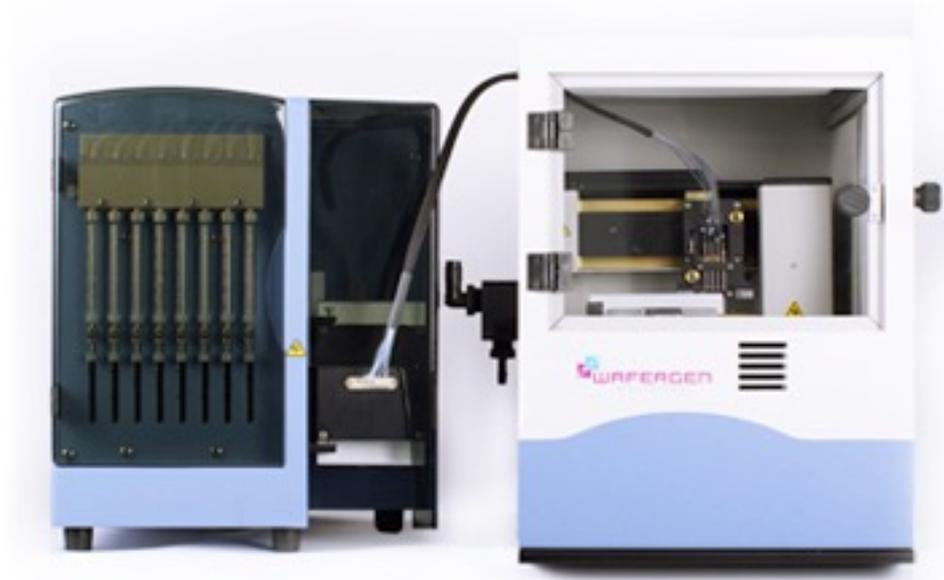
BioRad QX200 & ILMN system

QX200™ Droplet Digital™ PCR System

Chromium NGS
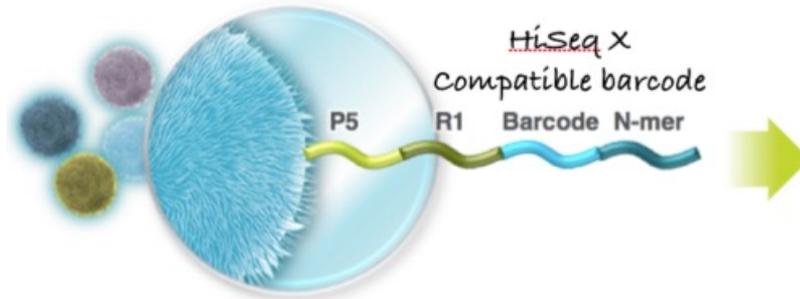
10X Genomics Single-Cell
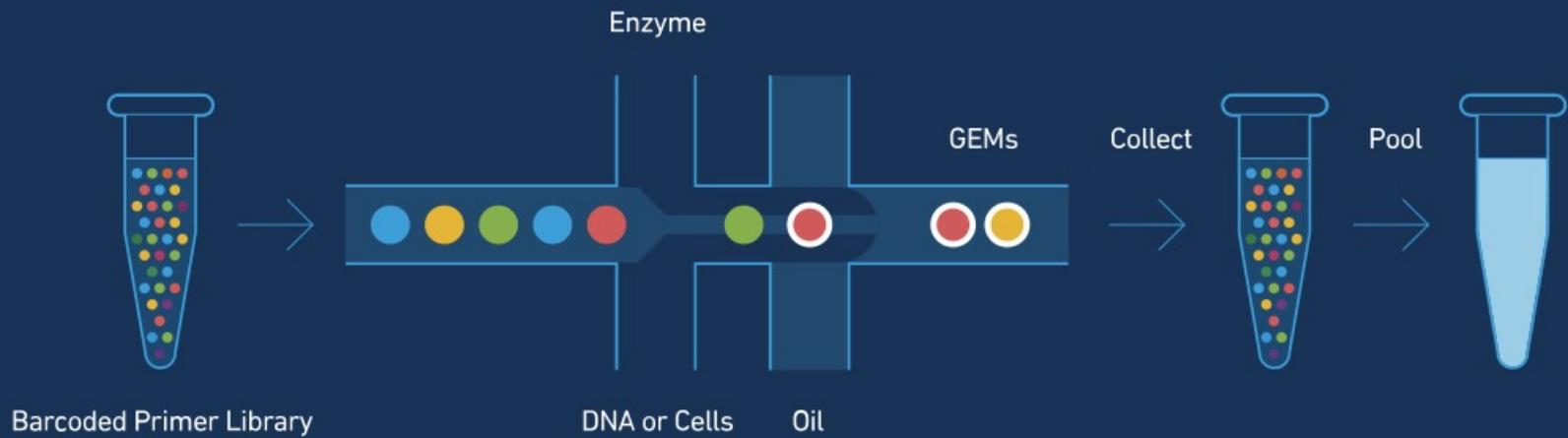
The explosion of scRNA-seq experiments

Svennson *et al.*, arXiv 2017
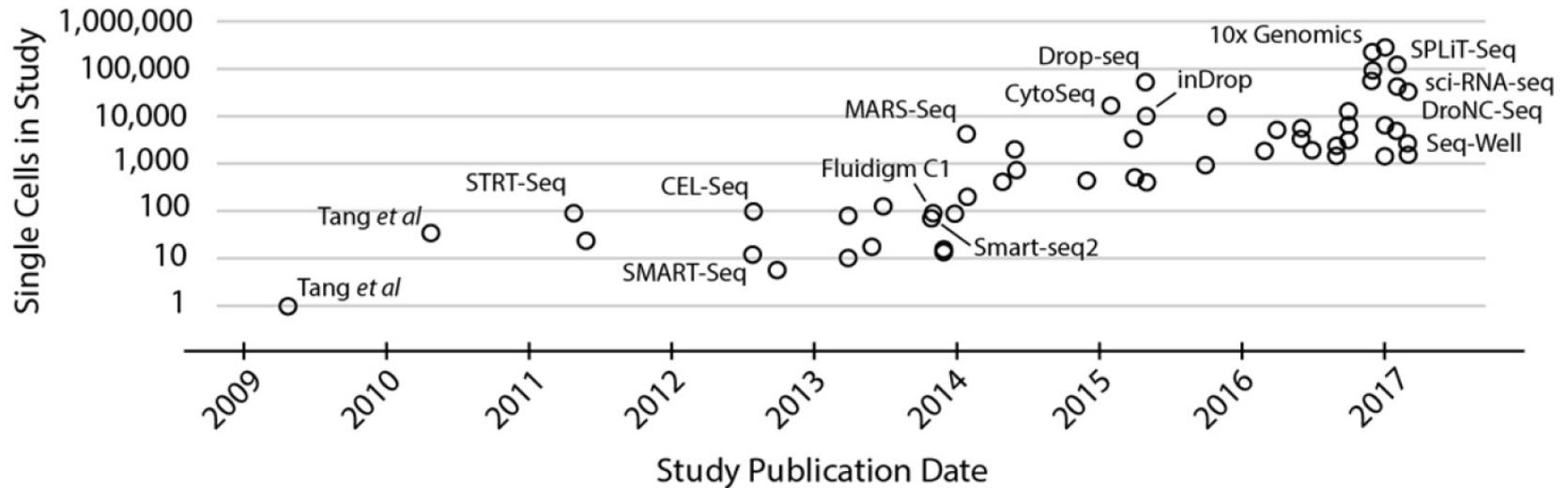
Many options today for single-cell sequencing

Source	Instrument	Number of Cells	input cells	est. cost per run	est. cost per cell	UMIs	Cell Phenotype	DNA	RNA	ATAC	3'	full cDNA	Size Range μ(m)
10X Genomics	Chromium	5,000	100,000	$ 1,290	$ 0.26	yes	no	yes	yes	yes	yes	no	1_60
Becton Dickinson	CSseq / BDPrecis	96	unk	$ 10,000	$ 104.17	yes	no	unk	unk	unk	unk	unk	5-100
Becton Dickinson	Resolve	10,000	50,000	$ 10,000	$ 1.00	yes	yes	unk	unk	unk	unk	unk	5-100
BioRad-ILMN	ddSeq	1,200	10,000	$ 1,200	$ 1.00	unk	no	no	yes	unk	unk	unk	unk
Drop-Seq	DropSeq	10,000	100,000	$ 1,000	$ 0.10	yes	no	no	yes	yes	yes	no	1-100
Fluidigm	C1	96	5,000	$ 1,900	$ 19.79	yes	no	yes	yes	yes	no	yes	5-10, 11-17, 17-24
Fluidigm	scRRBS	96	5,000	$ 1,900	$ 45.00	yes	no	yes	yes	yes	no	yes	5-10, 11-17, 17-24
Fluidigm	C1- high throughp	800	5,000	$ 4,000	$ 5.00	yes	no	yes	yes	yes	yes	no	5-10, 11-17, 17-24
Fluidigm	Polaris	800	5,000	$ 10,000	$ 12.50	no	yes	no	yes	no	yes	yes	5-10, 11-17, 17-24
In-Drop	custom	10,000	100,000	$ 5,000	$ 0.50	yes	no	no	yes	no	yes	no	5-100
Raindance	RainDrop	unk	unk	unk	unk	yes	no	unk	unk	unk	unk	unk	unk
QIAGEN	CellRaft	44,000	unk	unk	unk	unk	no	unk	unk	unk	unk	unk	unk
WaferGen	iCell8	1,800	40,000	$ 2,750	$ 1.53	yes	limited	soon	yes	unk	unk	maybe	5-100

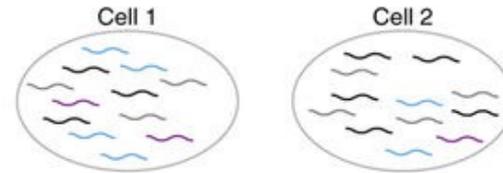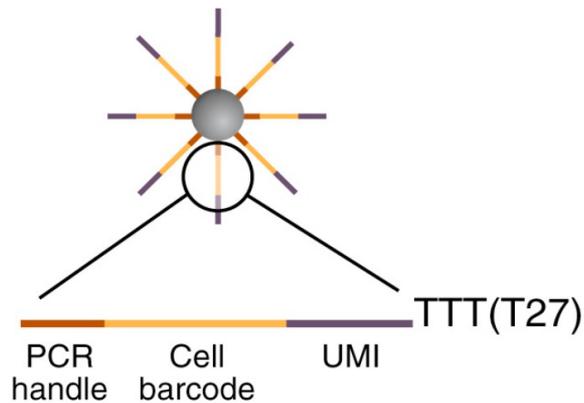# Single cell capture and RNA chemistry using nanodroplets

- Drop-seq

Beads

Cells + Enzymes

Oil

Single cell capture and RNA chemistry using nanodroplets

Beads

Cells + Enzymes

Oil

Barcoded beads

TTT(T27)

PCR handle   Cell barcode   UMI
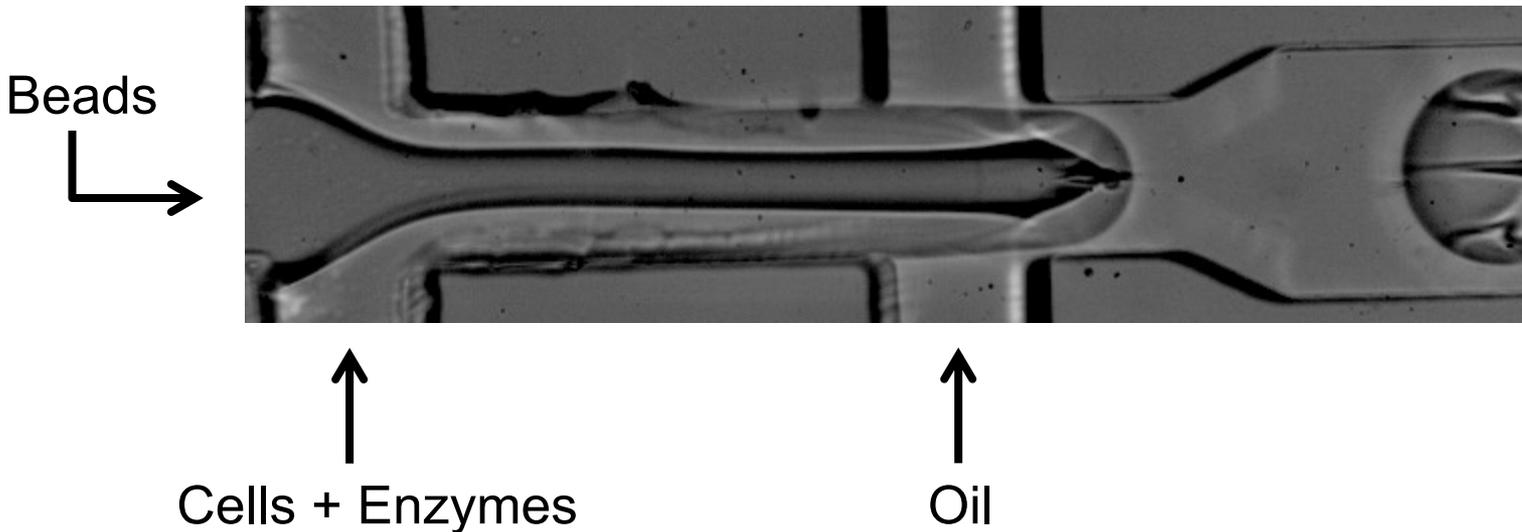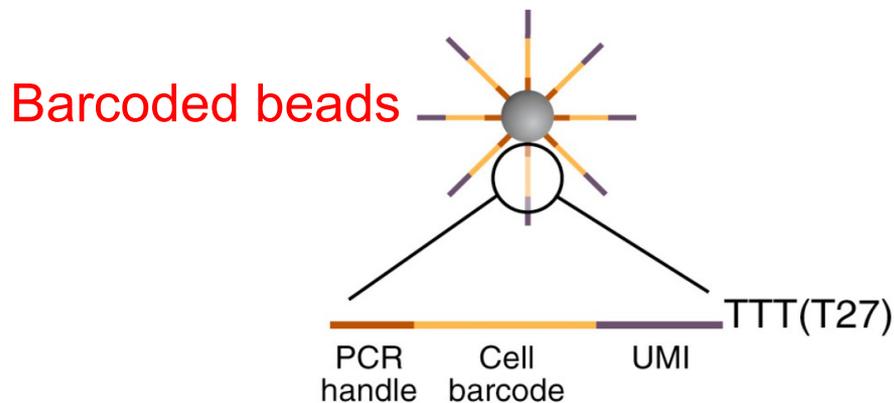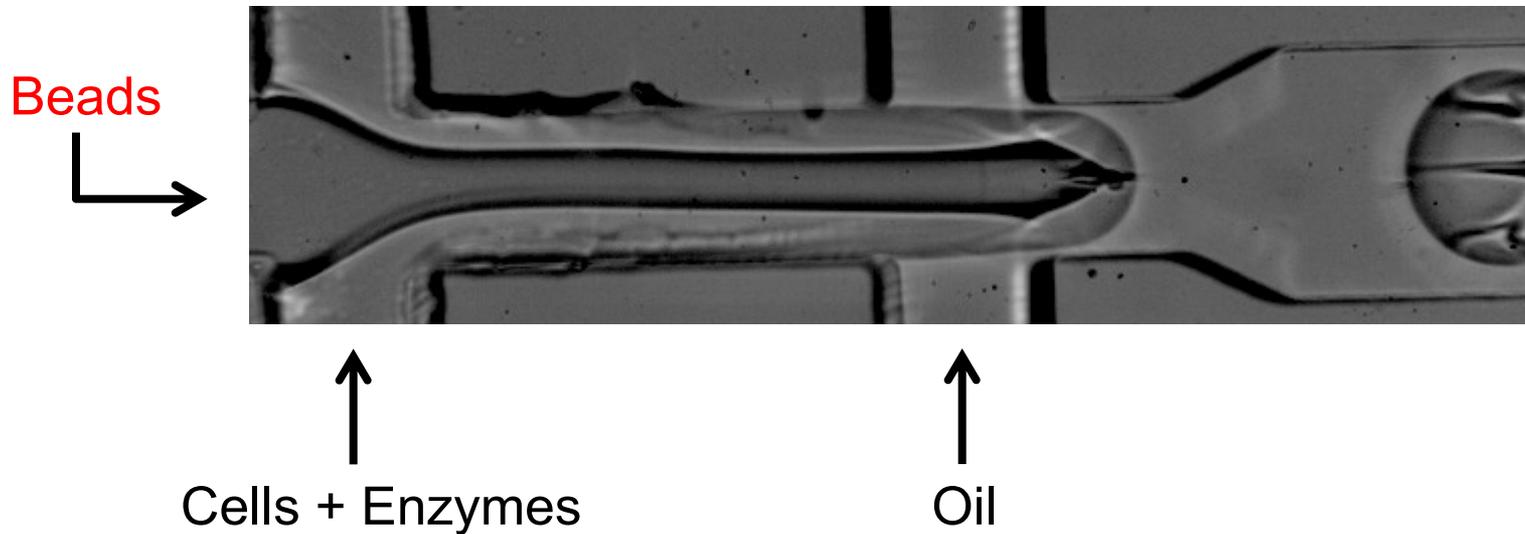
Unique Molecular Identifiers (UMIs)

Barcoded beads

Islam *et al.*, Nature Methods 2014

- CD45RA+ Naive T Cells
- CD4+ T Cells
- CD8+ T Cells
- CD14+ Monocytes
- CD19+ B Cells
- CD34+ Myeloid Progenitors
- CD56+ Natural Killer Cells
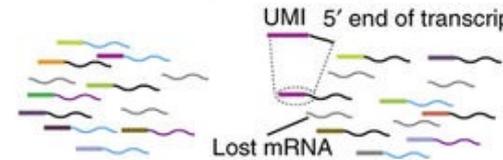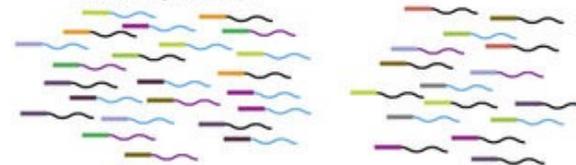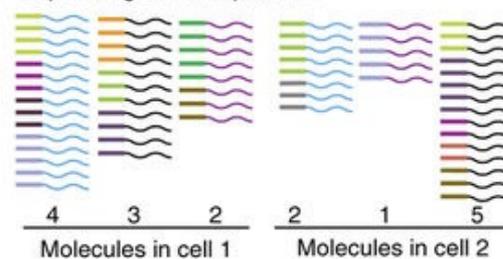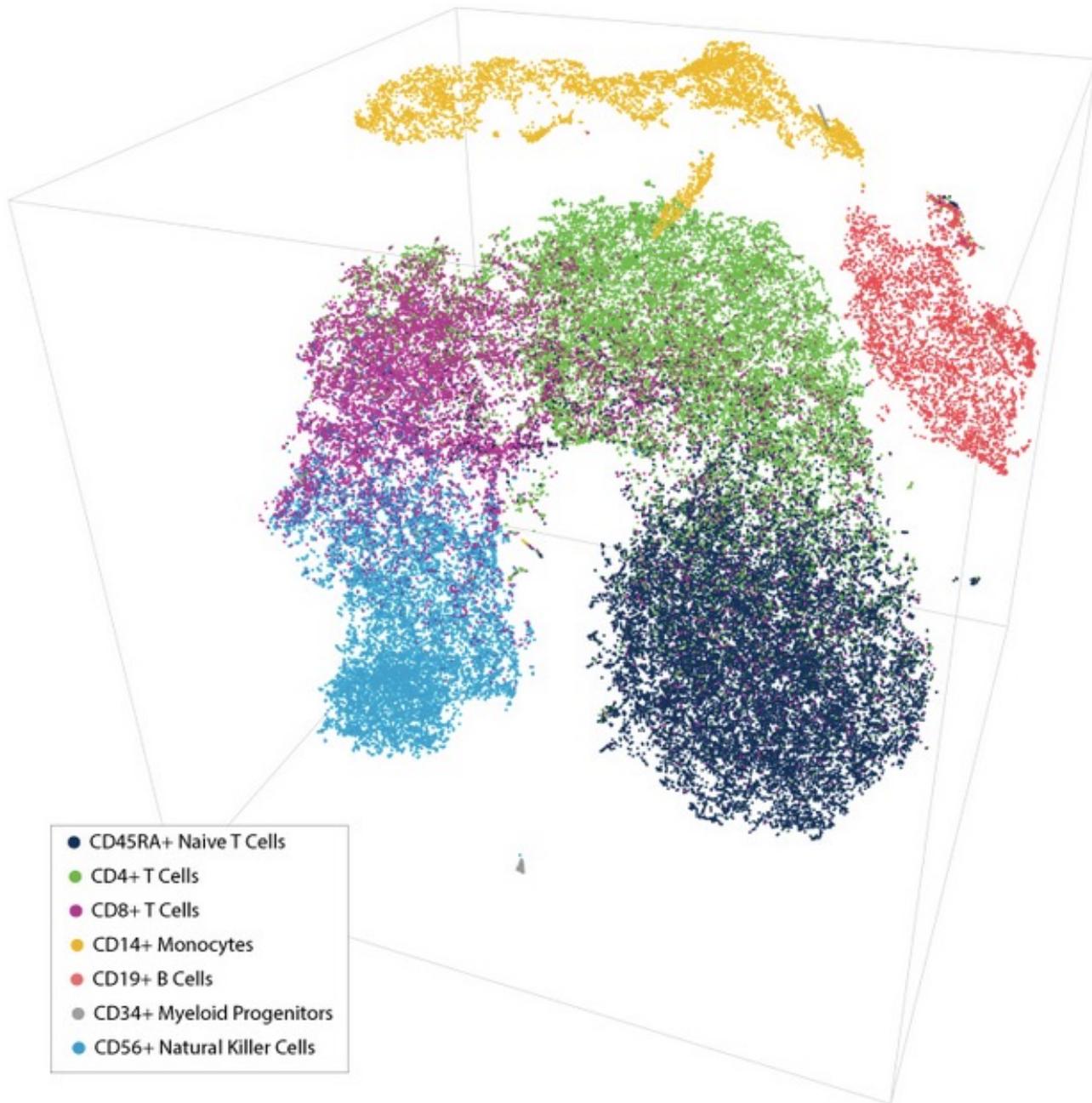
1.3 million neurons catalogued

1.3 million mouse embryonic brain cells, 10X Chromium
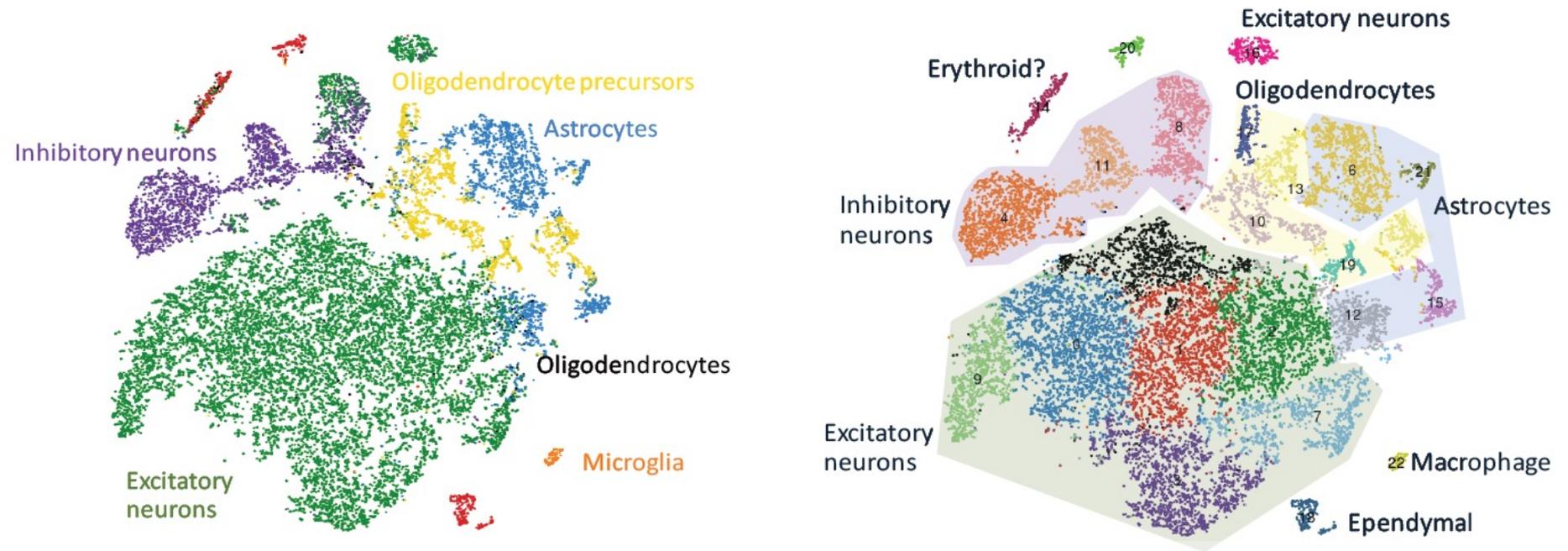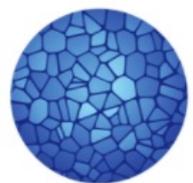
MISSION

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.

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	index sorting, CITE-seq
RNA + genome editing	Perturb-seq, CRISP-seq, CROP-seq

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

Affiliations | Contributions | Corresponding authors

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

📄 PDF | ⬇ Citation | 🗐 Reprints | 🔍 Rights & permissions | 📈 Article metrics

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.

ARTICLE PREVIEW

view full access options ▶

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

ARTICLE PREVIEW

view full access options ▶

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

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]

Other methods also emerging
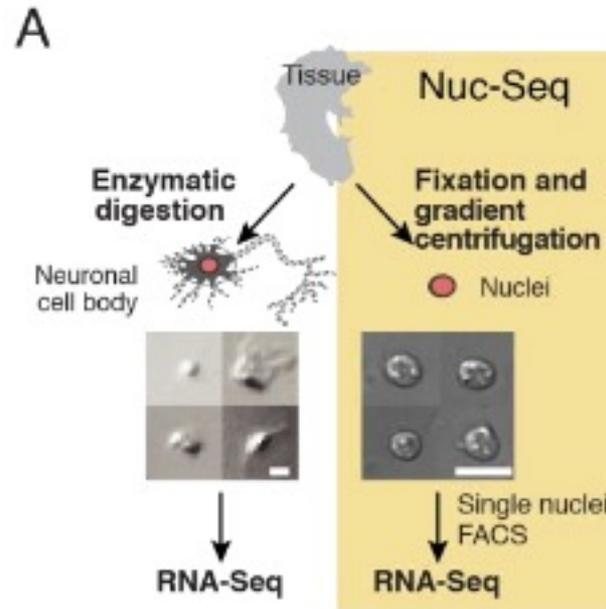
Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS

http://biorxiv.org/content/biorxiv/early/2016/03/27/045989.full.pdf

scATAC/RNA-seq

Indexed RT

Indexed transposition

Pool and sort

Cell lysis and split

Indexed PCR for RNA-seq

Indexed PCR for ATAC-seq

sci-RNA-seq library

P5 R1 UMI cDNA R2 P7

i5 RT oligo(dT)30 i7

PCR barcode ('i5'+ 'i7'), and RT barcode (RNA-seq) or linked Tn5 barcode (ATAC-seq) comprise a cellular index.

P5 i5 N5 R1 DNA R2 N7 i7 P7

Adaptor Adaptor

sci-ATAC-seq library

https://science.sciencemag.org/content/361/6409/1380
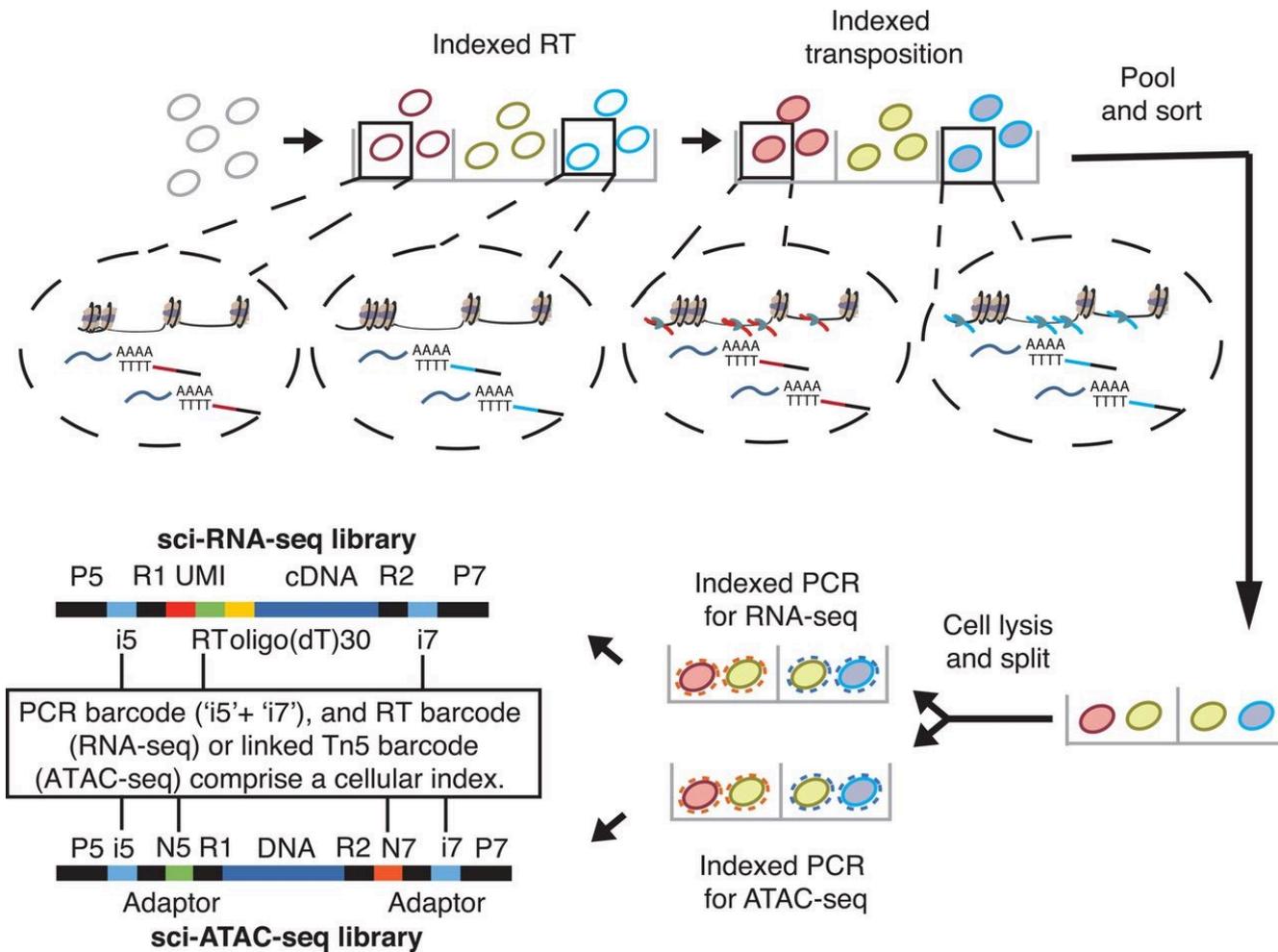
Analysis:
Structure of a generic pipeline
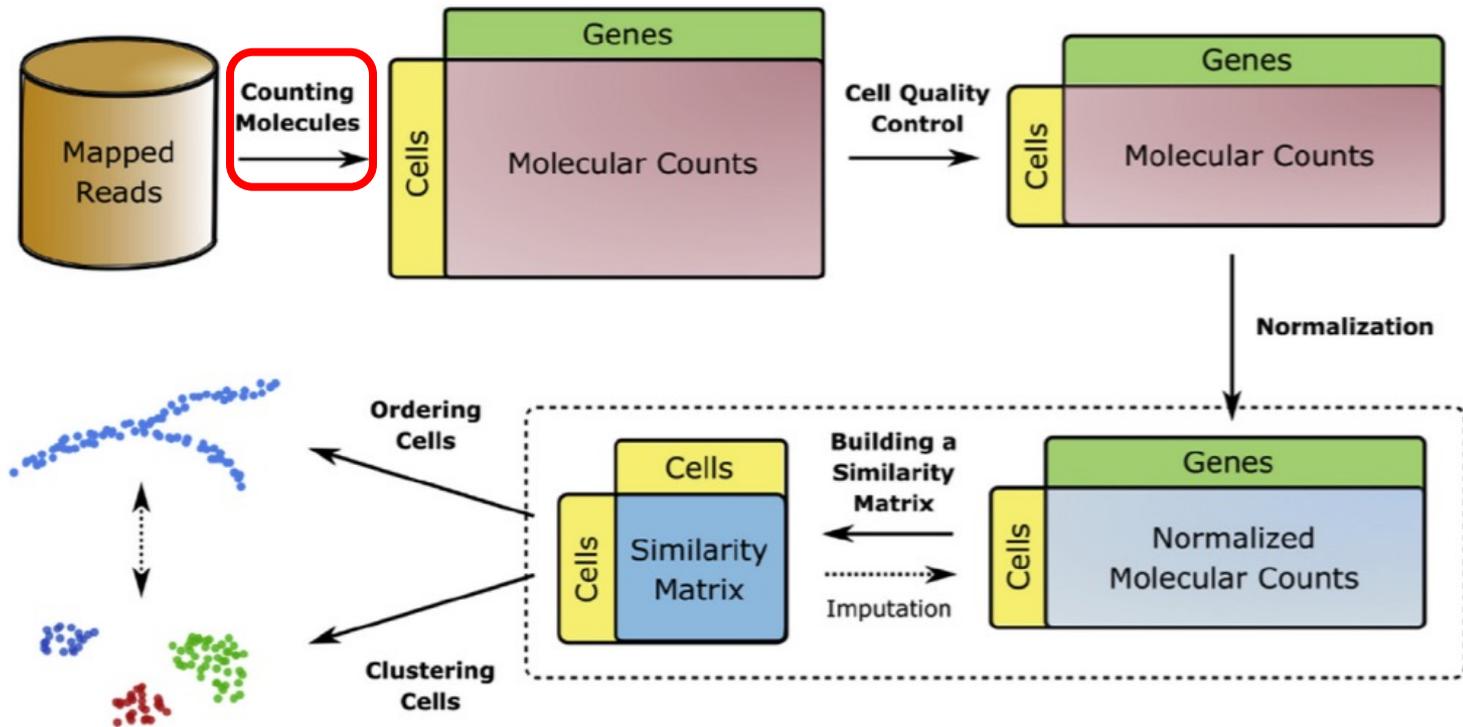
Counting Molecules

- ## 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
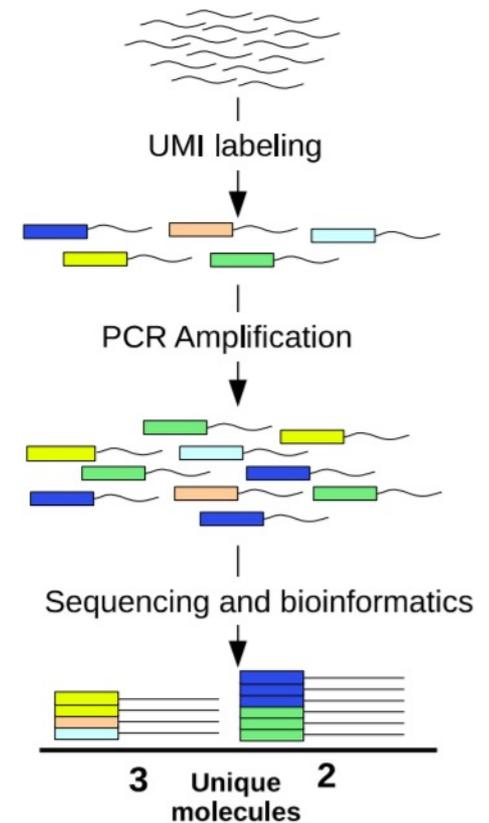
Commonly used open-source tools

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) – **featureCounts**

5. Count unique UMIs per gene – **UMI-tools**

6. QC – **fastqc, picard, multiqc, custom scripts**

Structure of a generic pipeline
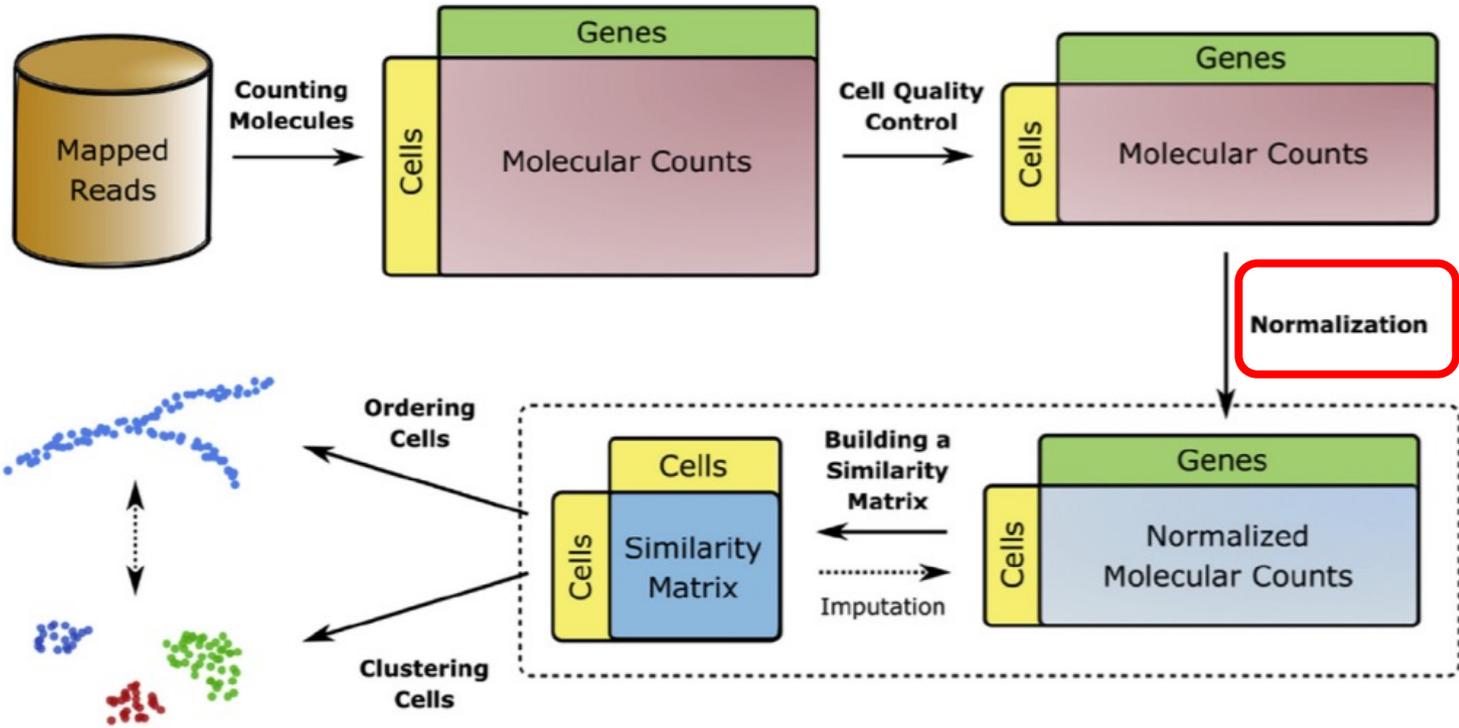
Normalization challenges
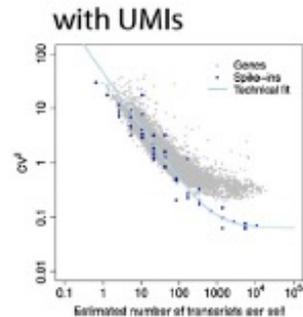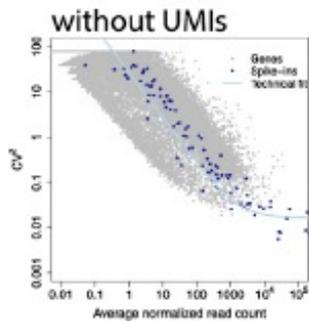
Kolodziejczyk *et al*., Briefings in Functional Genomics 2017
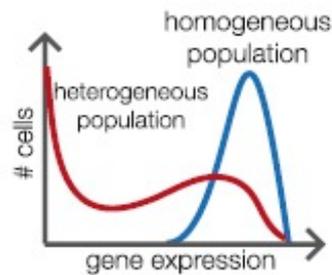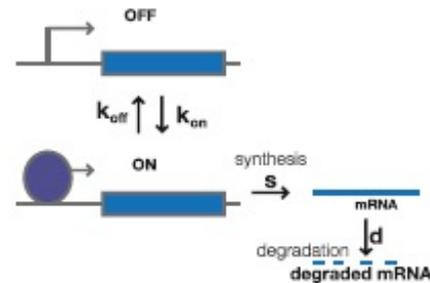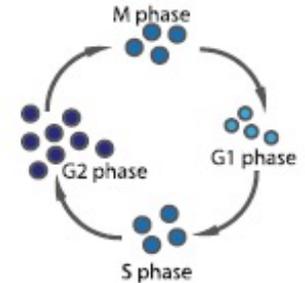
Normalization + Differential Expression Analysis

Soneson and Robinson, Nature Methods 2018
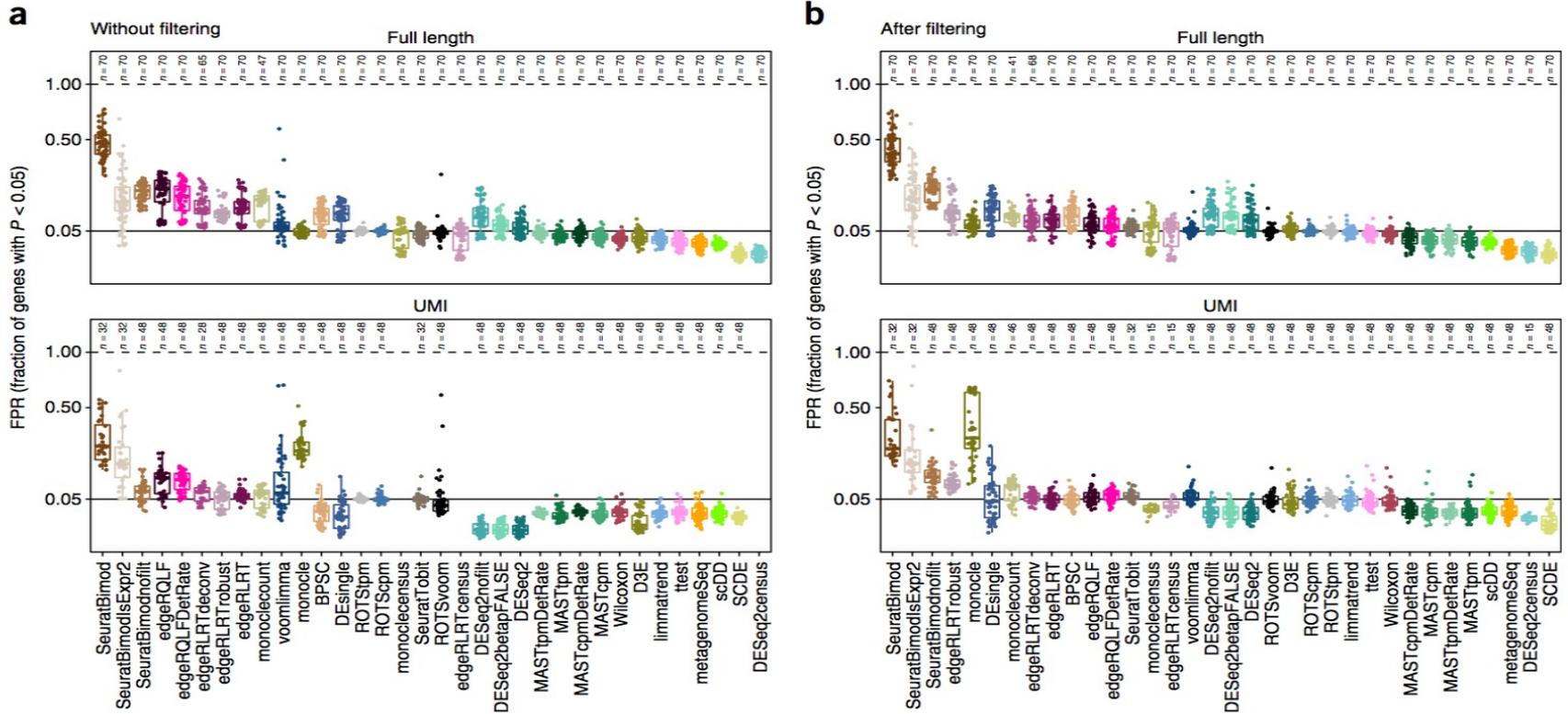
Structure of a generic pipeline

Gene Expression Imputation


Satija *et al*., Nature Biotechnology 2015
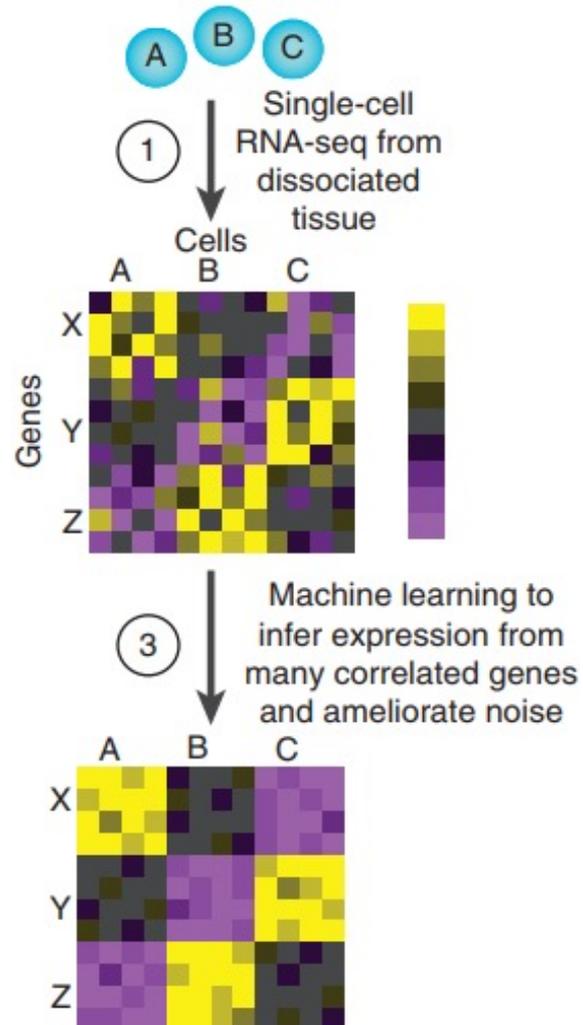
Gene Expression Imputation

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.

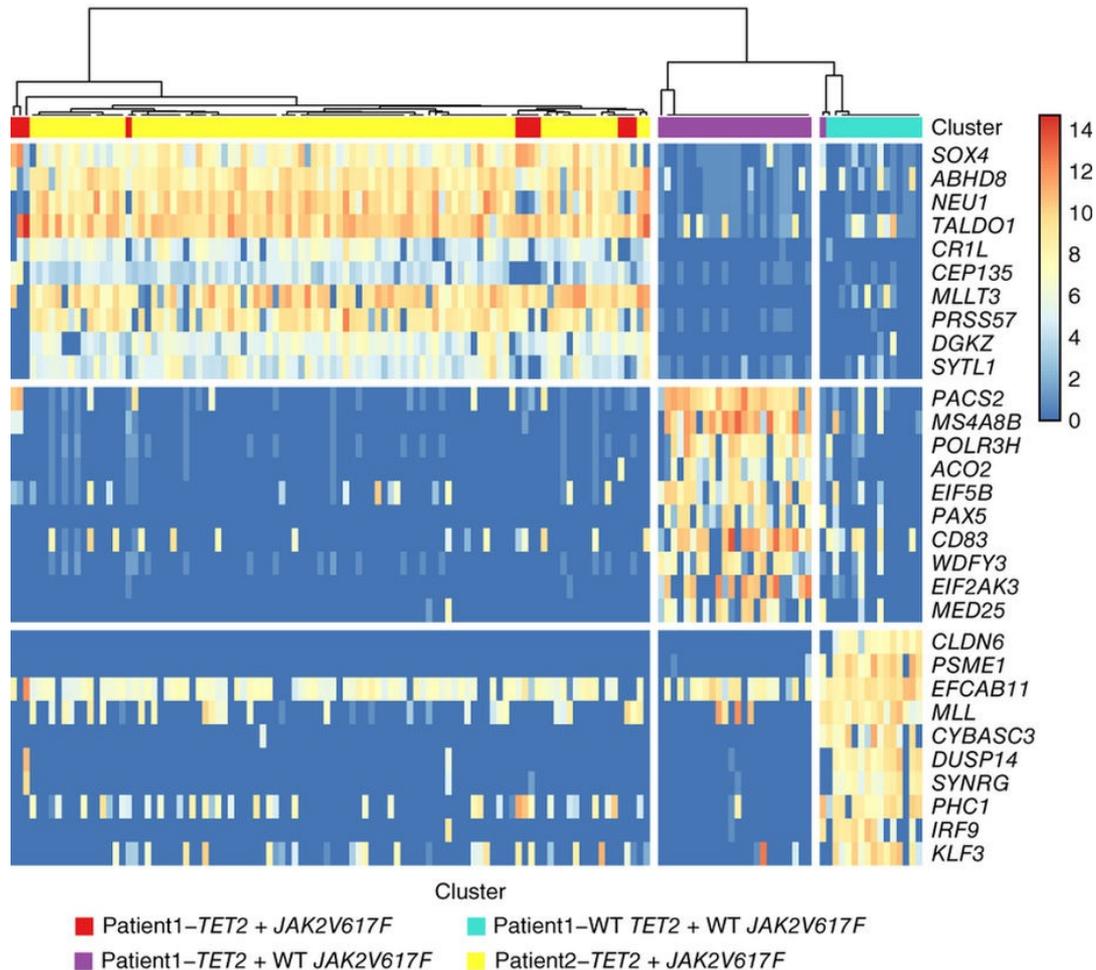Zhang and Zhang, Biorxiv 2017

Structure of a generic pipeline
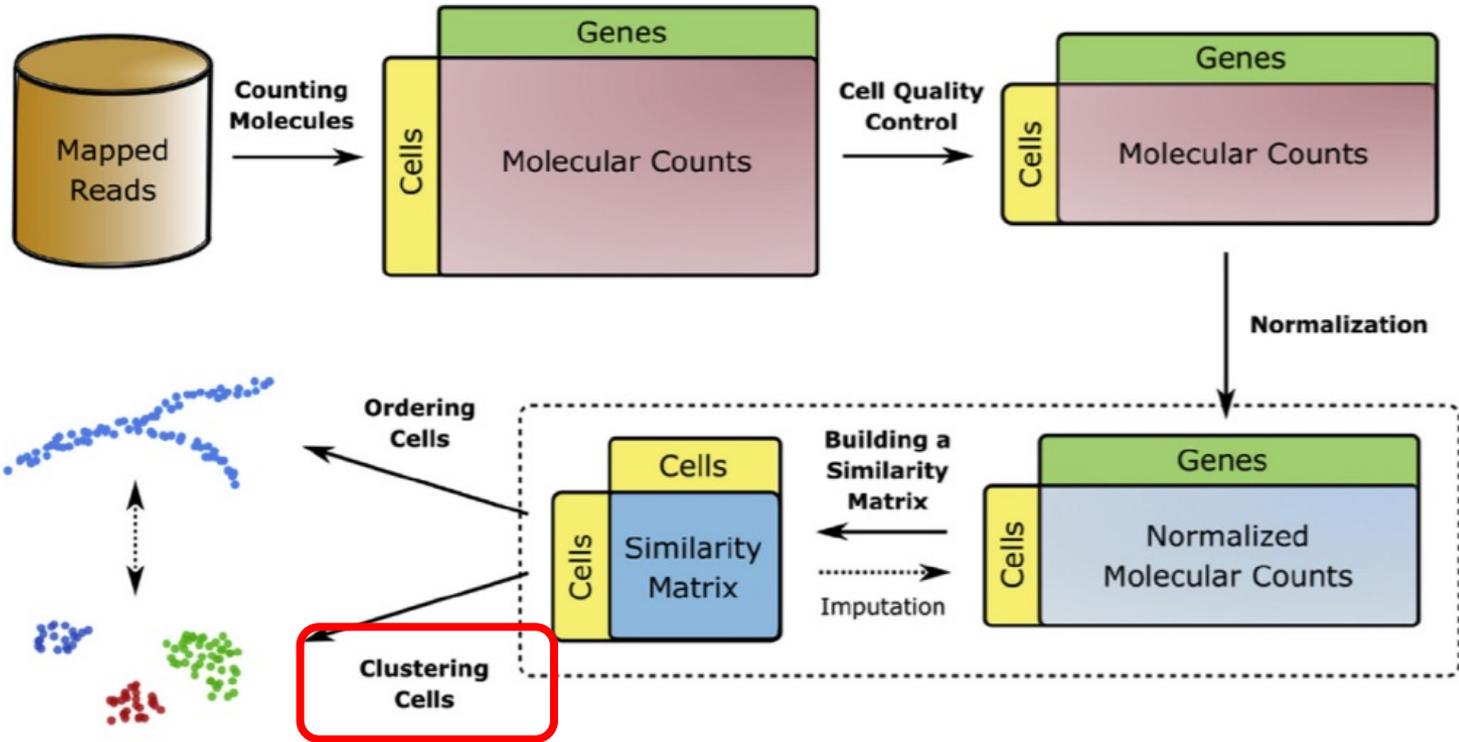
Clustering Cells

SC3: consensus clustering of single-cell RNA-seq data
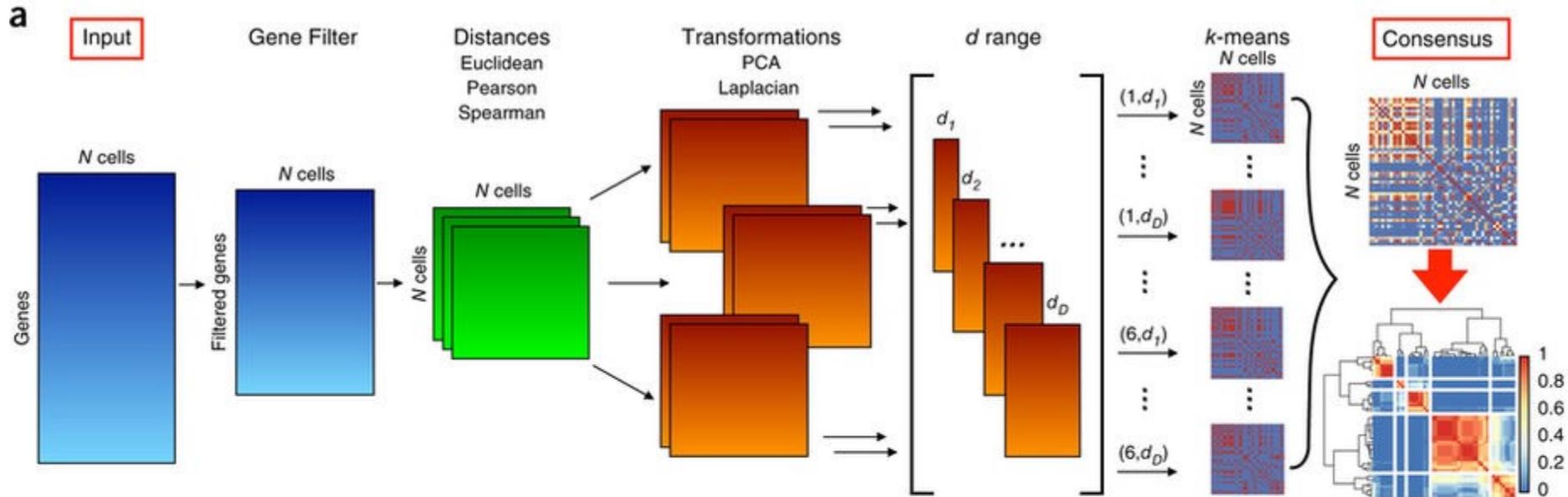
Differential Expression Analysis

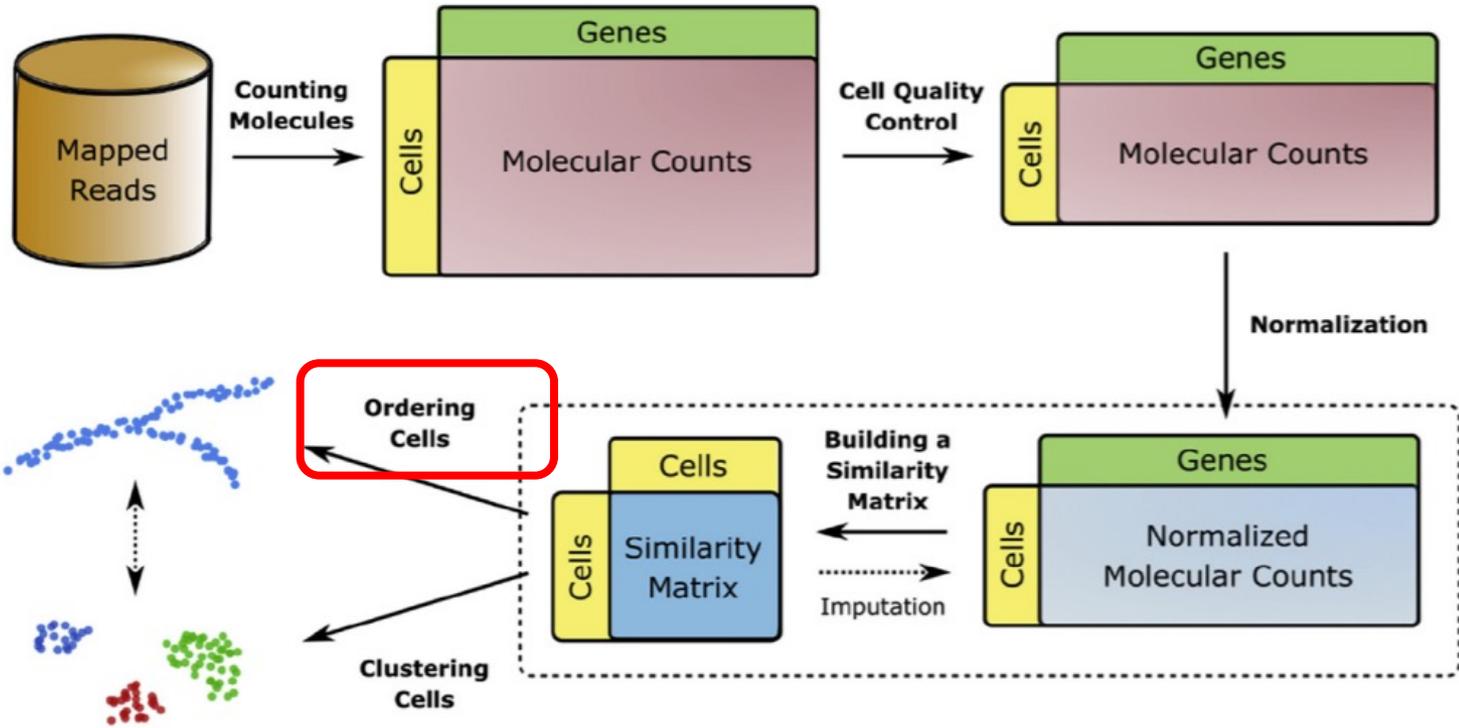SC3: consensus clustering of single-cell RNA-seq data

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
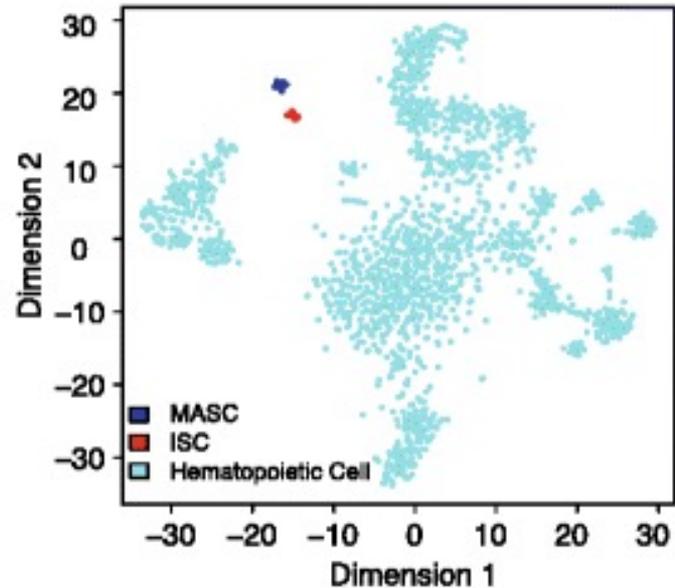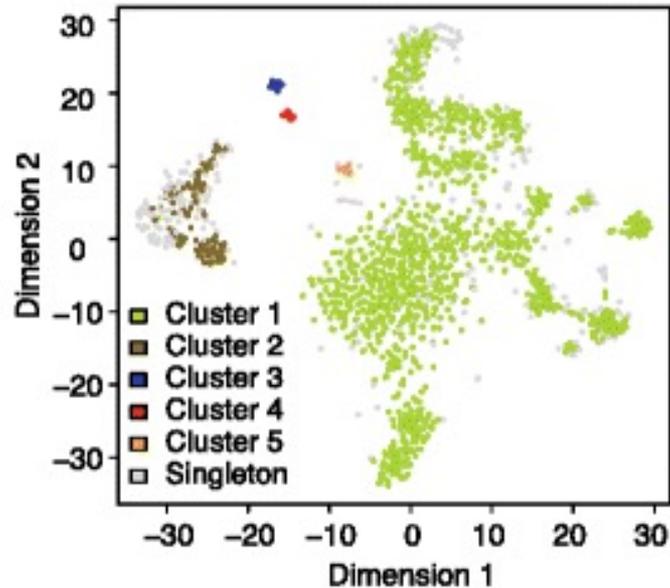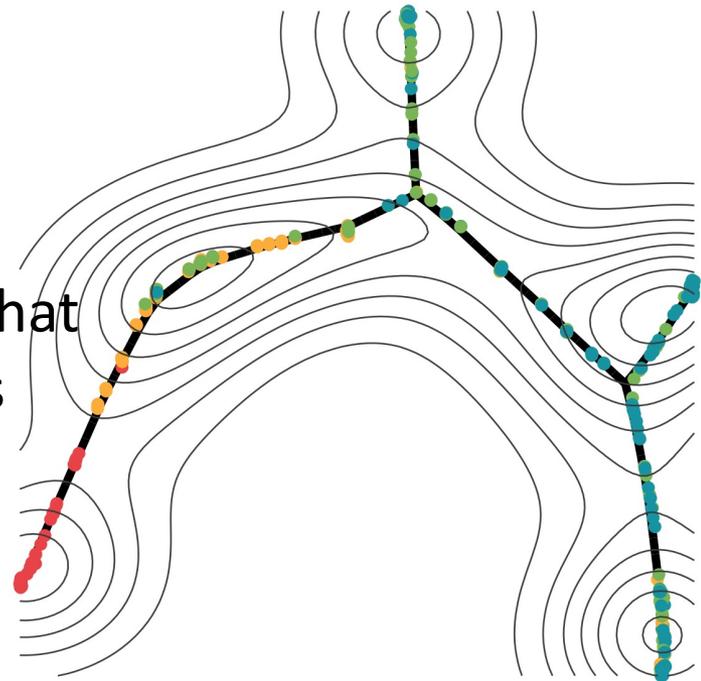
Structure of a generic pipeline

Single Cell Trajectory Inference

- "Pseuodotime" 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

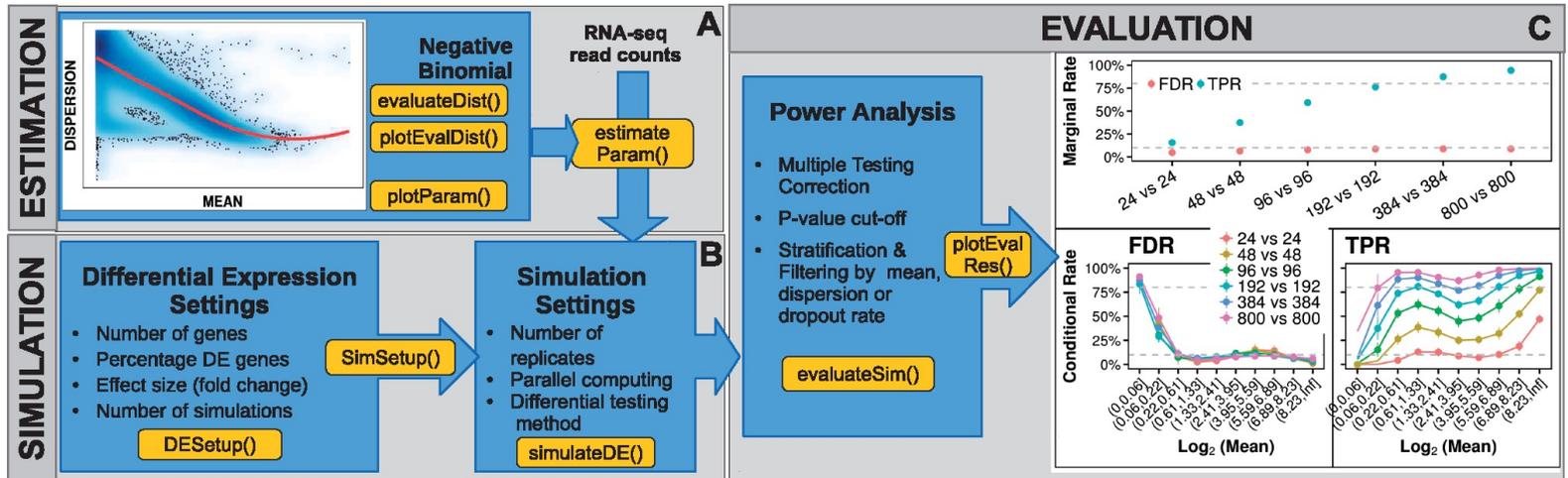- "Pseuodotime" introduced in Trapnell *et al.*, Nature Biotechnology 2014 (Monocle)

- Steps:

  1. (Optional) Choose genes that define a biological process

  2. Reduce dimensionality
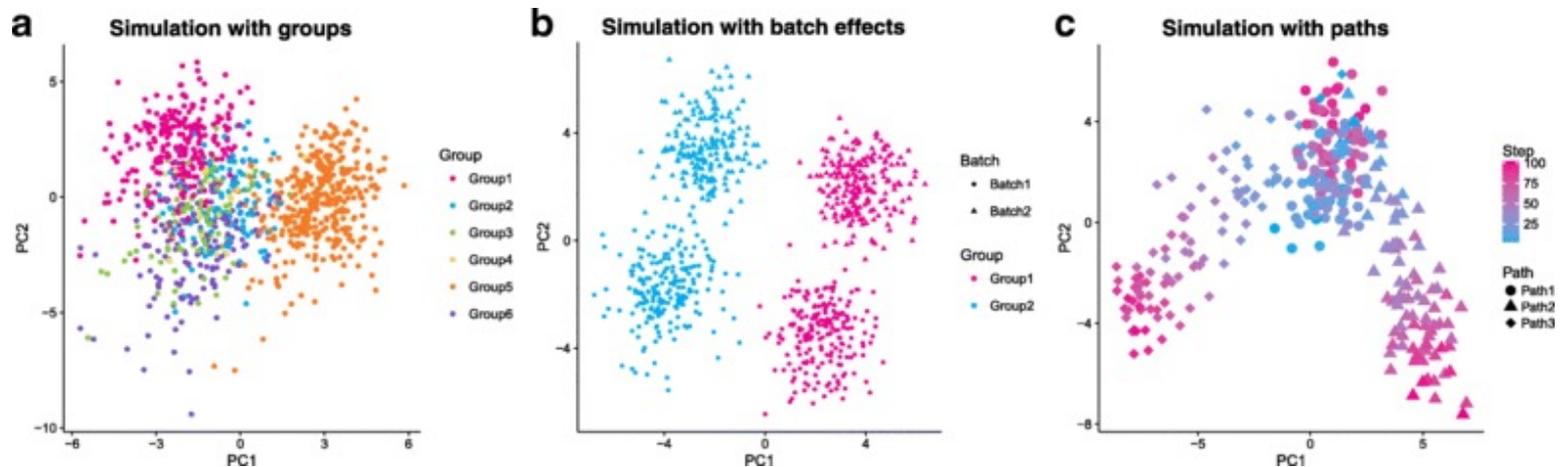
Differential Expression Analysis using Monocle

Qiu *et al.*, Nature Methods 2017
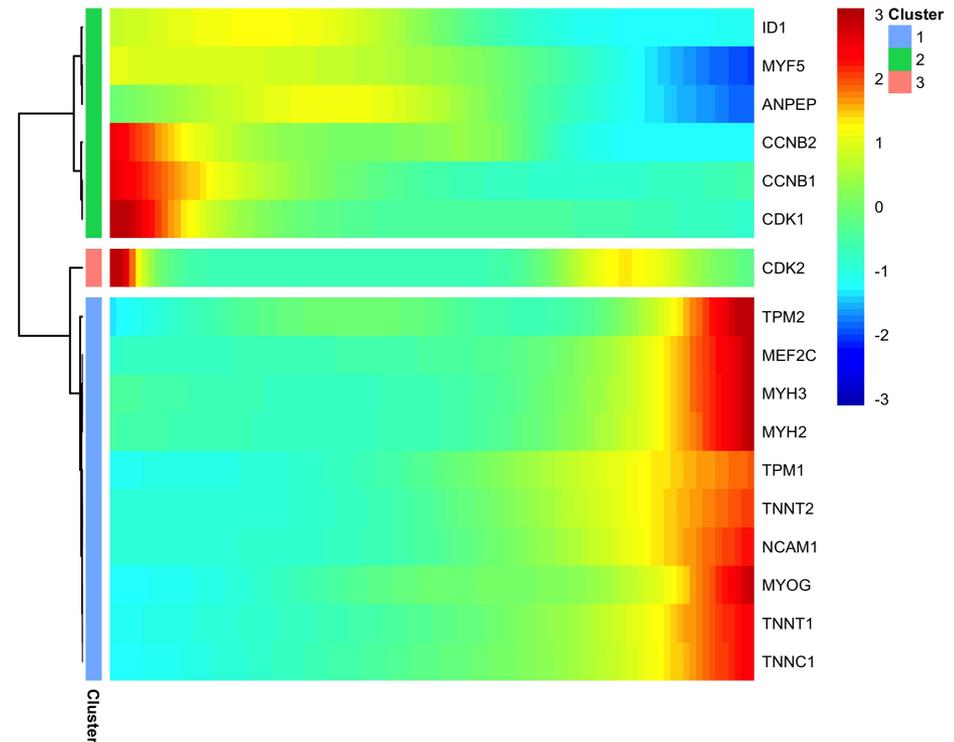
Simulating scRNA-seq data

PowSimR

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*** (iD) ○, **Robrecht Cannoodt*** (iD) ○, Helena Todorov ○, *Yvan Saeys* ○
> bioRxiv:276907 doi:10.1101/276907

https://github.com/dynverse/dynverse

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

GeoMx™
Digital Spatial Profiler

NanoString's GeoMx™ Digital Spatial Profiler provides morphological context with high-plex protein or gene expression profiling from a single slide.

RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERF

Jeffrey Moffitt

FEI CHEN
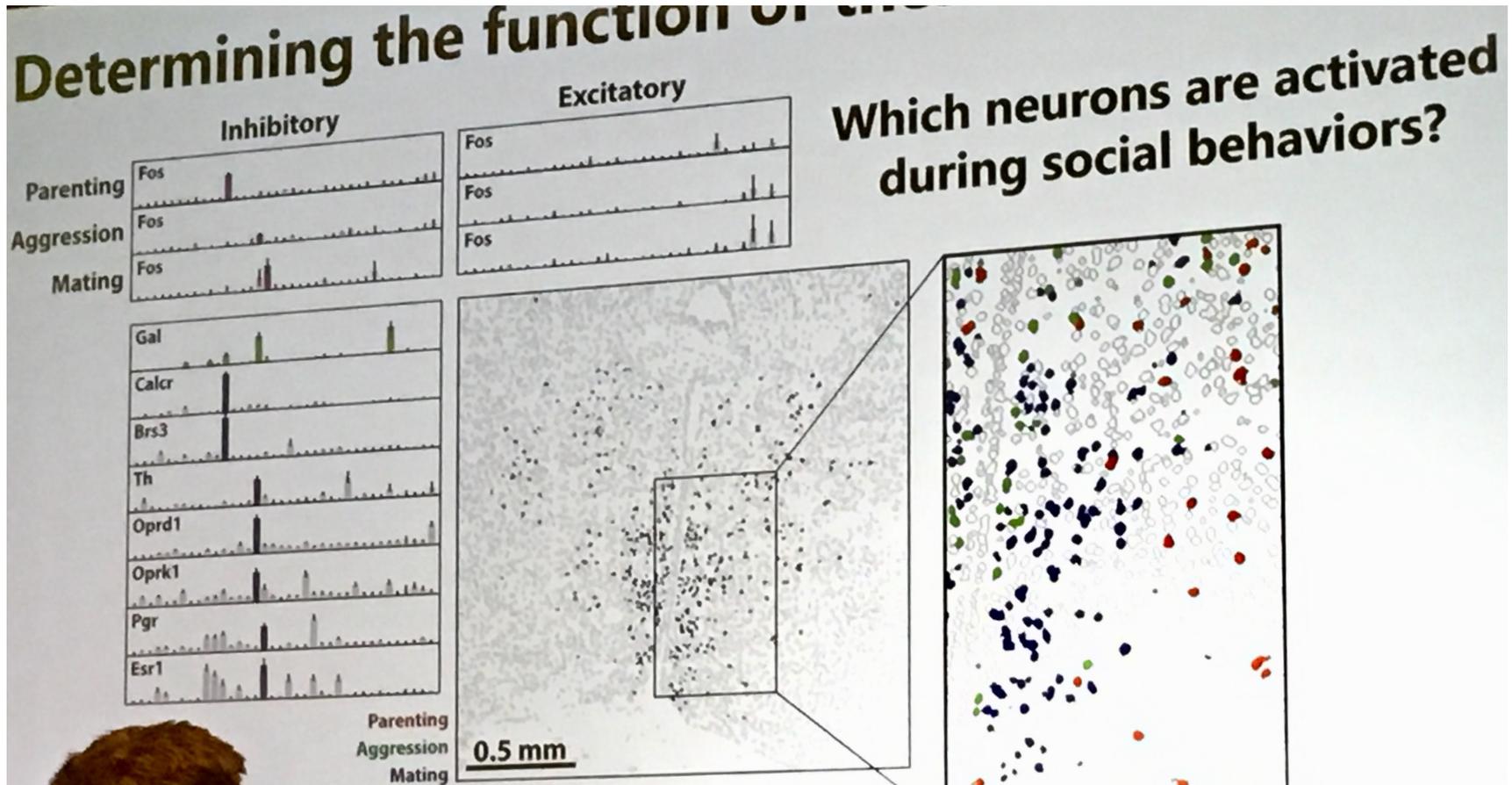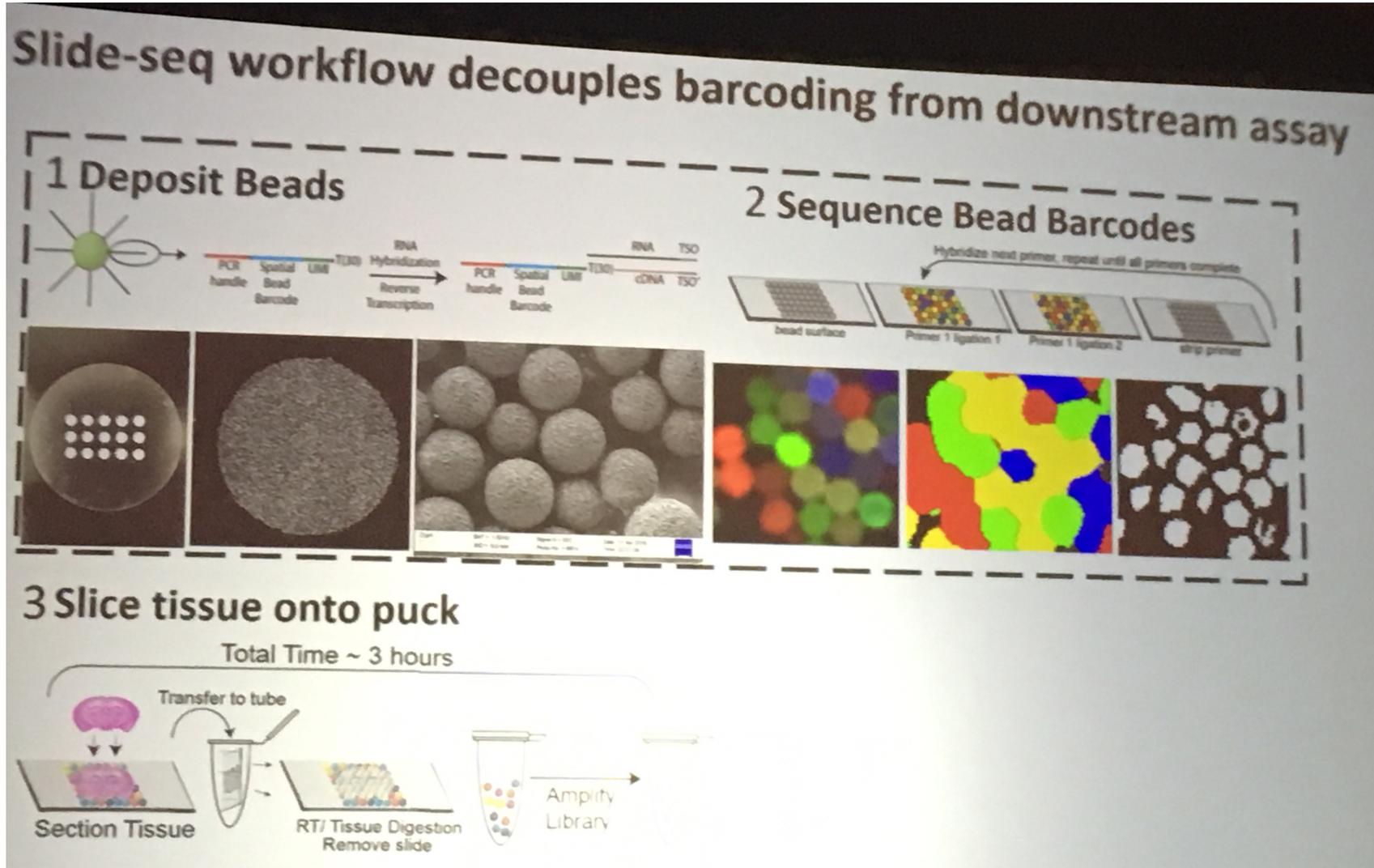
Questions?

Thanks also to Dr. Priyanka Vijay!