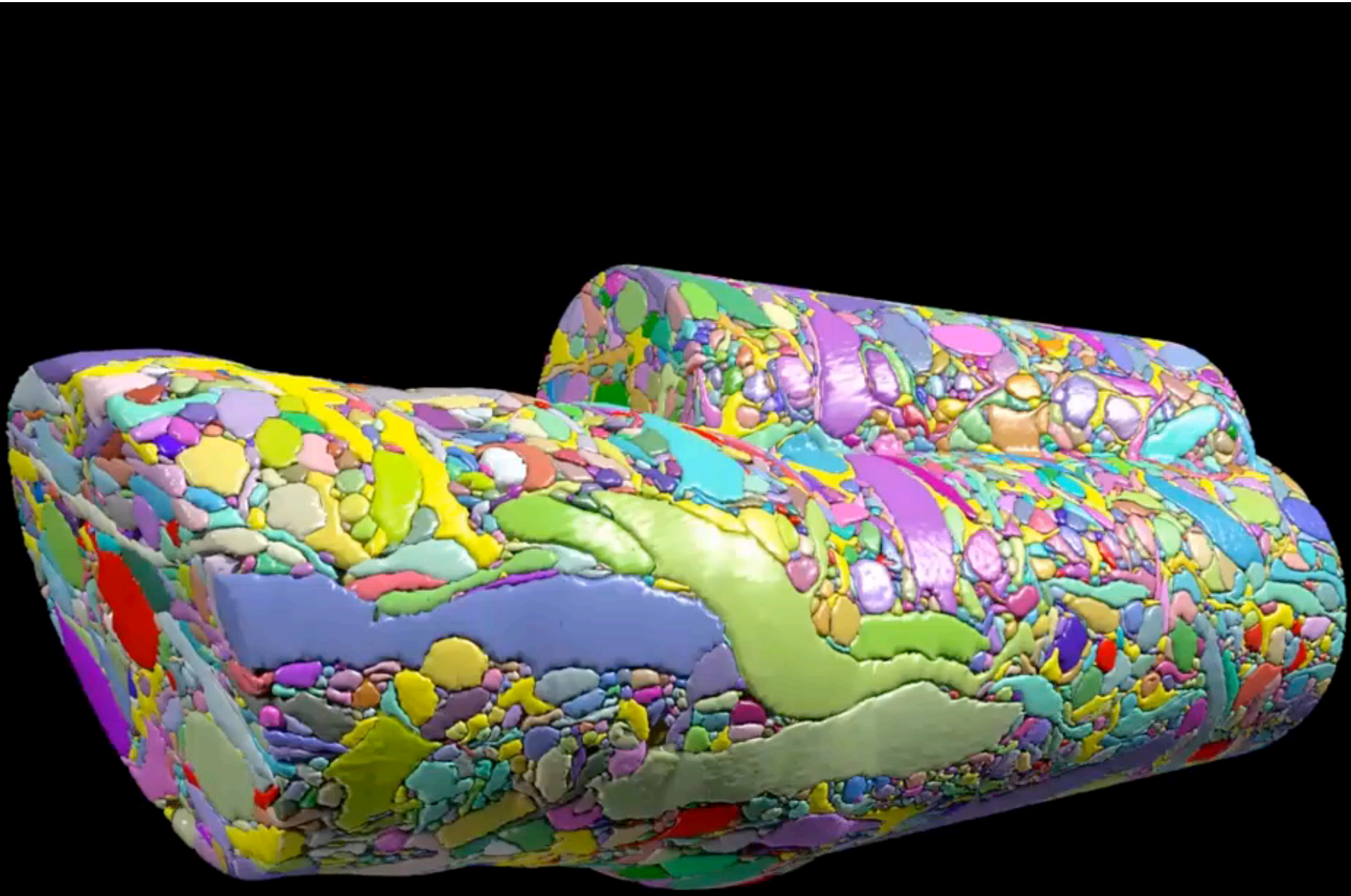
integrated analysis of the transcriptome, translatome, and proteome

Rob Kitchen

GPMTG

18 Feb 2016



Lichtman 2015 Cell

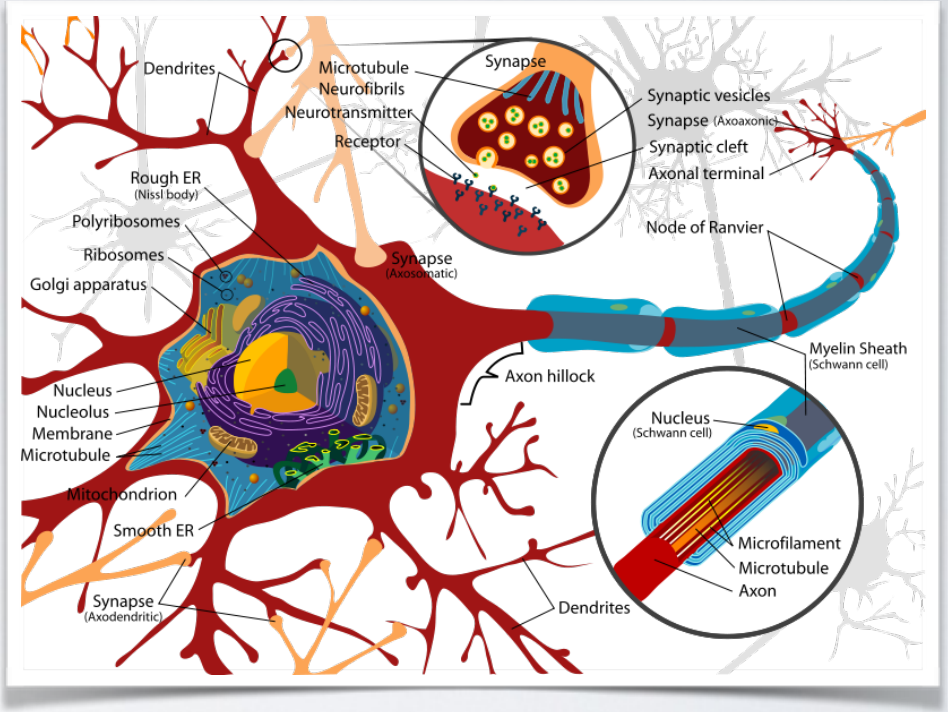
dealing with brain complexity

most highly evolved organ:

~ **10^{11} neurons**

~ **10^{14} synapses** (connections)

~ **5-50Hz firing rate**



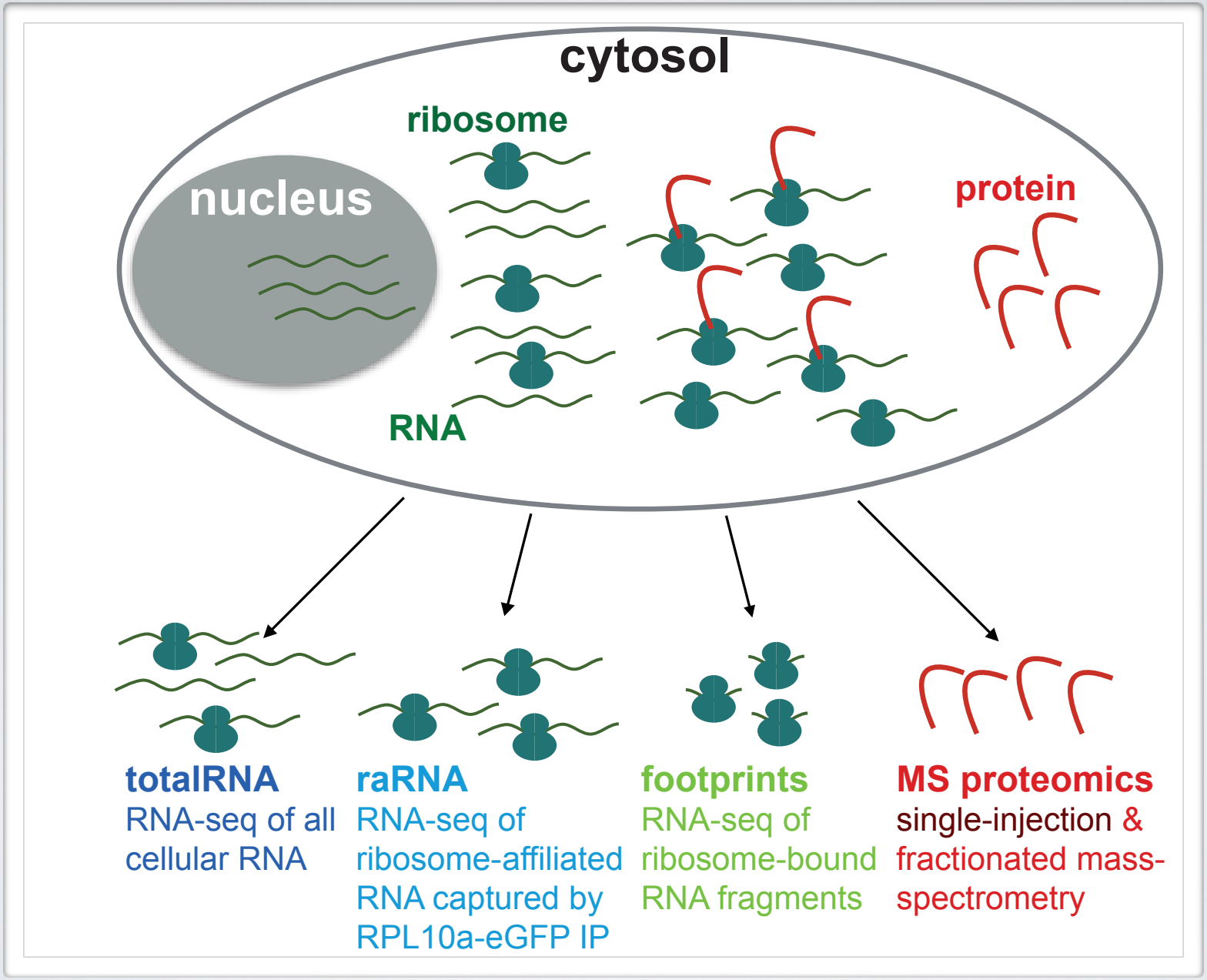
in almost every tissue, cells continually grow and divide but **most neurons stop growing** in early childhood

there is large interest in both **long-term** processes (Alzheimer's, Parkinson's, etc) and **short-term** processes (reward, addiction, PTSD)

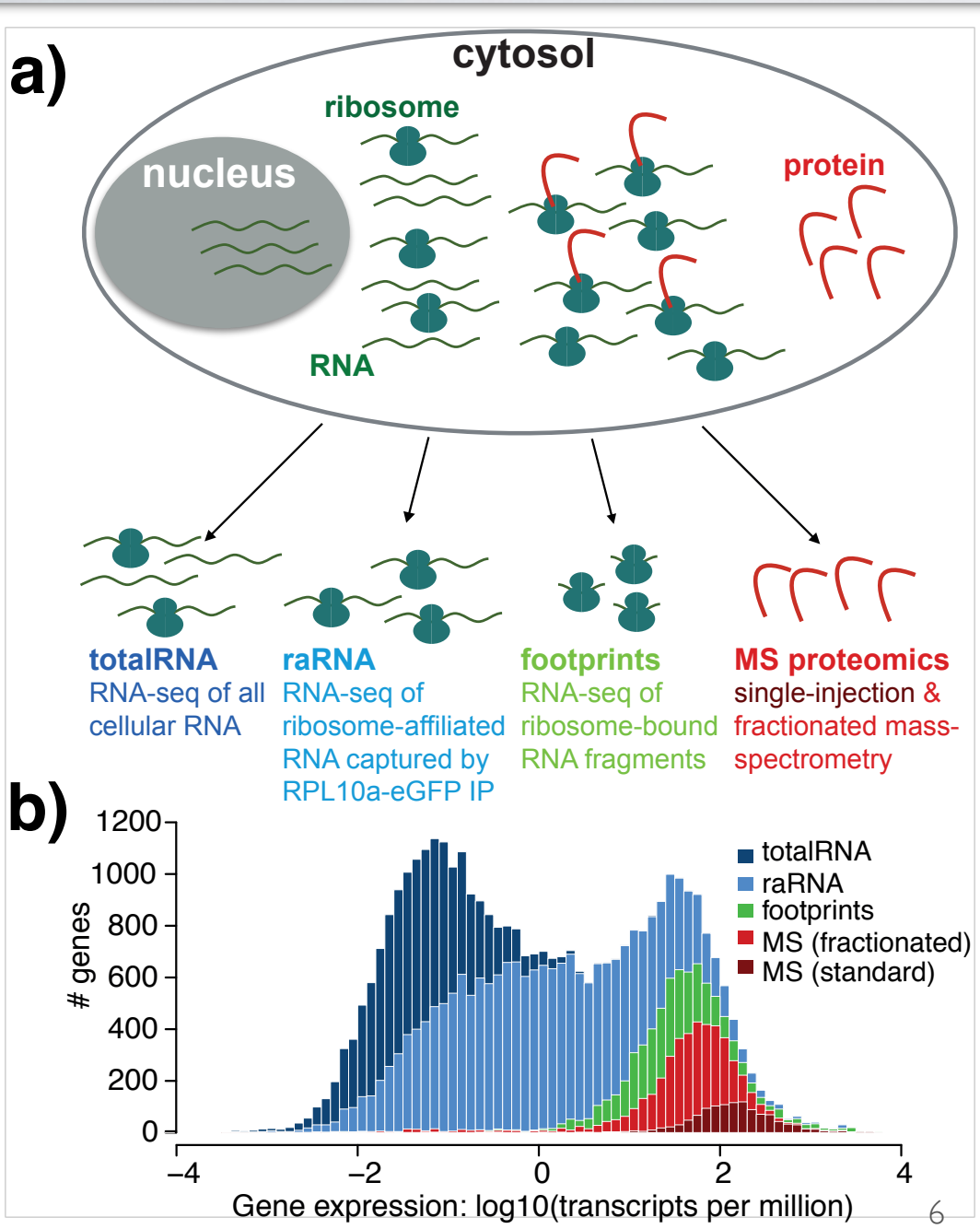
maybe more than any other organ, **post-transcriptional regulation is extremely important**

as is the scope for **alternative splicing…**

parallel, multi-level observations of gene expression



parallel, multi-level observations of gene expression



…although it is possible to assay translation and steady-state protein levels, the yield of these experiments is much lower than RNA-seq

easier to measure transcription than translation

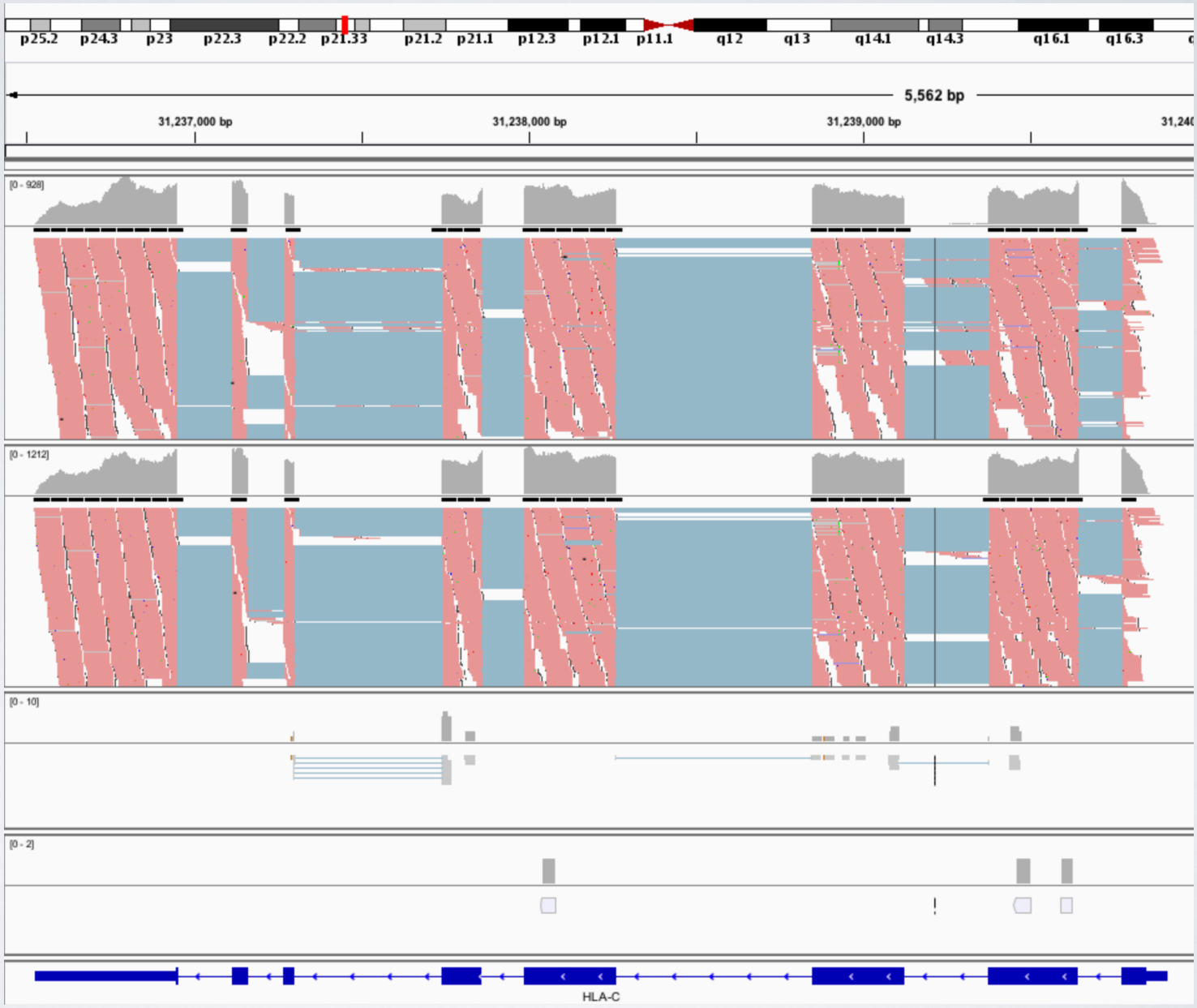
coverage

totalRNA

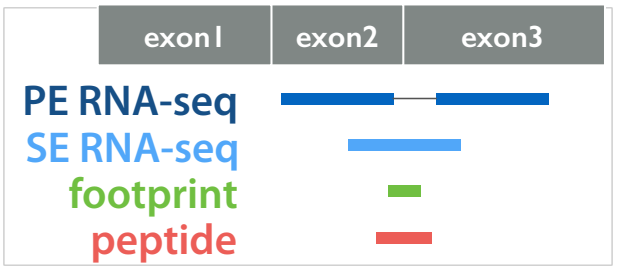
raRNA

ribosome footprints

MS/MS proteomics



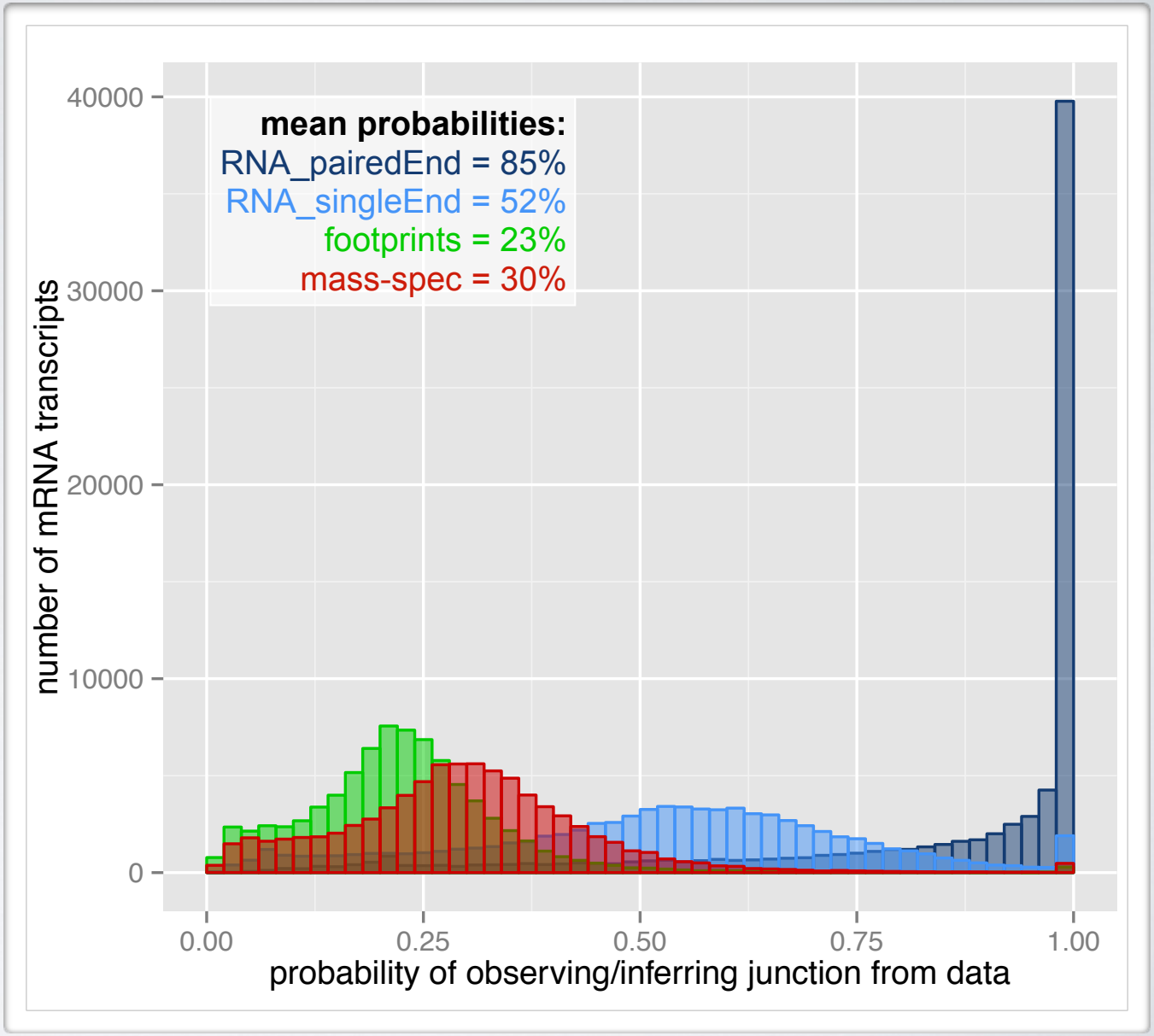
junction observability



RNA-seq will eventually produce single-molecule, whole transcript reads.

Not the case for footprints and not clear if MS will ever be able to profile intact protein

-> how can we integrate RNA, footprint, and MS data for mutual gain?



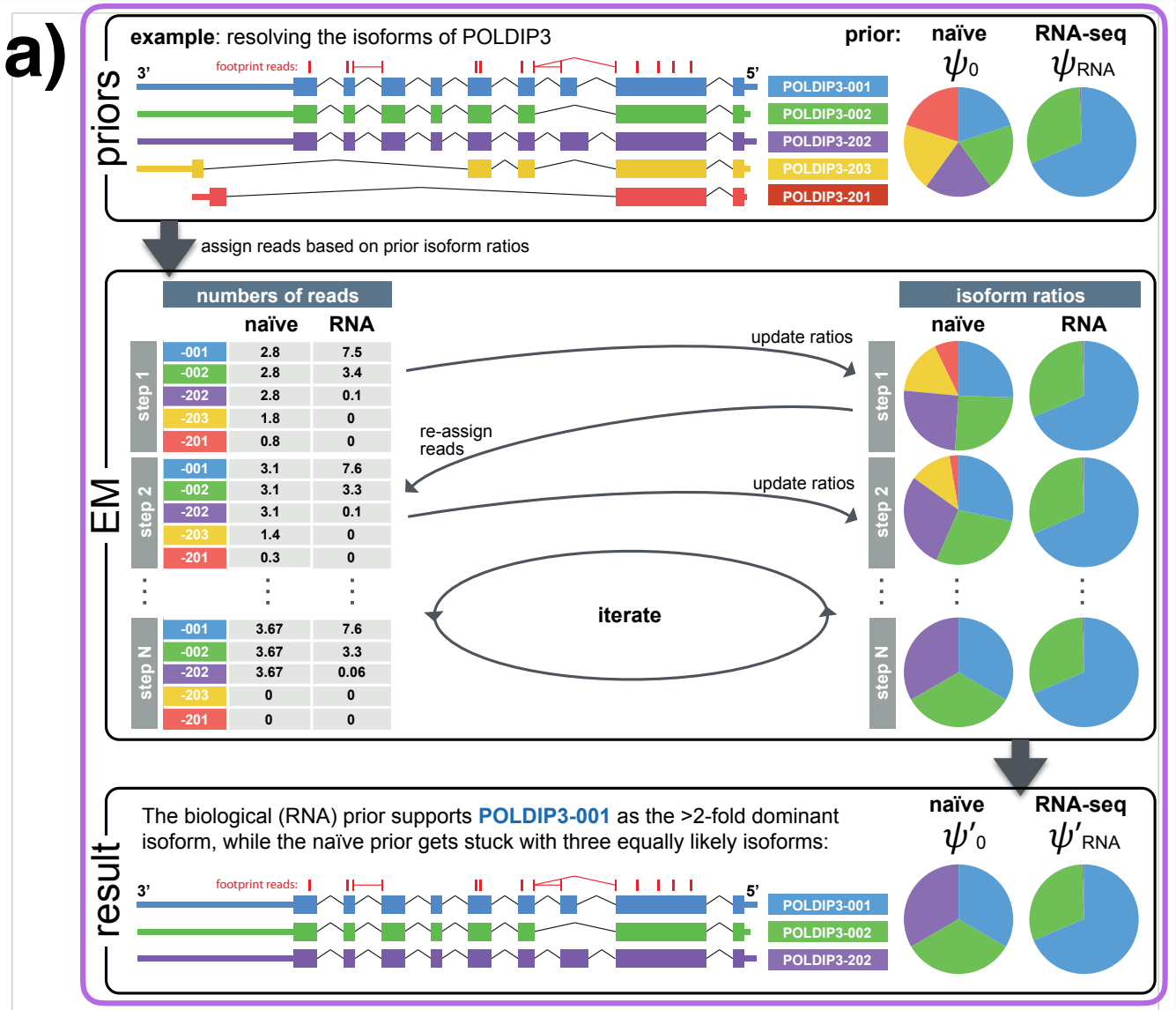
EM algorithm

- for each gene:
- maximise the likelihood of $1..N$ observed reads or peptides, R, given isoform abundances, ψ, of $1..K$ isoforms:

$$P(R_{1:N}|\psi) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(R_n|I_k) P(I_k|\psi)$$

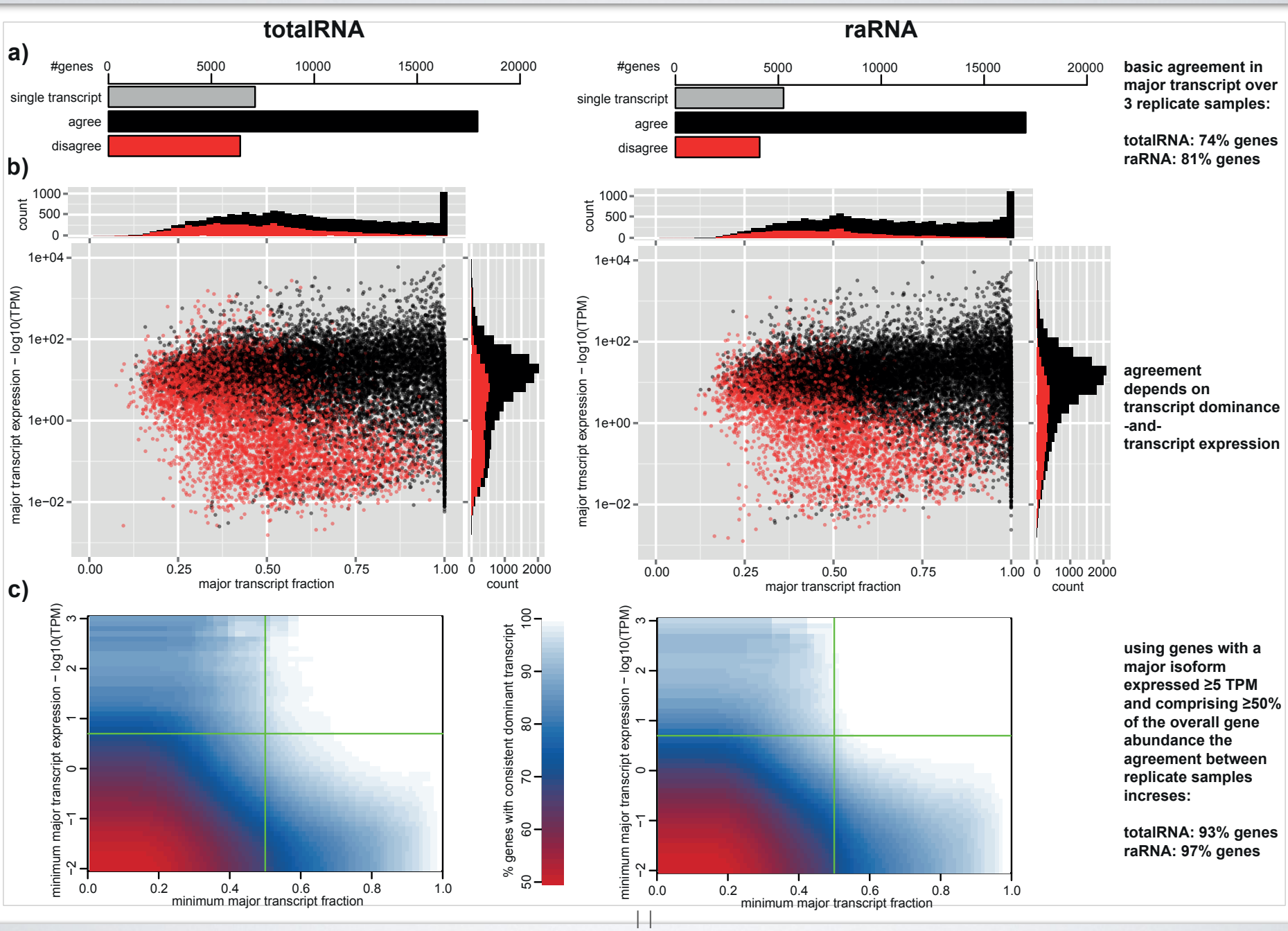
- for the naïve prior: $\psi_k = K^{-1}$
- probability that the j^{th} isoform will contribute a footprint or peptide is based on its CDS length, l, and abundance, ψ:

$$P(I_j|\psi) = \frac{\psi_j l_j}{\sum_{k=1}^{K} \psi_k l_k}$$

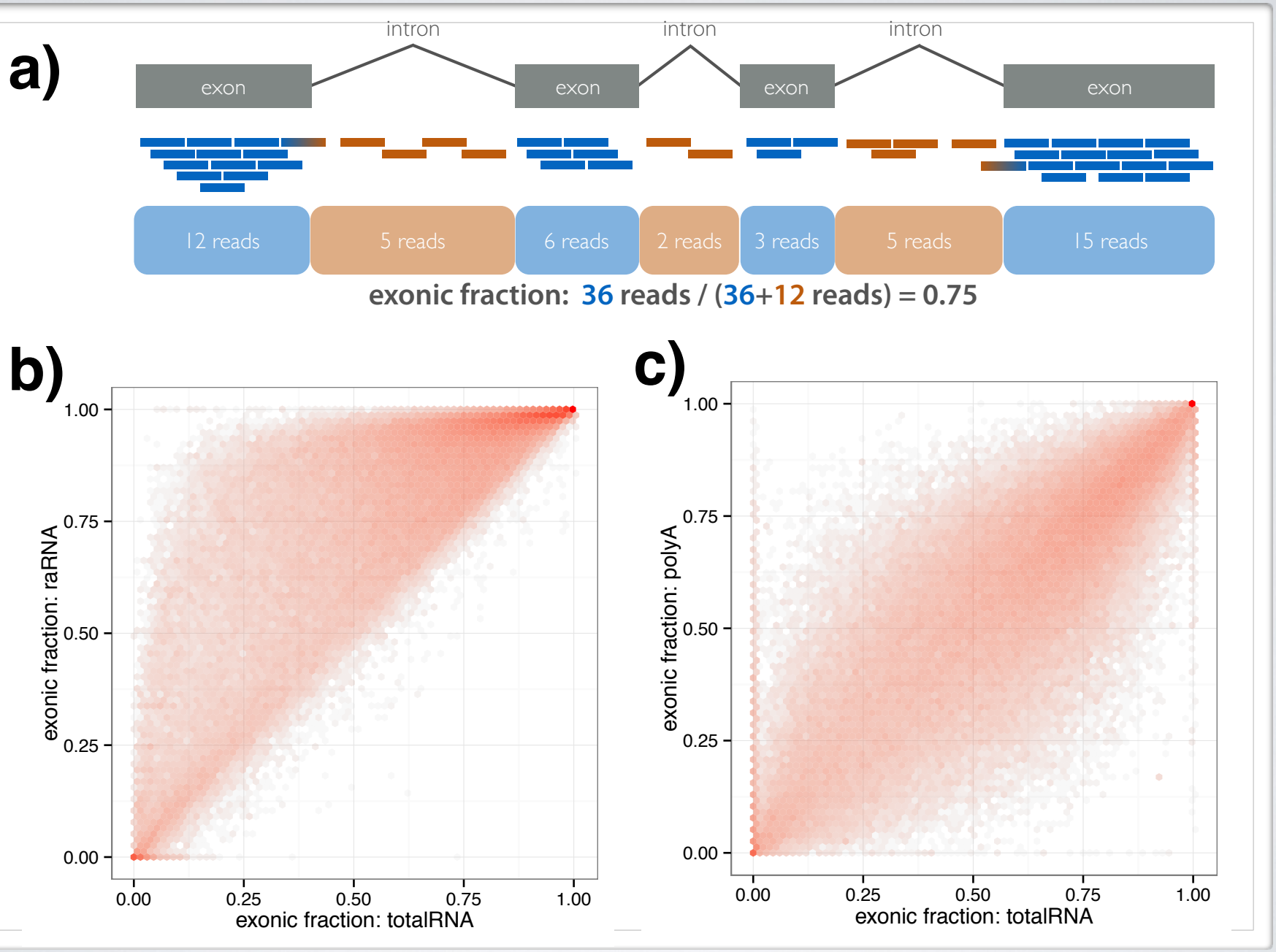


aside: higher quality data improves our chances for successful integration

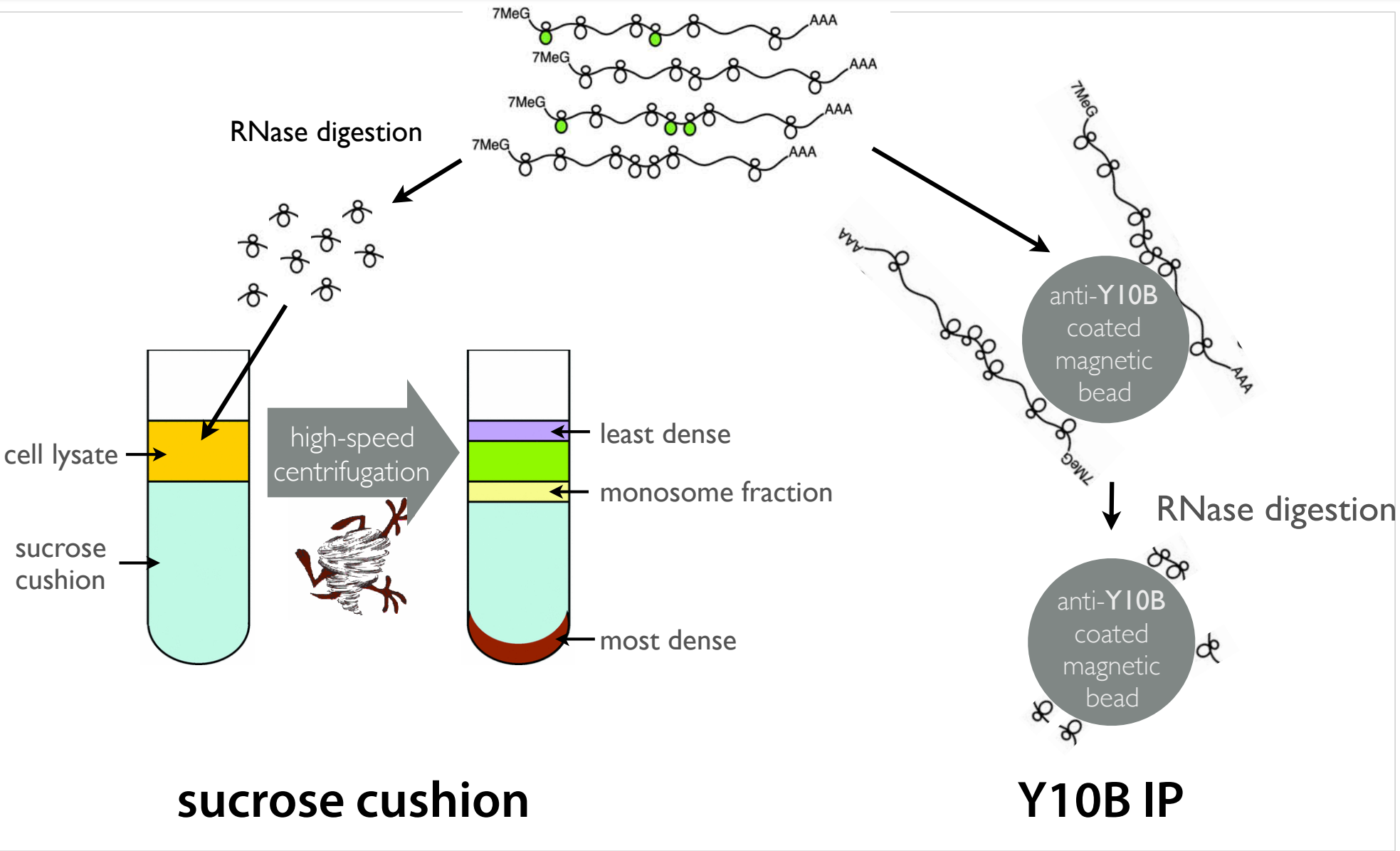
ribosome-affiliated RNA (raRNA) > totalRNA



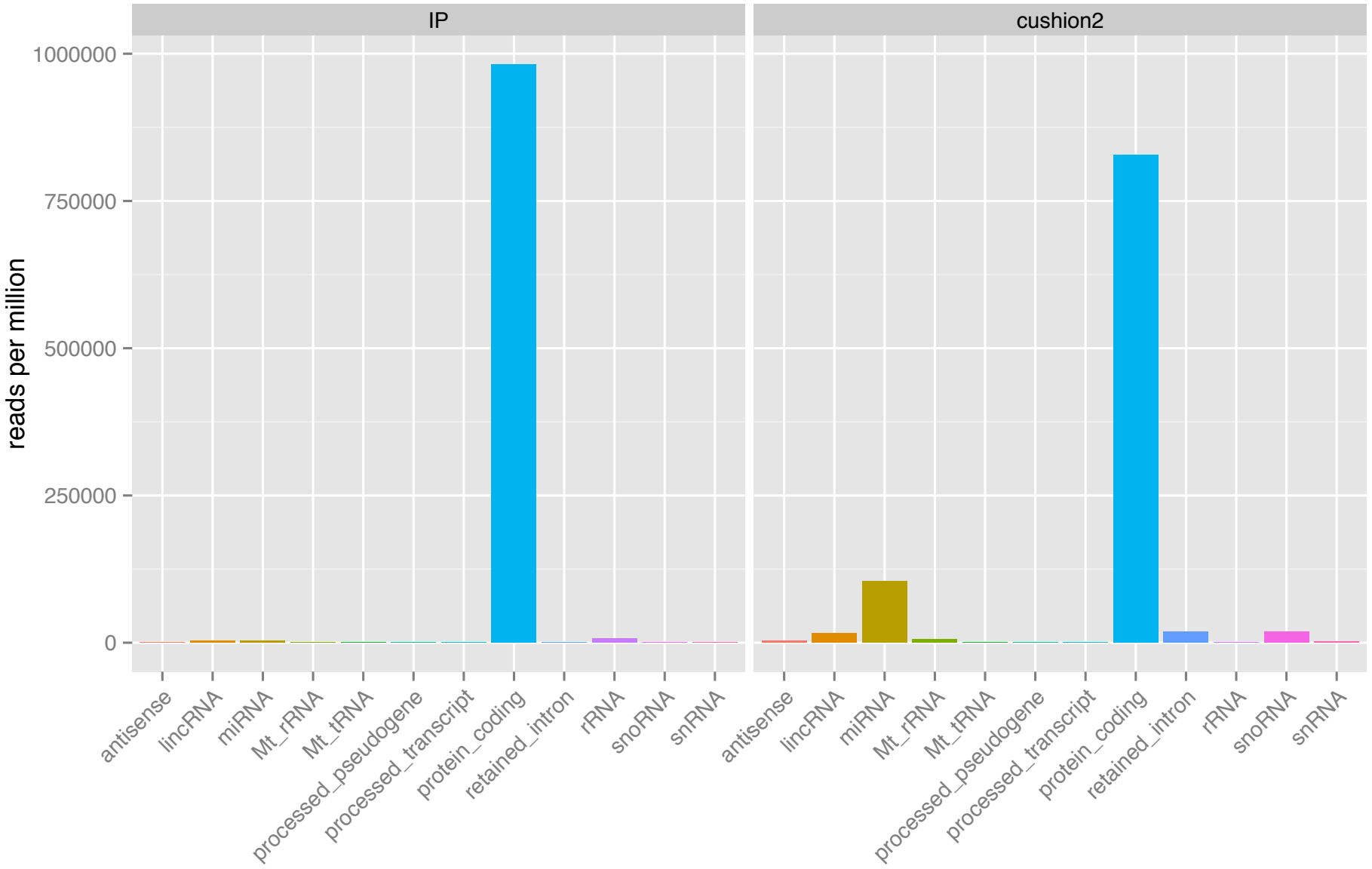
raRNA suffers less intronic contamination than totalRNA



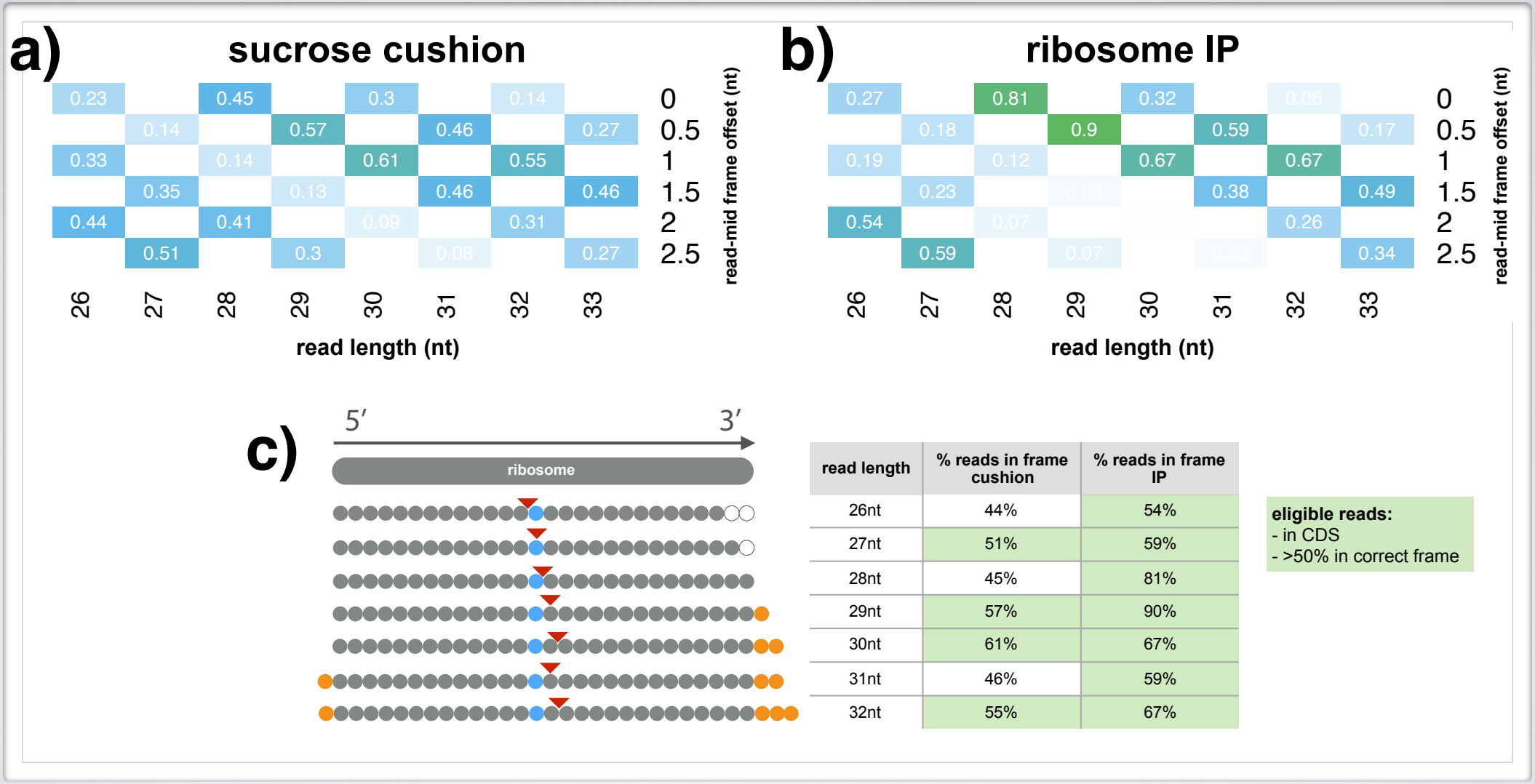
ribosome footprinting | wet-lab methods



>98% of IP footprints map to mRNA CDS'



IP footprints are more consistently in correct frame

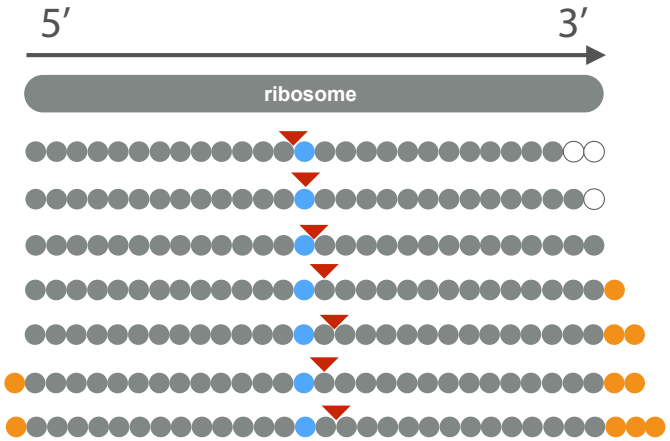


read length	% reads in frame cushion	% reads in frame IP
26nt	44%	54%
27nt	51%	59%
28nt	45%	81%
29nt	57%	90%
30nt	61%	67%
31nt	46%	59%
32nt	55%	67%

eligible reads:
- in CDS
- >50% in correct frame

IP footprints better able to predict correct frame

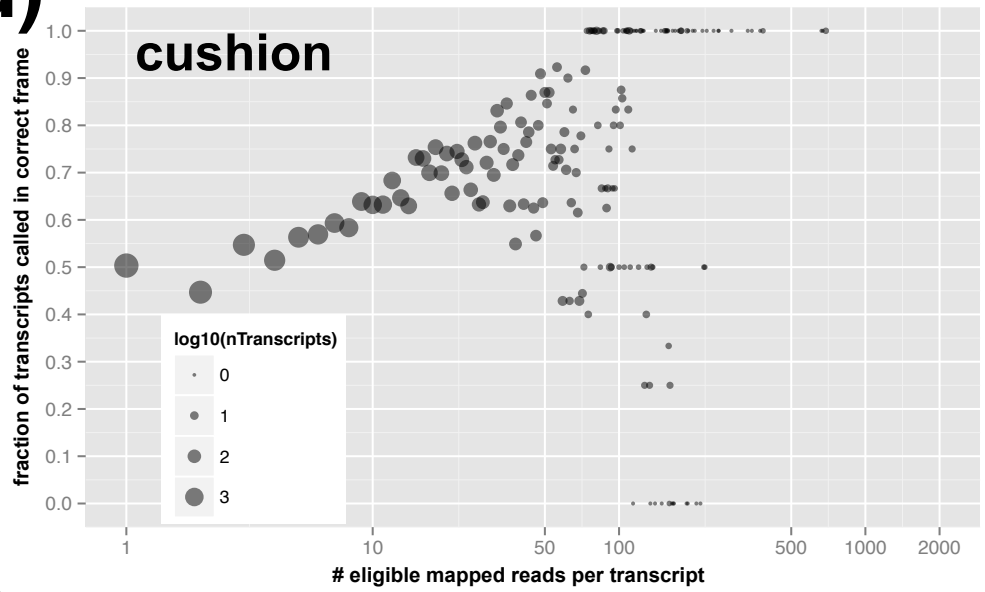
c)



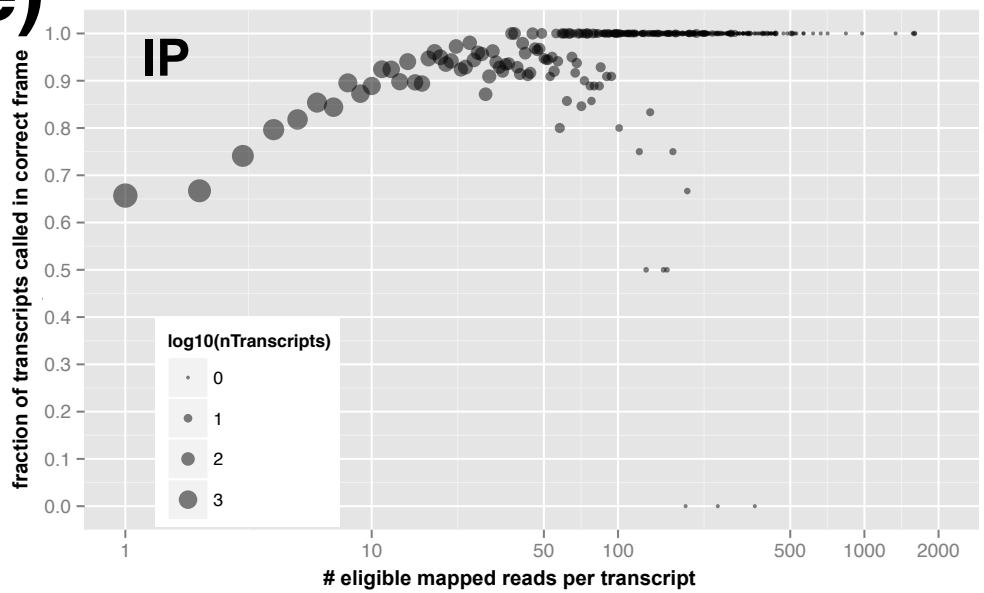
read length	% reads in frame cushion	% reads in frame IP
26nt	44%	54%
27nt	51%	59%
28nt	45%	81%
29nt	57%	90%
30nt	61%	67%
31nt	46%	59%
32nt	55%	67%

eligible reads:
- in CDS
- >50% in correct frame

d)



e)



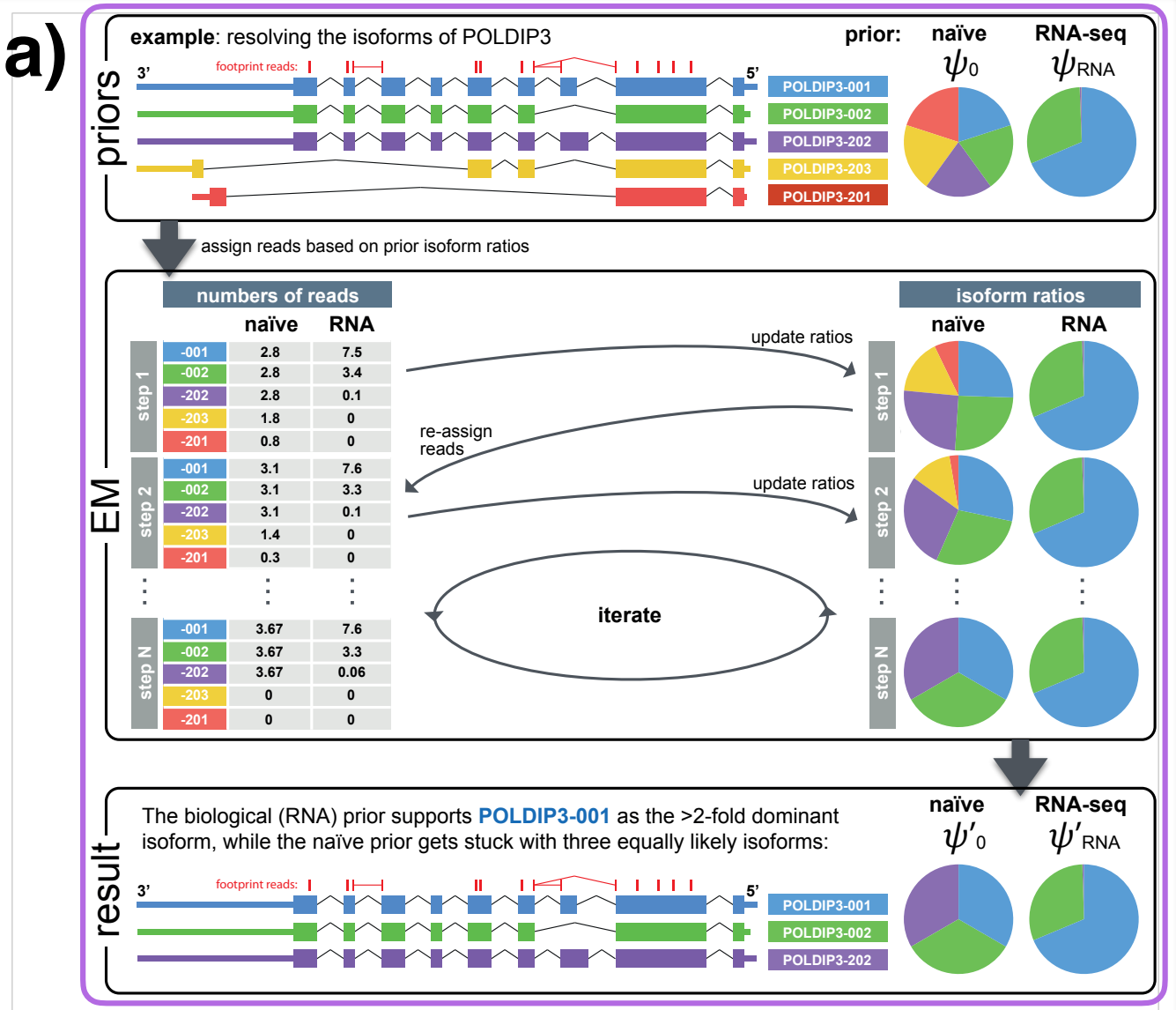
EM algorithm

- for each gene:
- maximise the likelihood of $1..N$ observed reads or peptides, R, given isoform abundances, ψ, of $1..K$ isoforms:

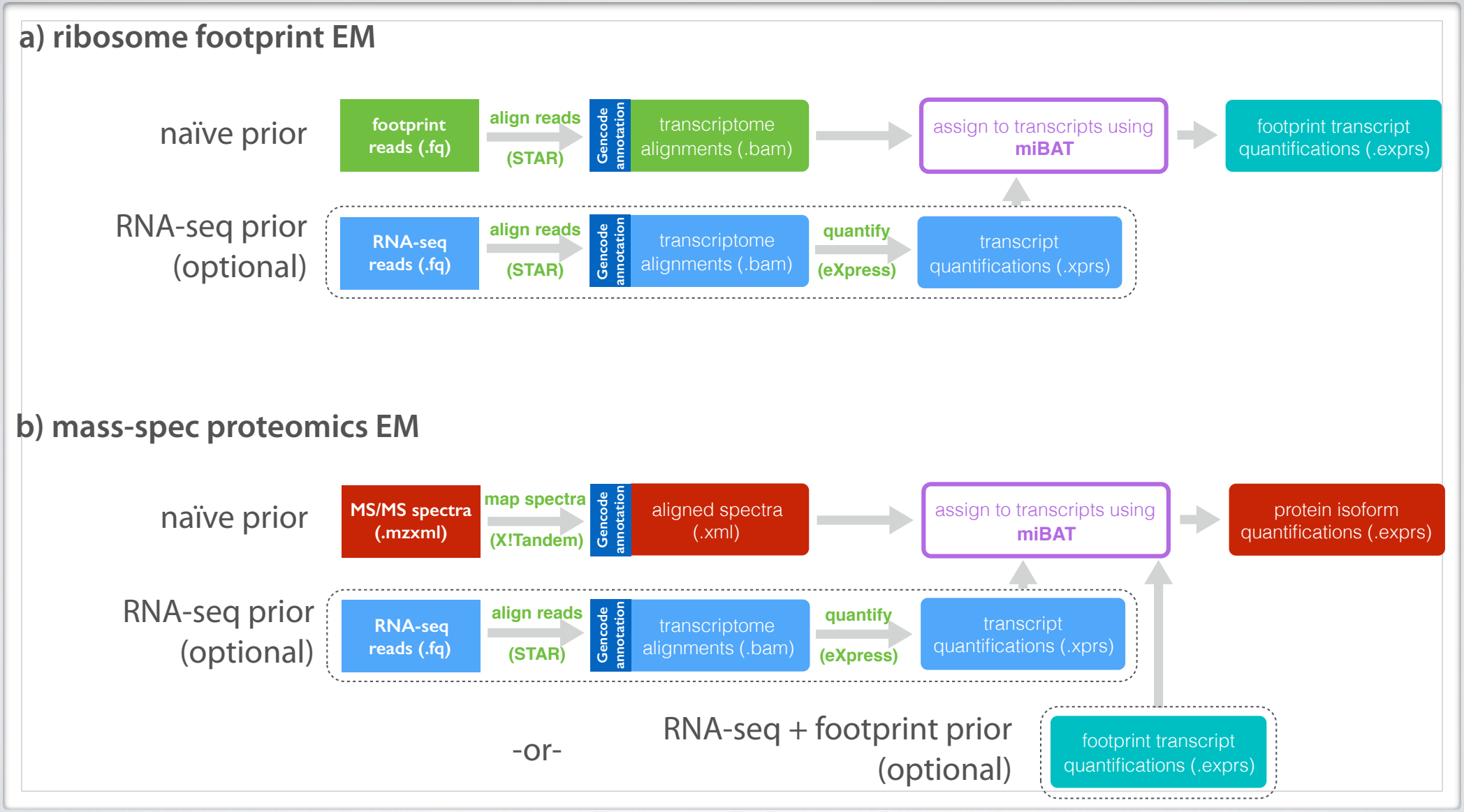
$$P(R_{1:N}|\psi) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(R_n|I_k) P(I_k|\psi)$$

- for the naïve prior: $\psi_k = K^{-1}$
- probability that the j^{th} isoform will contribute a footprint or peptide is based on its CDS length, l, and abundance, ψ:

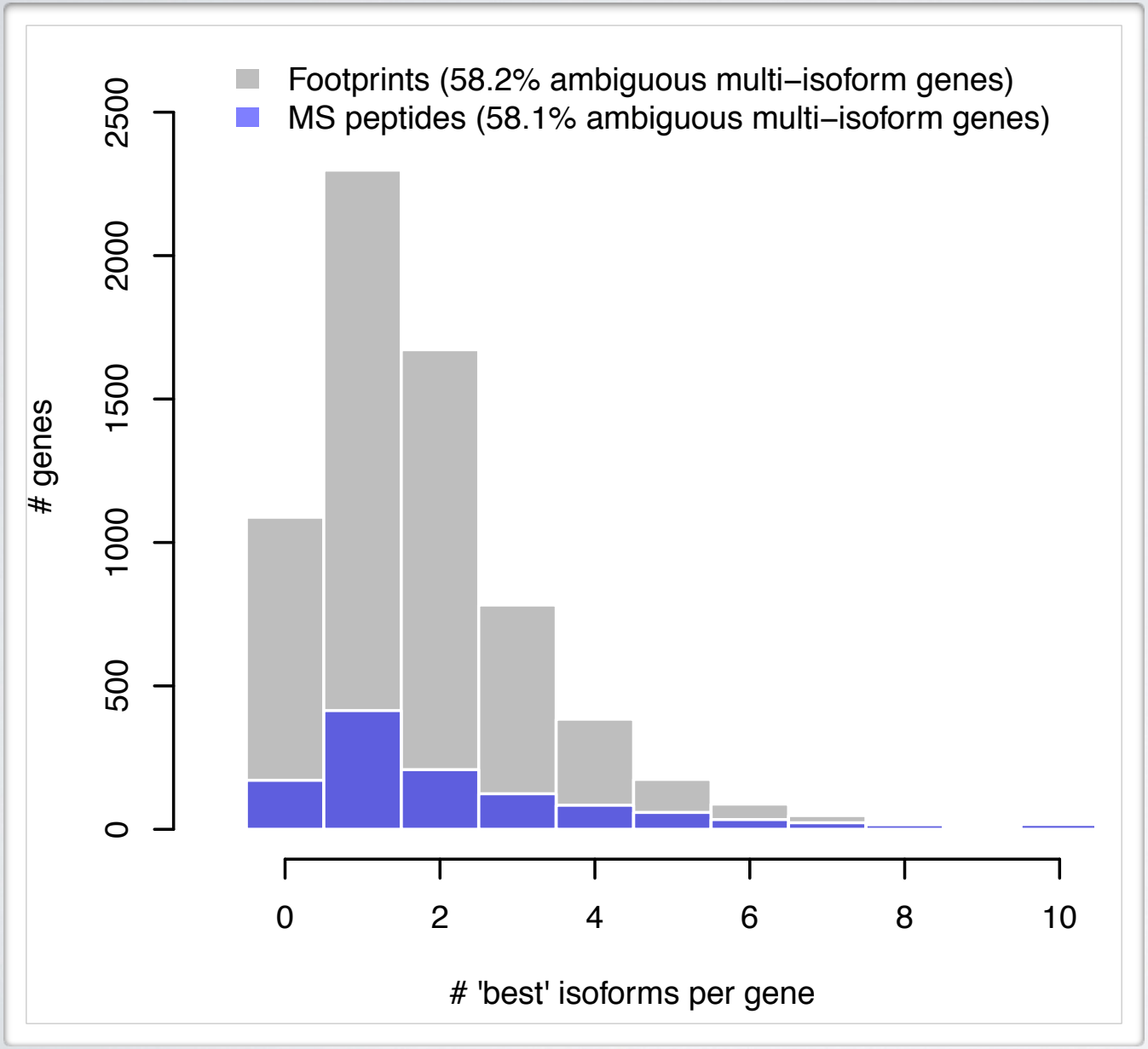
$$P(I_j|\psi) = \frac{\psi_j l_j}{\sum_{k=1}^{K} \psi_k l_k}$$



implementation: 'miBAT'

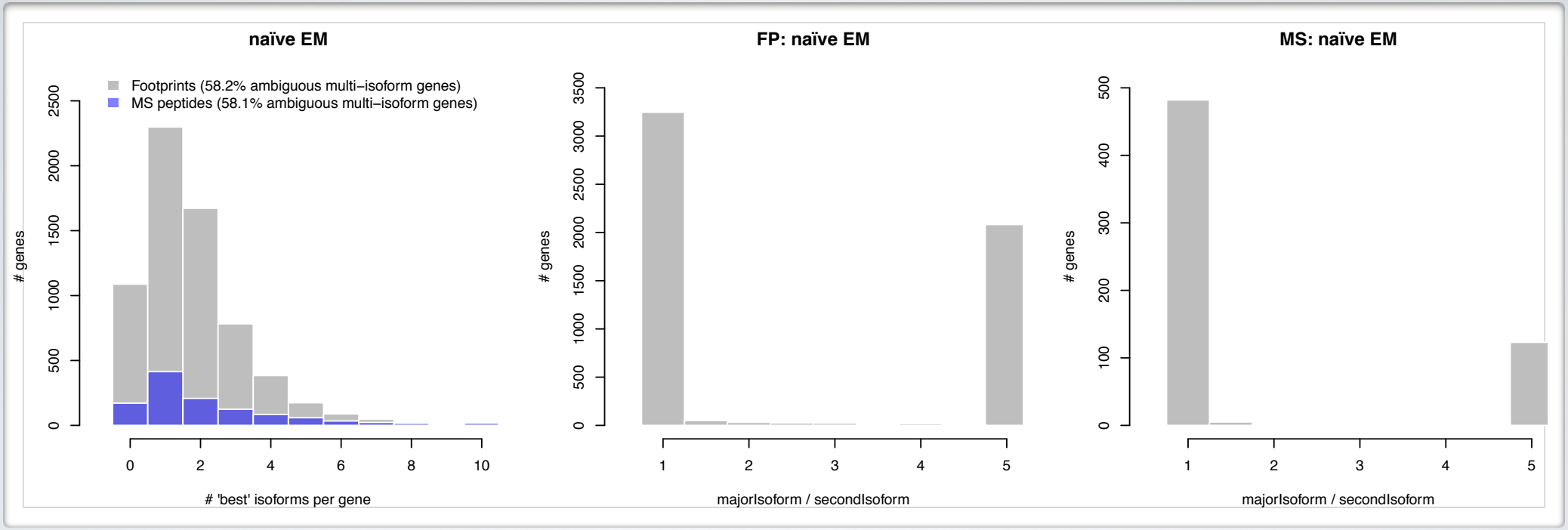


naïve prior



ribosome footprints or MS/MS peptides alone cannot distinguish isoforms for the majority of genes

naïve prior

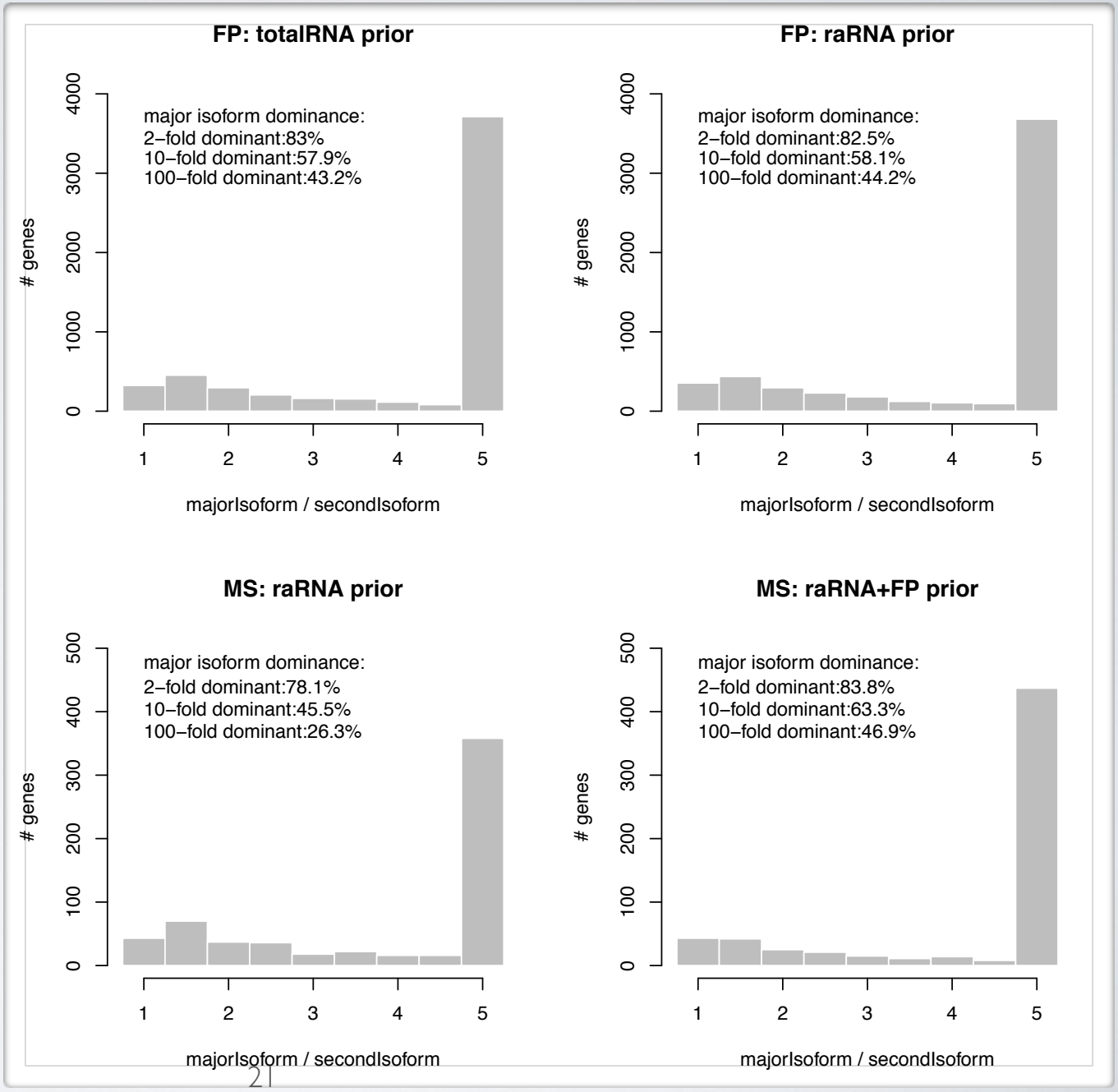


ratio of 'best' to 'second-best' isoform reveals all-or-nothing style behaviour of the naïve EM

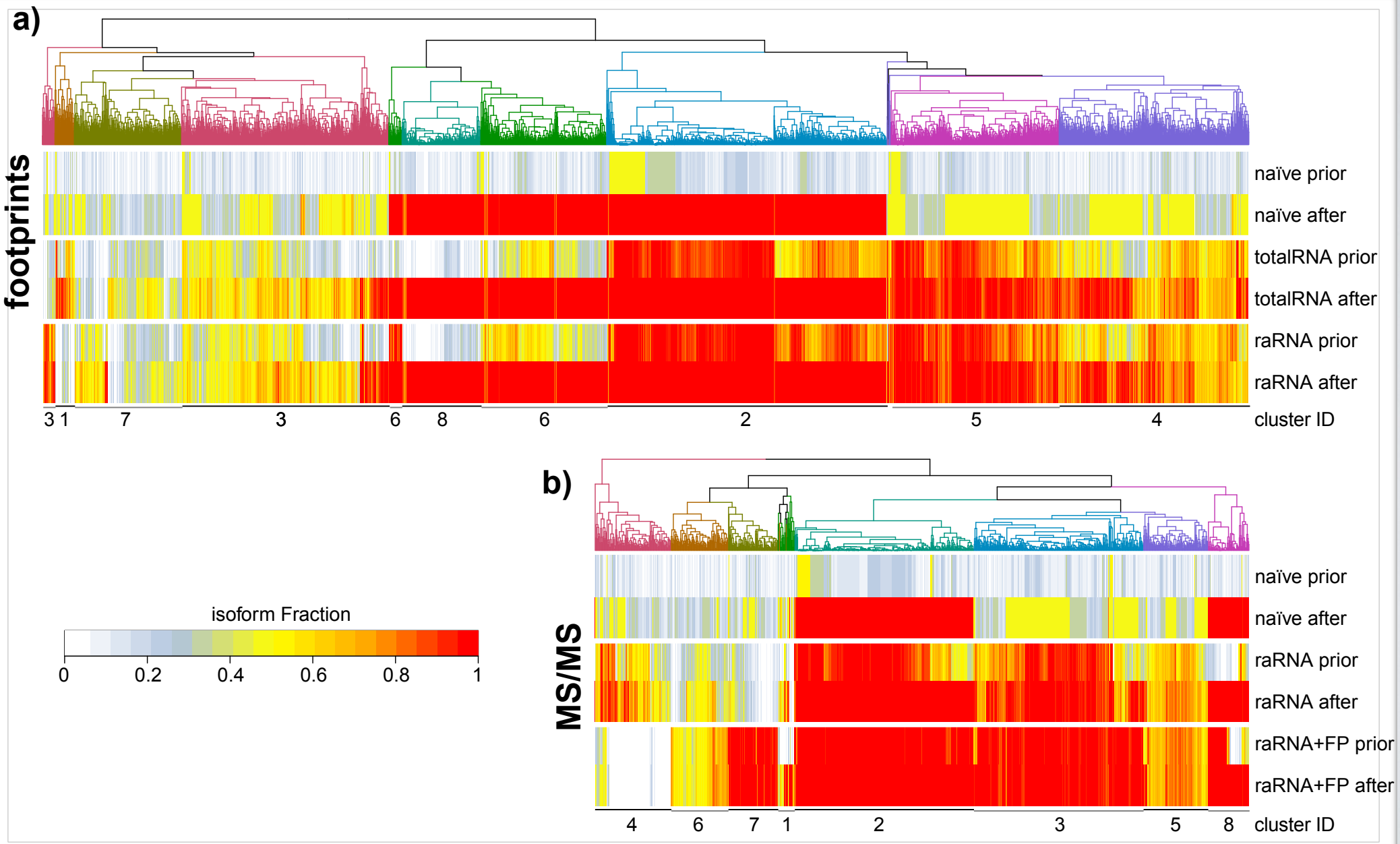
informative prior

ribosome footprints benefit from totalRNA or raRNA-seq prior

MS/MS peptides also benefit from raRNA-seq or footprint prior

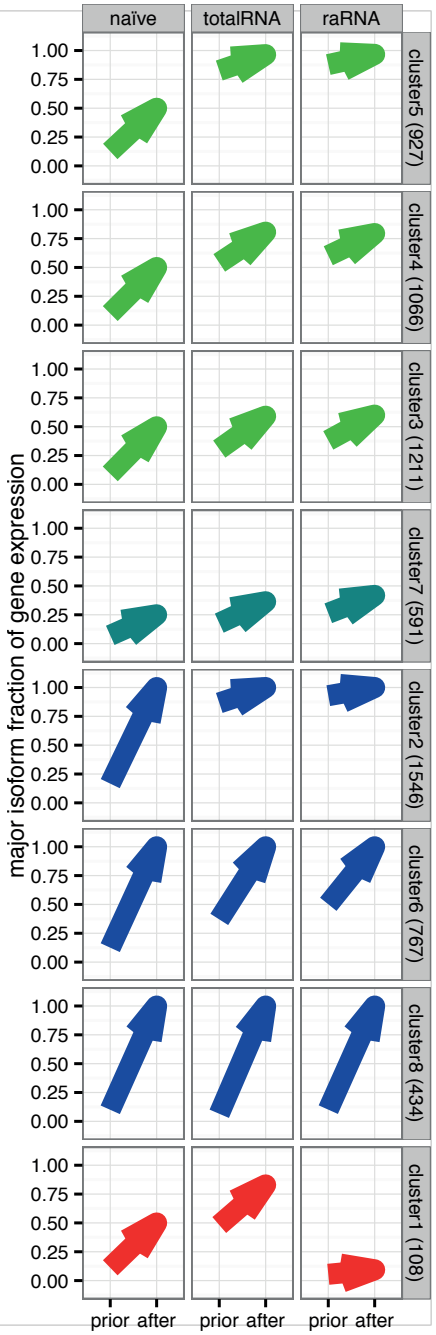


global clustering



cluster profiles

ribosome footprints

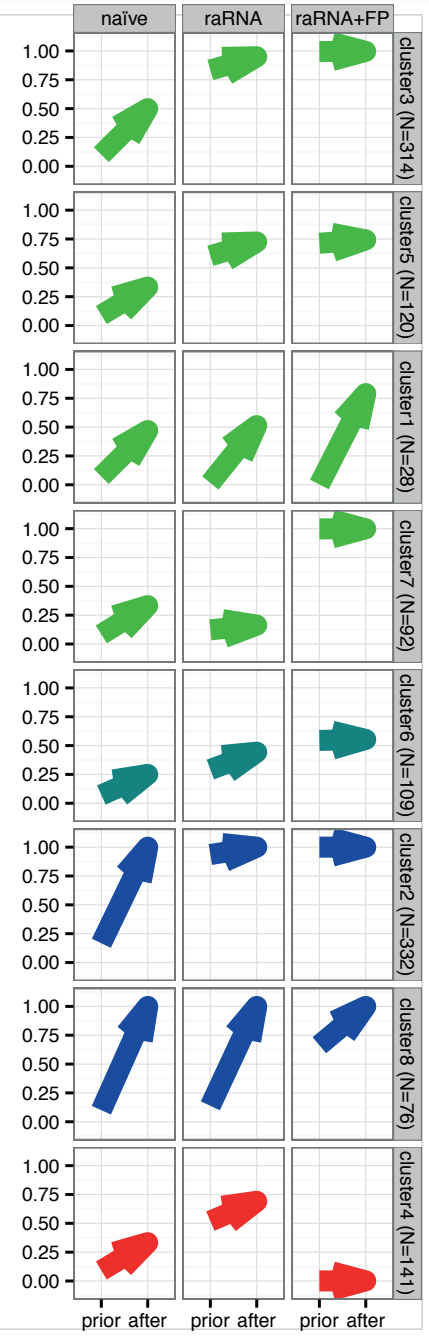


informative prior helps determine isoform

footprints / peptides drive the isoform ID

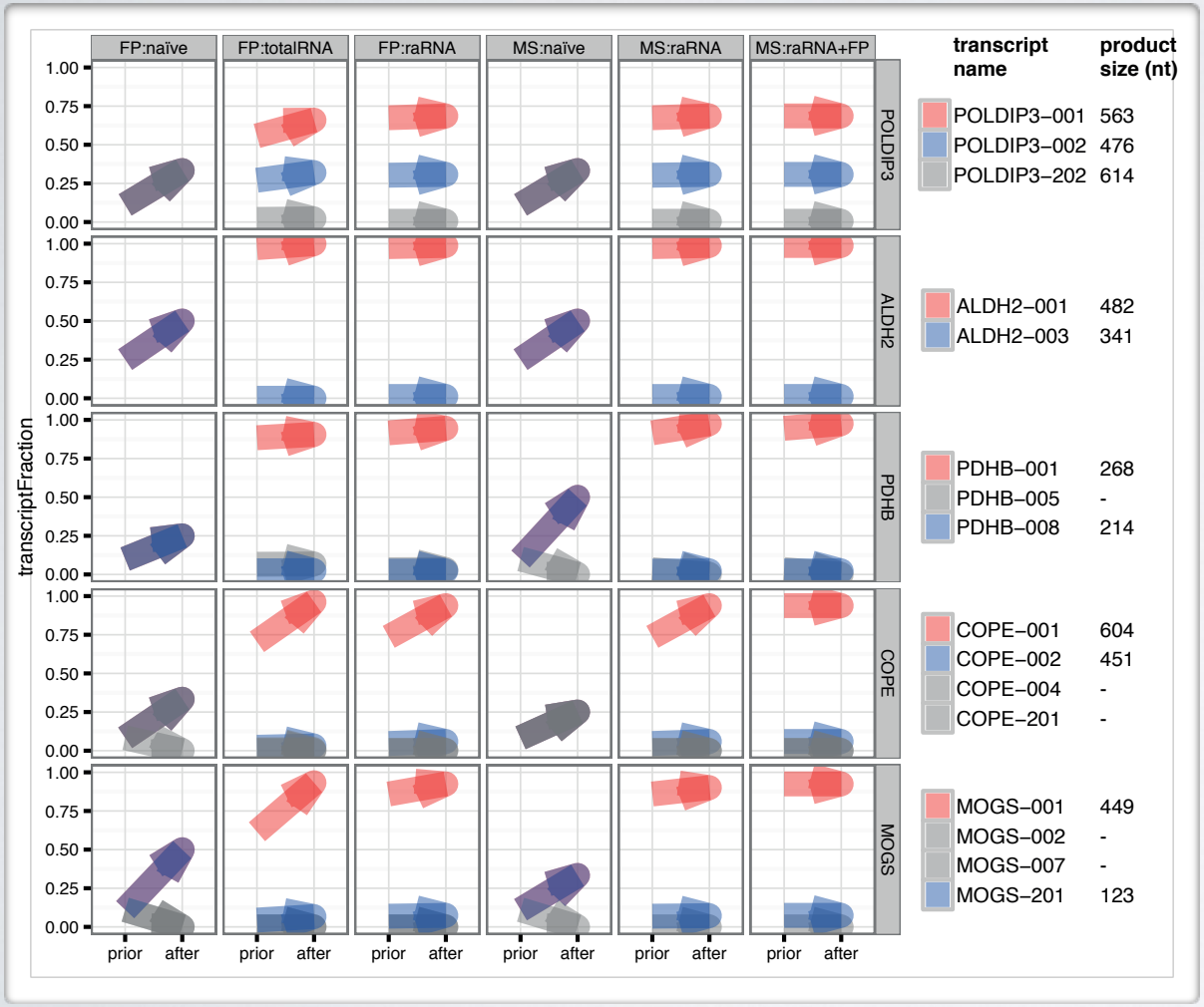
isoform choice changes with different prior

MS/MS peptides

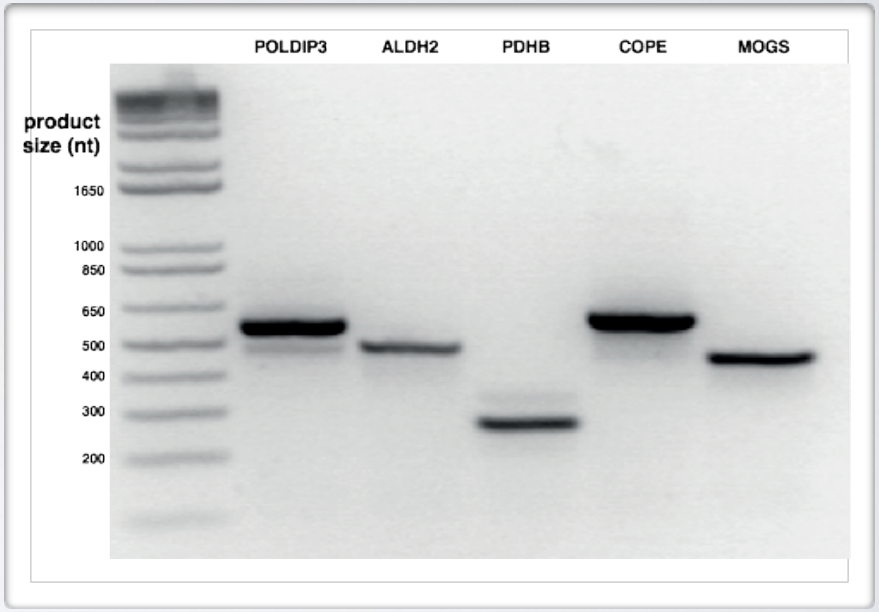


validation

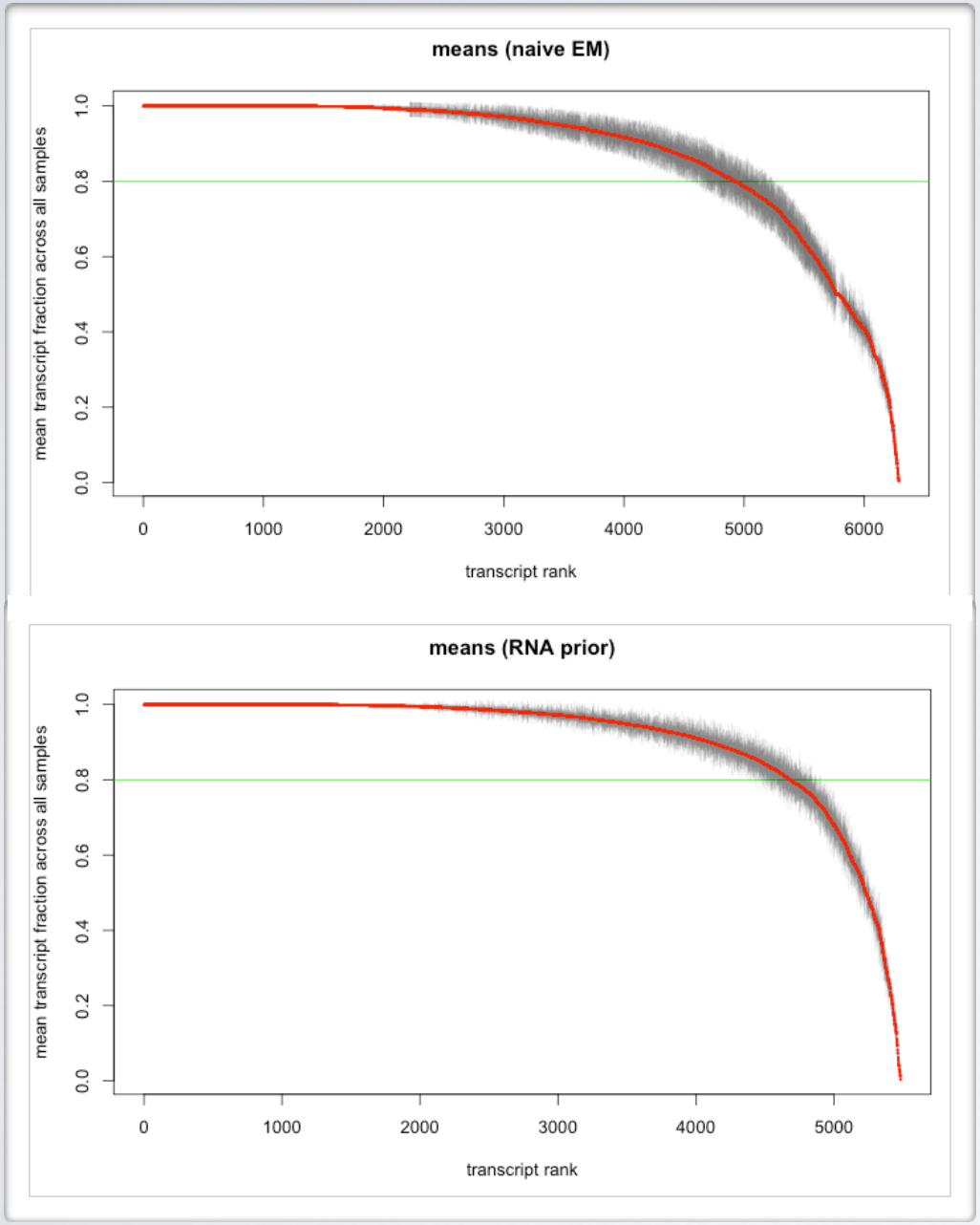
6 specific examples of genes/isoforms resolved by RNA-seq:

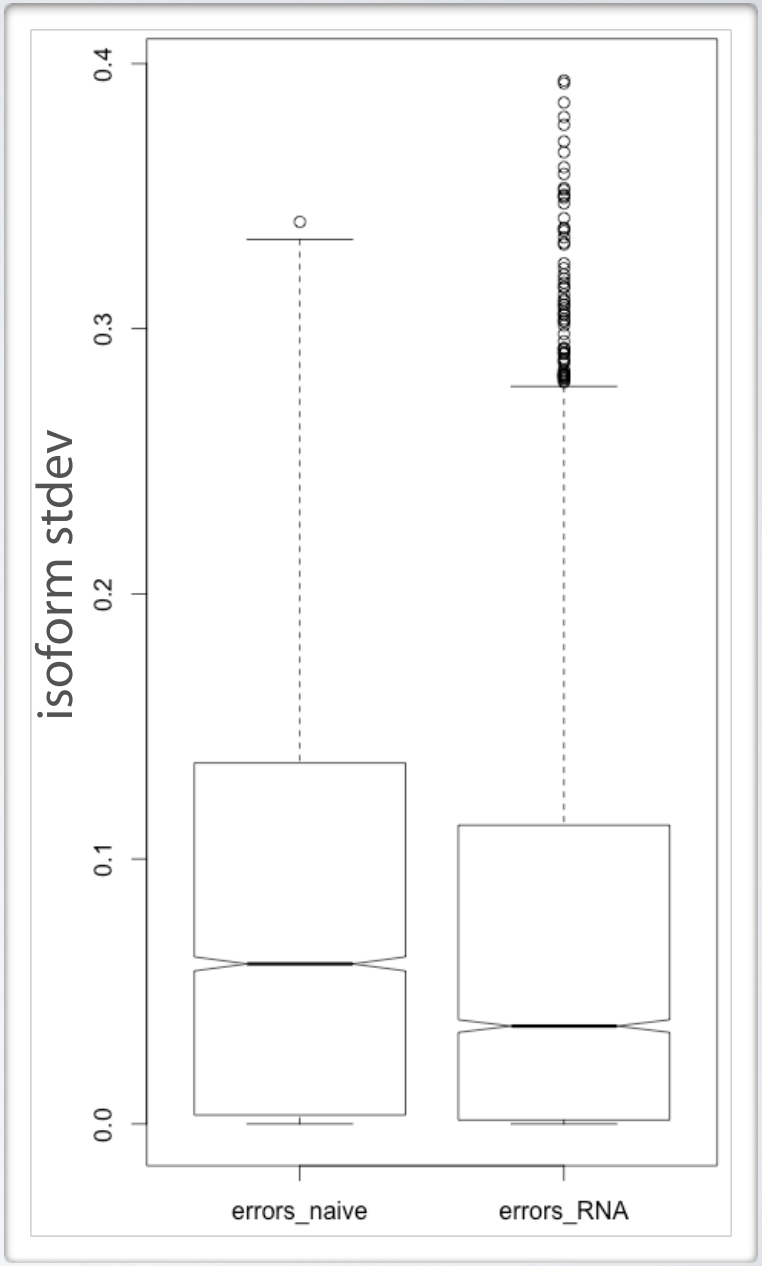


PCR:



54 1000genomes RNA-seq & footprint samples



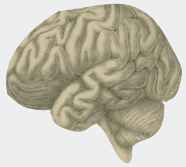


applications

1 - proteomic profiling of human brain (psychENCODE)
2 - cell-type specific footprinting and proteomics (CEBRA)

1 - proteomic profiling of human brain (psychENCODE)

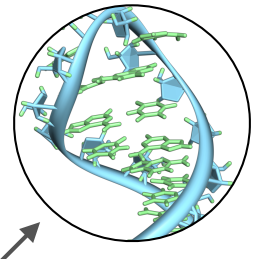
a)



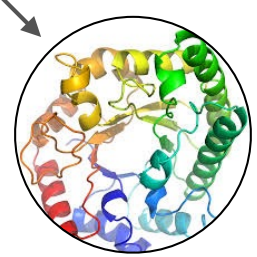
x5 adult human brains

x7 brain regions

DFC	dorsolateral prefrontal cortex
V1C	primary visual cortex
HIP	hippocampus
STR	striatum
MD	thalamus (mediodorsal nucleus)
AMY	amygdala
CBC	cerebellum



per-sample quantification
(all regions, all brains)
poly-A RNA-seq

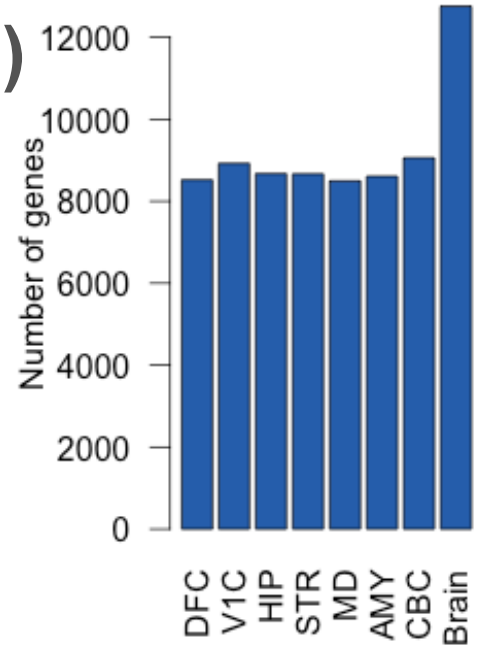


protein

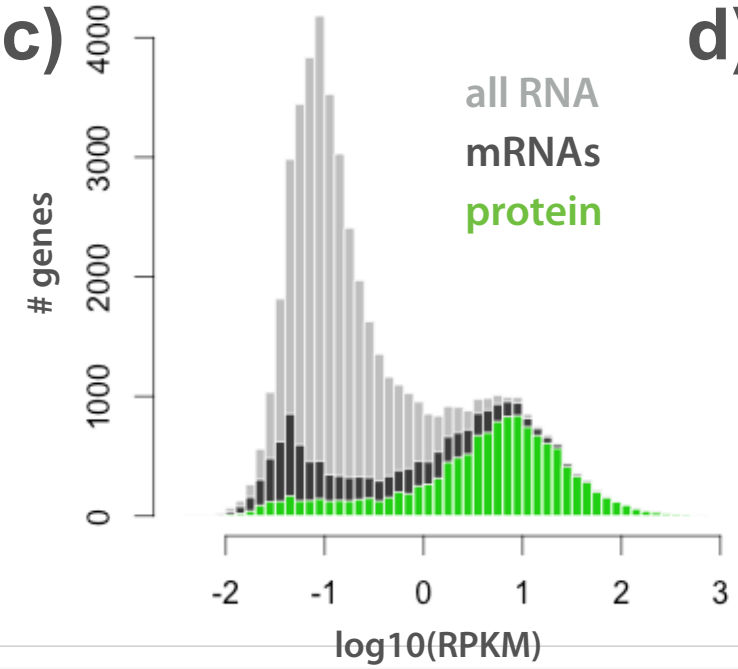
per-region survey
(all regions, pooled brains)
high-pH reversed-phase chromatography
fractionated LC-MS/MS

per-sample quantification
(all regions, all brains)
single-run LC-MS/MS

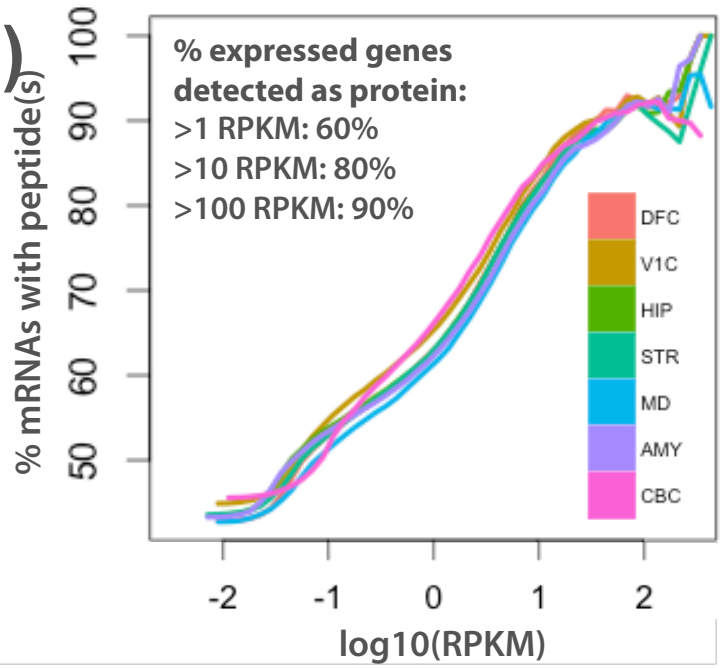
b)



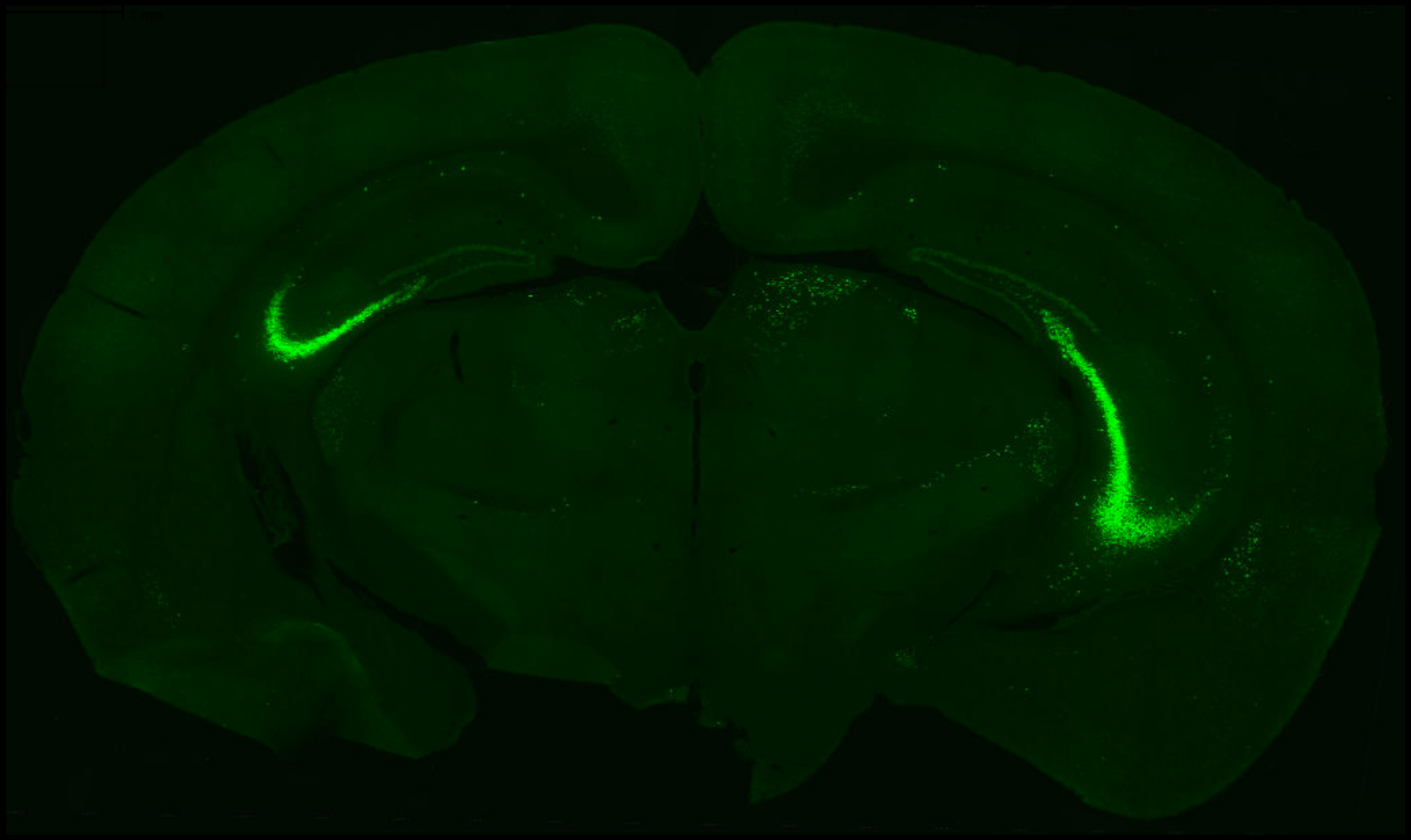
c)



d)

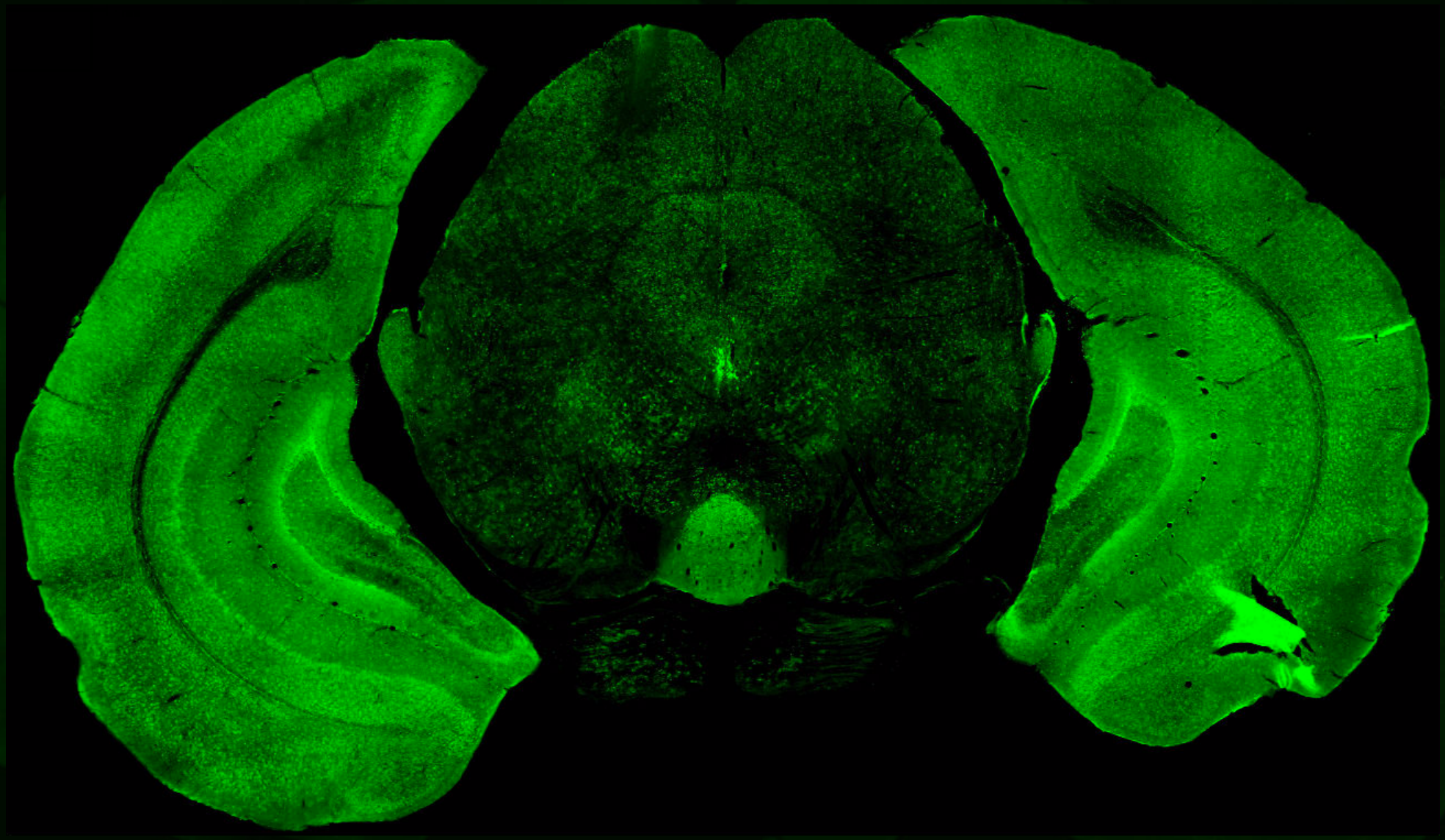


2 - cell-type specific footprinting and proteomics (CEBRA)



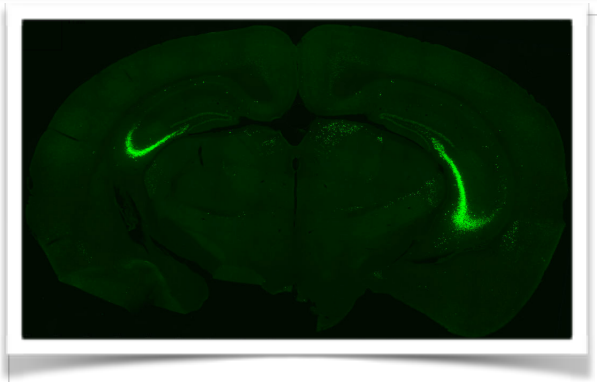
Htr6

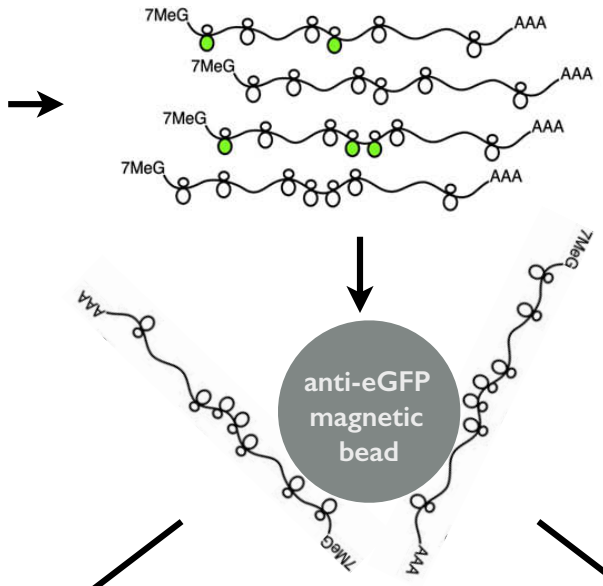
2 - cell-type specific footprinting and proteomics (CEBRA)

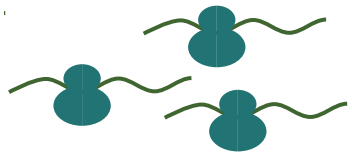


Lhx5

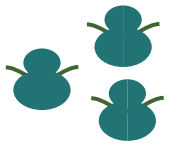
2 - cell-type specific footprinting and proteomics (CEBRA)







raRNA
RNA-seq of ribosome-affiliated RNA captured by RPL10a-eGFP IP



footprints
RNA-seq of ribosome-bound RNA fragments

MS proteomics
single-injection & fractionated mass-spectrometry

summary

- **immunoprecipitation of ribosomes** provides very clean [ra]RNA-seq and ribosome footprint data

- developed flexible framework (**miBAT**) to not only assign footprints/peptides to isoforms, but also to allow these data to update isoform abundances estimated by RNA-seq

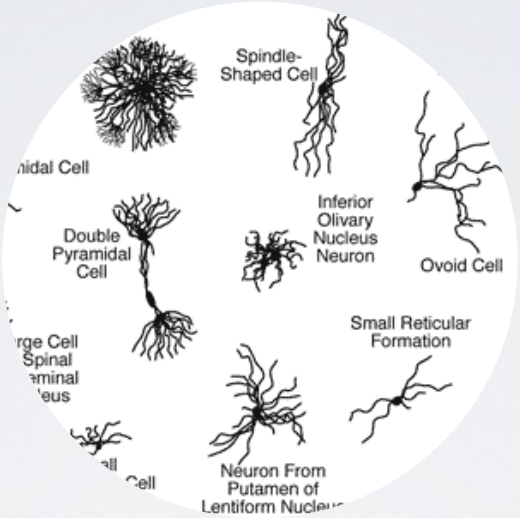
- **application to psychENCODE** project (integrating RNA-seq and MS/MS proteomics) to provide isoform-resolution map of the human brain

- **application to specific cell-types** to monitor translation (using RNA-seq, footprints, & MS/MS proteomics) in response to acute cocaine administration

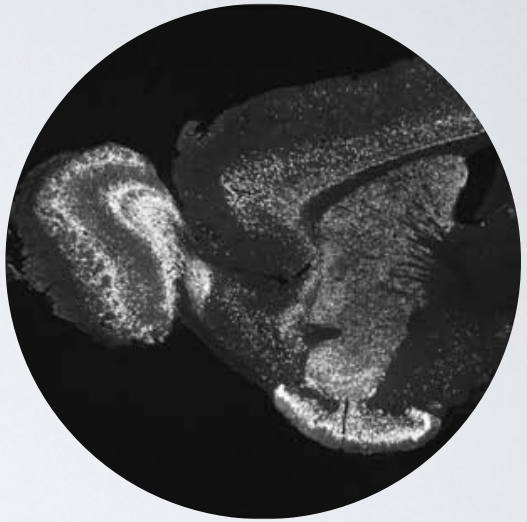
cell-type specific profiling of the mouse CNS

what defines a neural cell-type?

location



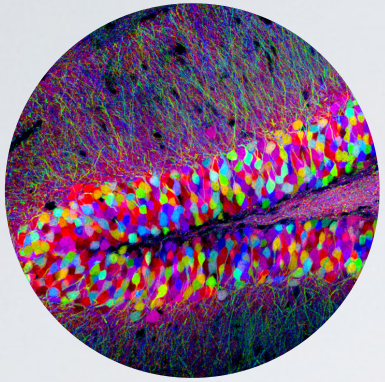
shape
&
connectivity



genes
&
neurotransmitters

use combinations of these properties to select specific cell-types

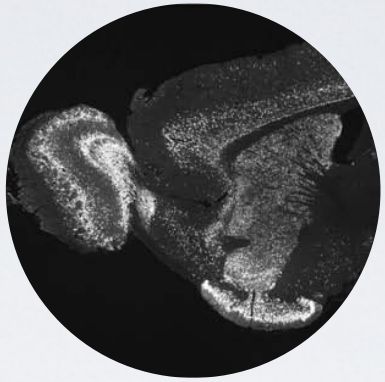
the case for cell-types



mixed cells

dissection

- ✔ easy to obtain
- ✔ plenty of material
- ✘ ++ white matter
- ✘ not sensitive



single cell-type

bacTRAP / FACS

- ✔ easy to obtain…
- ✘ …only in model organism
- ✔ plenty of material
- ✔ promoter-specific
- ✔ sensitive

single cell

LCM / FACS

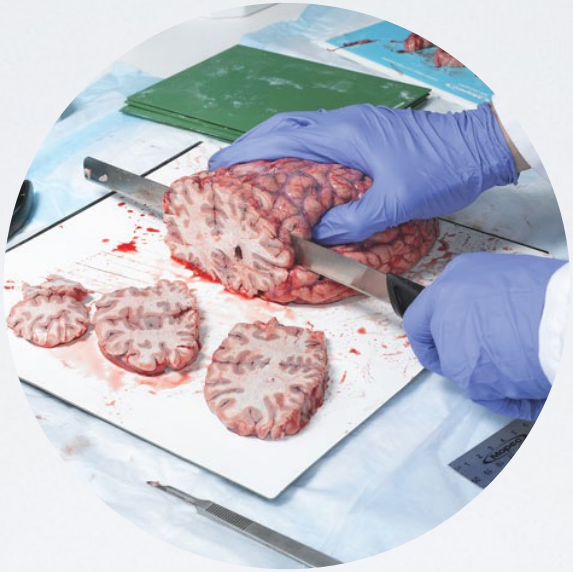
- ✘ difficult to obtain
- ✘ little material
- ✘ lose processes
- ✔ sensitive
- ✘ ChIP & MS/MS hard

specificity

limitations of human transcriptomics

major problems associated with profiling RNA abundance in the human brain

death
&
PMI



sample quality
&
biopsy precision

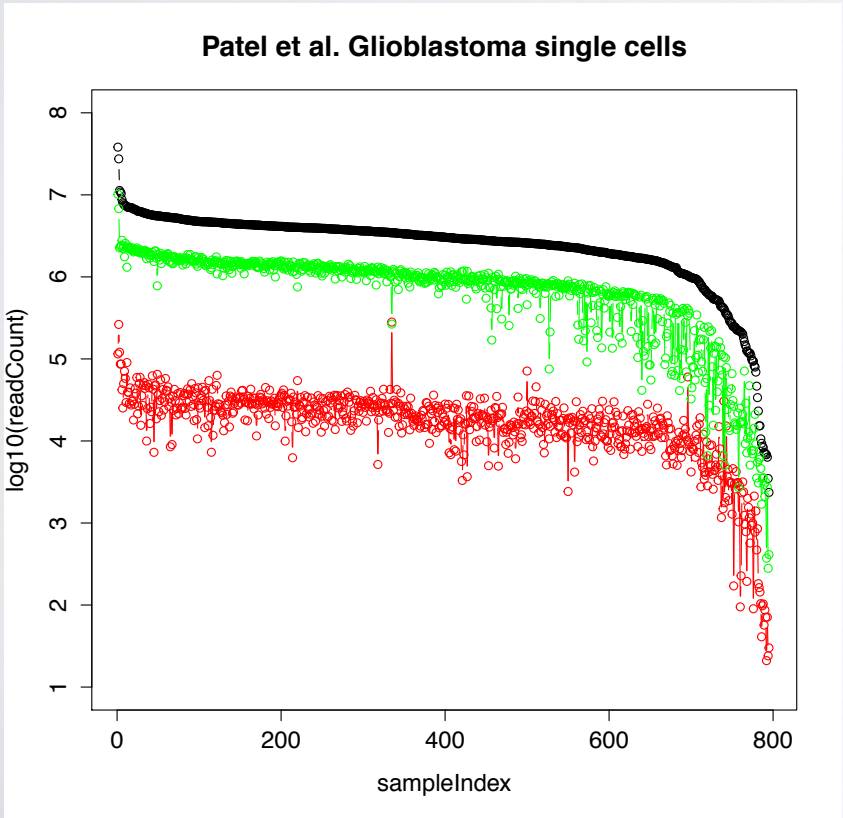
cell-types
&
transgenics

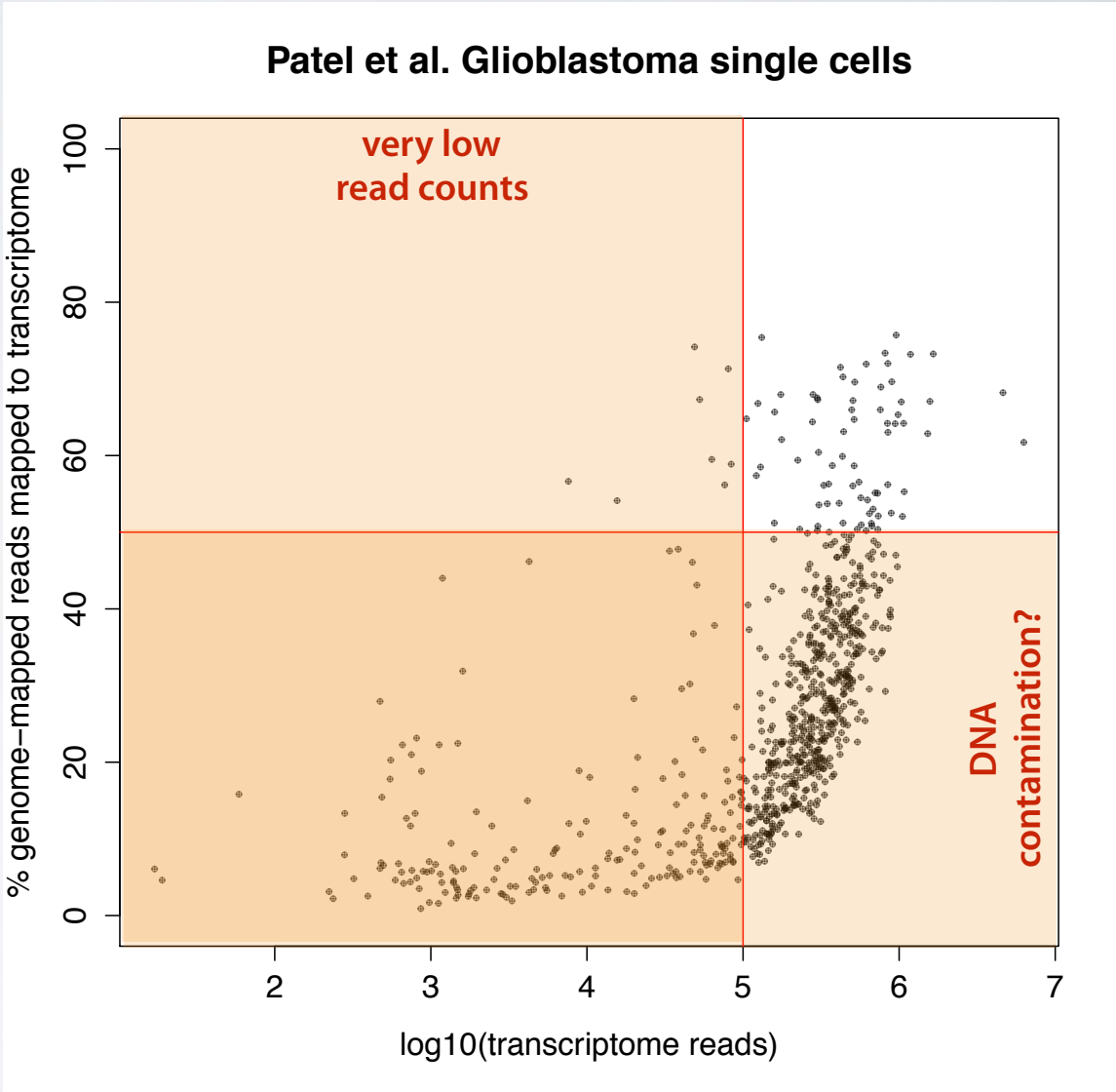
single-cell RNA-seq | bad [human] experiment

CANCER GENOMICS

Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma

Anoop P. Patel,[*1,2,3,4] Itay Tirosh,[*3] John J. Trombetta,[3] Alex K. Shalek,[3] Shawn M. Gillespie,[2,3,4] Hiroaki Wakimoto,[1] Daniel P. Cahill,[1] Brian V. Nahed,[1] William T. Curry,[1] Robert L. Martuza,[1] David N. Louis,[2] Orit Rozenblatt-Rosen,[3] Mario L. Suvà,[2,3]†‡ Aviv Regev,[3,4,5]†‡ Bradley E. Bernstein[2,3,4]†‡





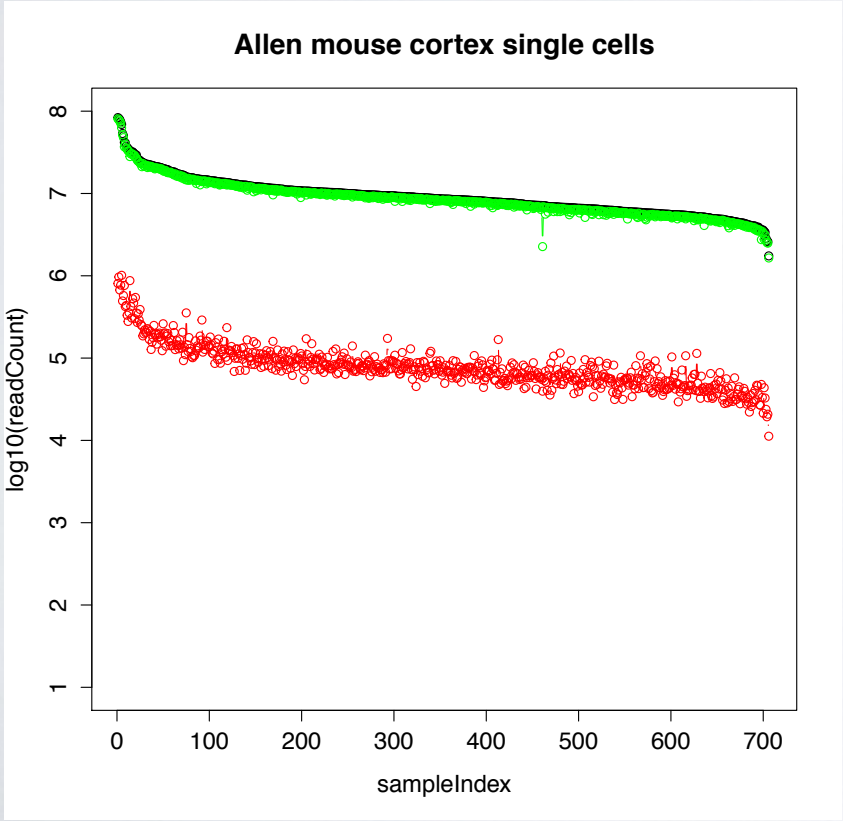
single-cell RNA-seq | good [mouse] experiment

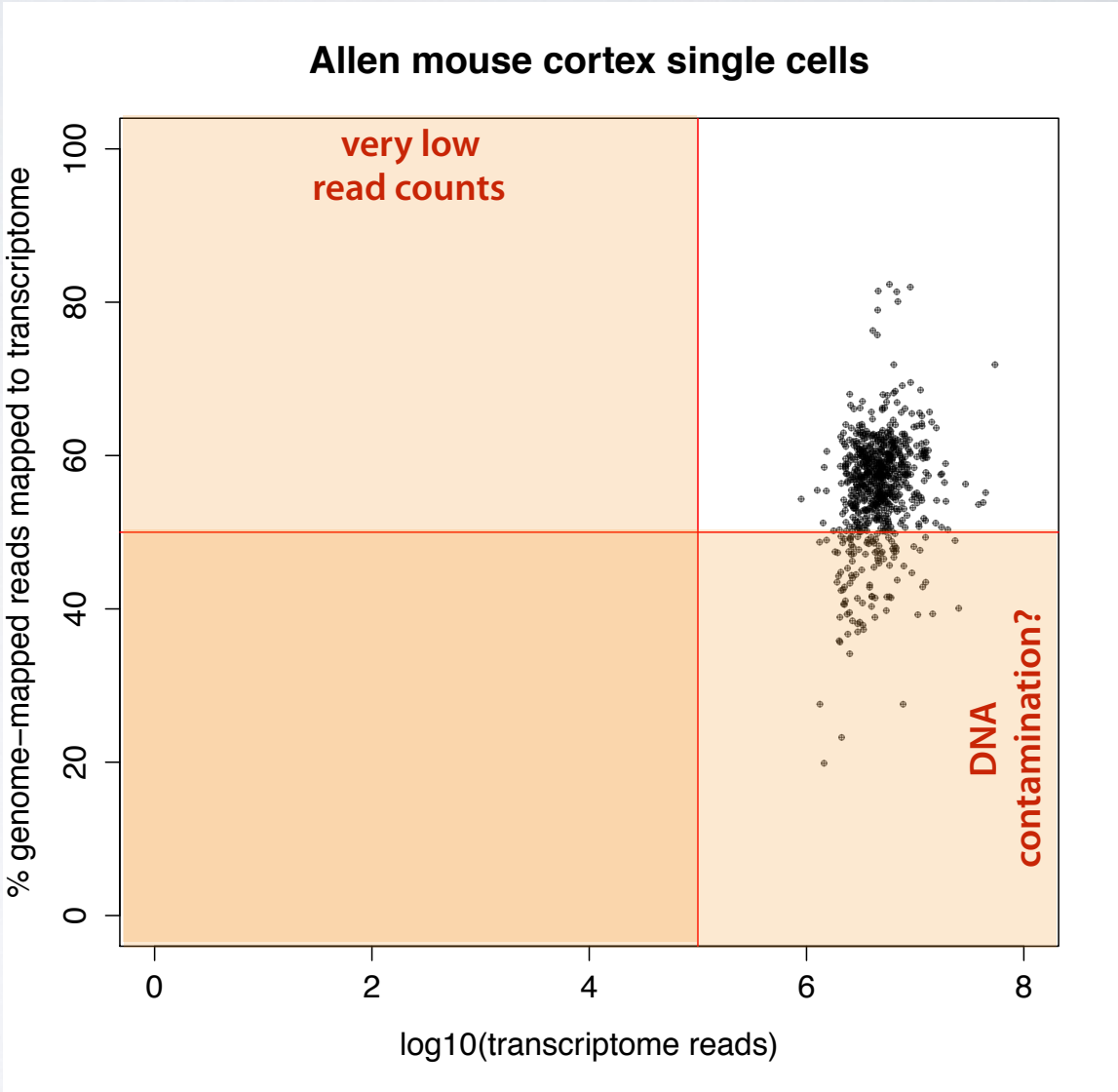
RESOURCE

nature neuroscience

Adult mouse cortical cell taxonomy revealed by single cell transcriptomics

Bosiljka Tasic[1,2], Vilas Menon[1,2], Thuc Nghi Nguyen[1], Tae Kyung Kim[1], Tim Jarsky[1], Zizhen Yao[1], Boaz Levi[1], Lucas T Gray[1], Staci A Sorensen[1], Tim Dolbeare[1], Darren Bertagnolli[1], Jeff Goldy[1], Nadiya Shapovalova[1], Sheana Parry[1], Changkyu Lee[1], Kimberly Smith[1], Amy Bernard[1], Linda Madisen[1], Susan M Sunkin[1], Michael Hawrylycz[1], Christof Koch[1] & Hongkui Zeng[1]

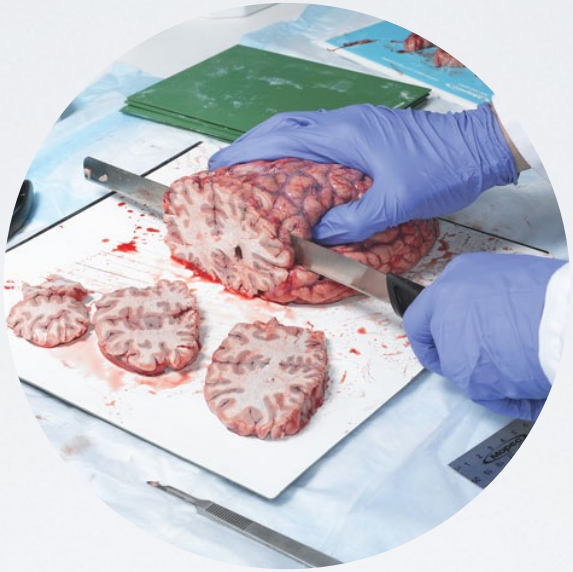




limitations of human transcriptomics

major problems associated with profiling RNA abundance in the human brain

death
&
PMI



sample quality
&
biopsy precision

cell-types
&
transgenics

limitations of human transcriptomics

major problems associated with profiling RNA abundance in the human brain

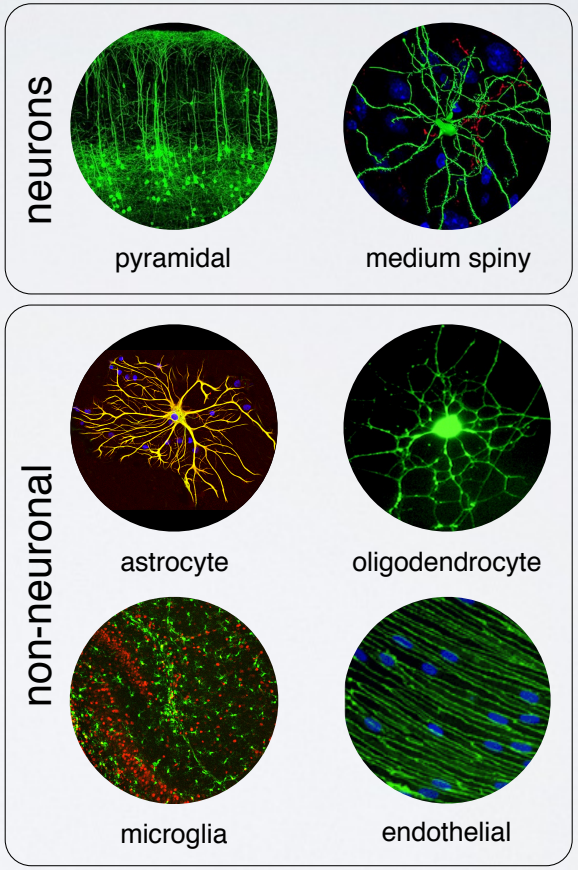
death & PMI

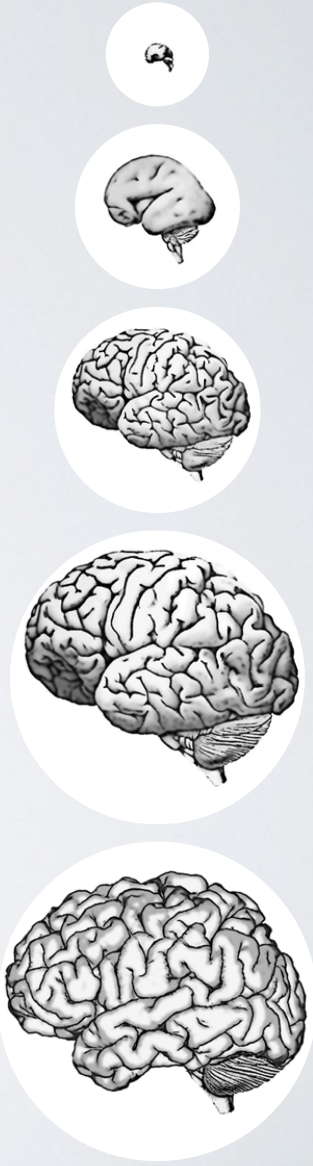
sample quality & biopsy precision

cell-types & transgenics

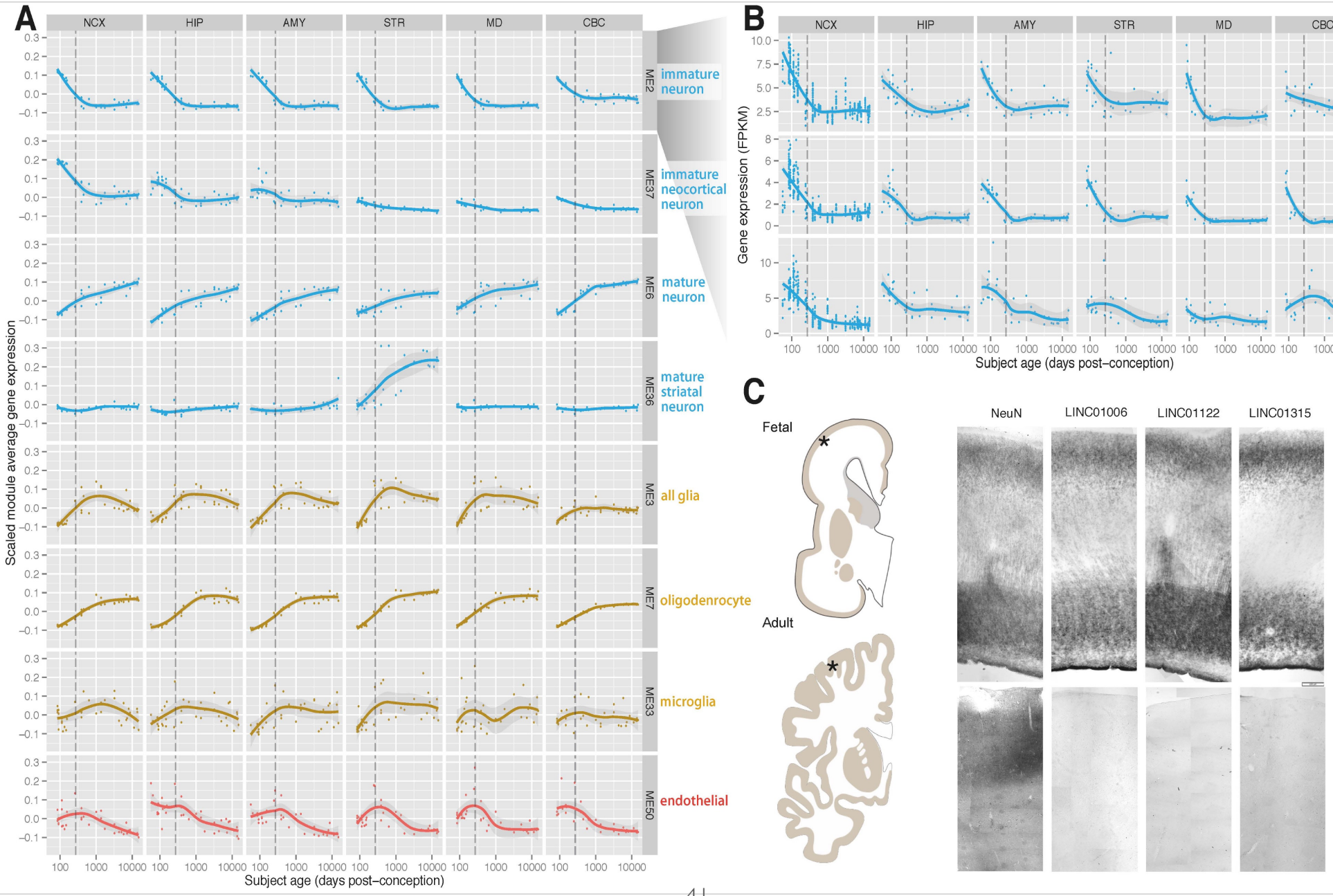
overcoming limitations of human transcriptomics

how can we use high-quality, high-resolution mouse data to learn more about the development of the human brain?

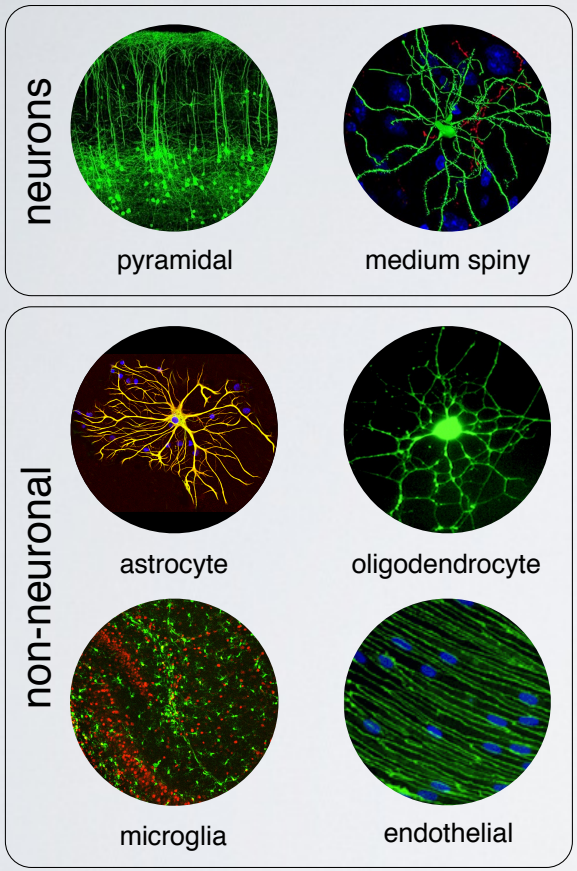


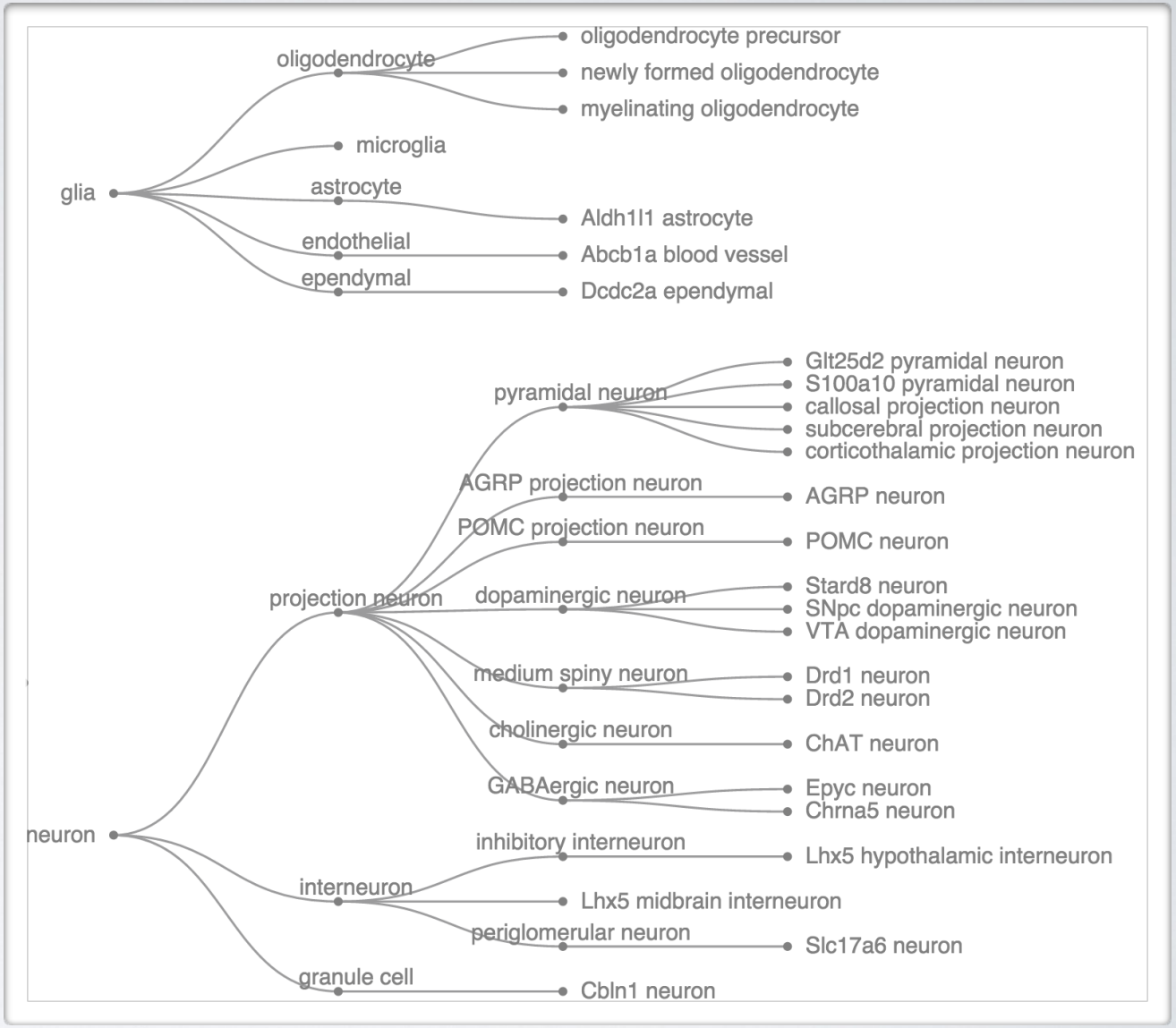


early fœtal, neuron-specific, human lincRNAs

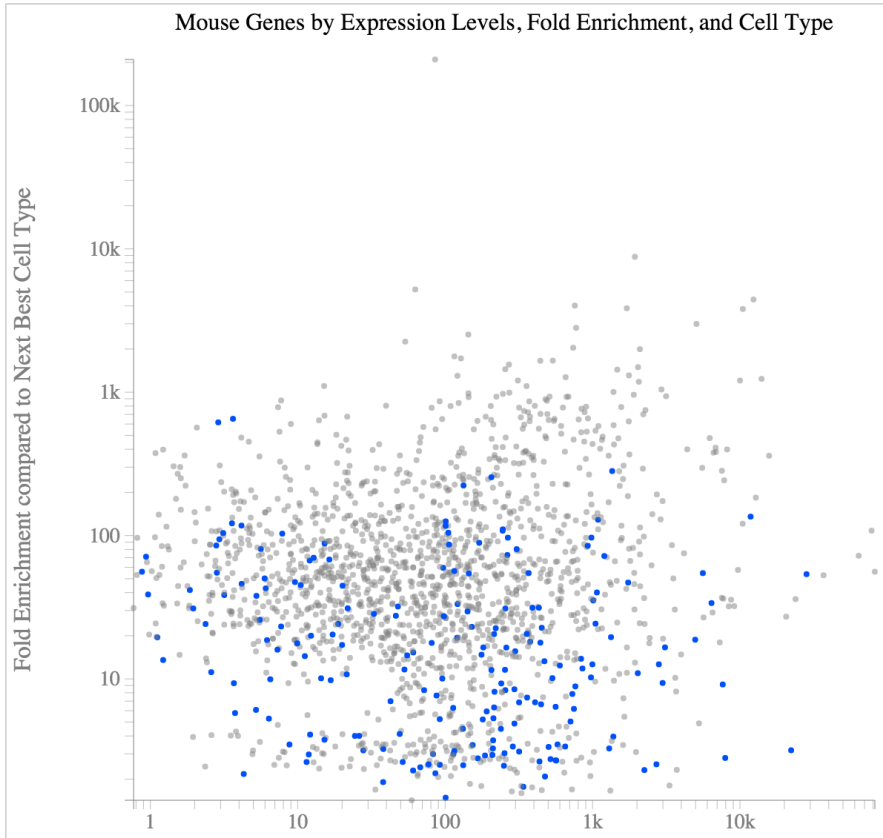


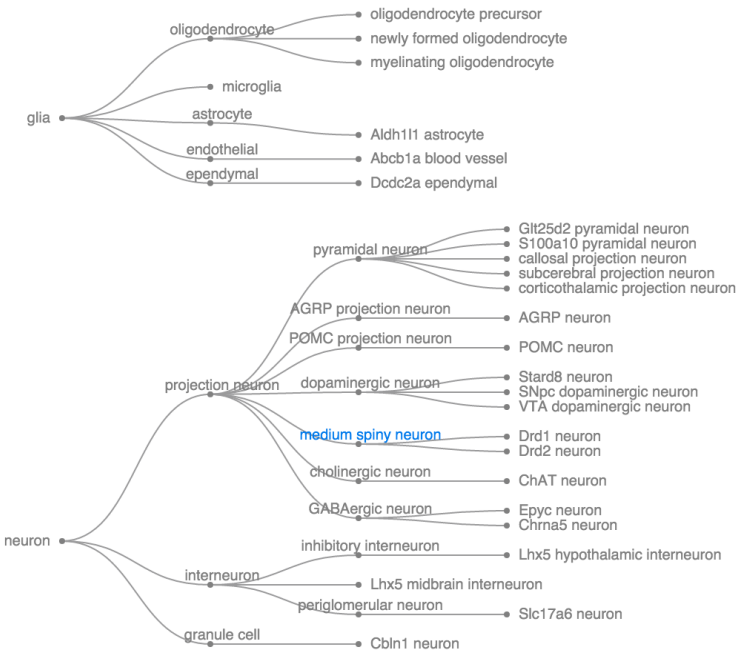
current database of cell-type specific raRNA-seq



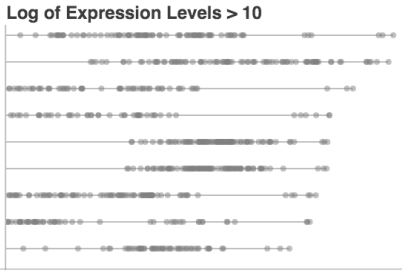


interactive web resource of cell-type enriched genes





Gene Id	Cell	Fold Enrichment
ENSMUSG00000028785.10	Drd2 neuron	54
ENSMUSG00000090223.1	medium spiny neuron	3
ENSMUSG00000045573.9	Drd2 neuron	136
ENSMUSG00000061718.10	medium spiny neuron	34
ENSMUSG00000001270.8	Drd2 neuron	3
ENSMUSG00000030122.10	Drd2 neuron	9
ENSMUSG00000032503.15	medium spiny neuron	19
ENSMUSG00000032259.7	Drd2 neuron	55
ENSMUSG00000064288.4	Drd2 neuron	17



uses of cell-type enriched gene catalogue

- **clustering and single-cells**

  - use known cell-types to inform clustering of single cell based on expression profiles

- **deconvolution!**

  - use to interpret gene expression profiling of human CNS

- **cell-type specific PPI + coexpression networks**

  - use to find cell-type specific hubs/bottlenecks

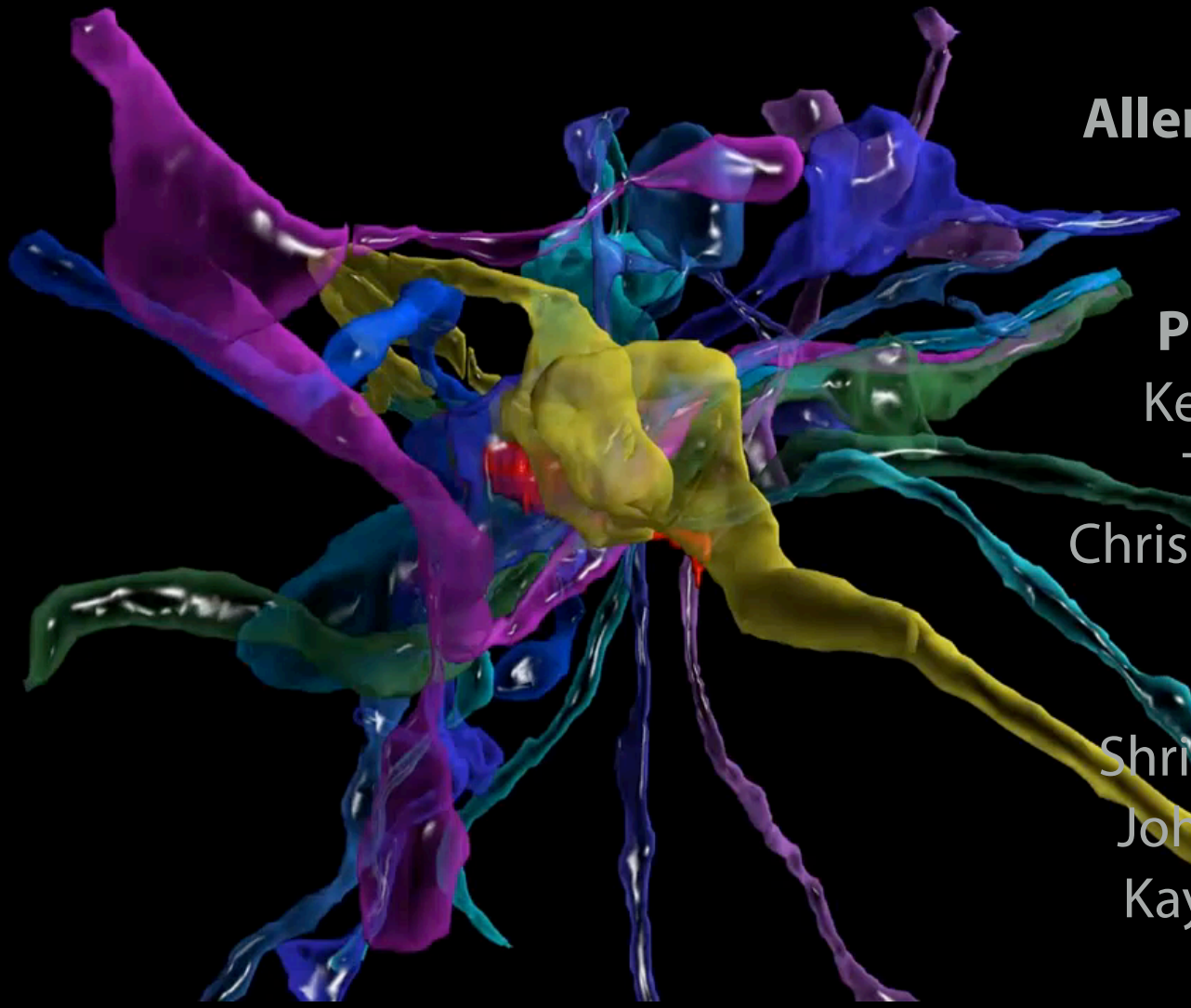
  - use to refine cell-type hierarchy

acknowledgements

Gerstein lab
Mark Gerstein
Arif Harmanci
Joel Rozowsky
Anurag Sethi
Jing Zhang

Nairn lab
Angus Nairn
Becky Carlyle

Sestan lab
Nenad Sestan
Yuka Kawasawa
Mingfeng Li



Allen Institute
Ed Lein

Proteomics
Ken Williams
Tukiet Lam
Chris Colangelo

YCGA
Shrikant Mane
John Overton
Kaya Bilguvar

Lichtman 2015 Cell