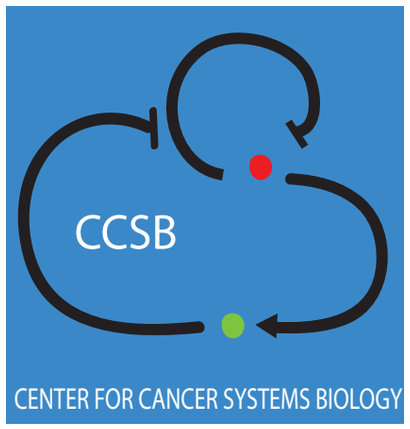
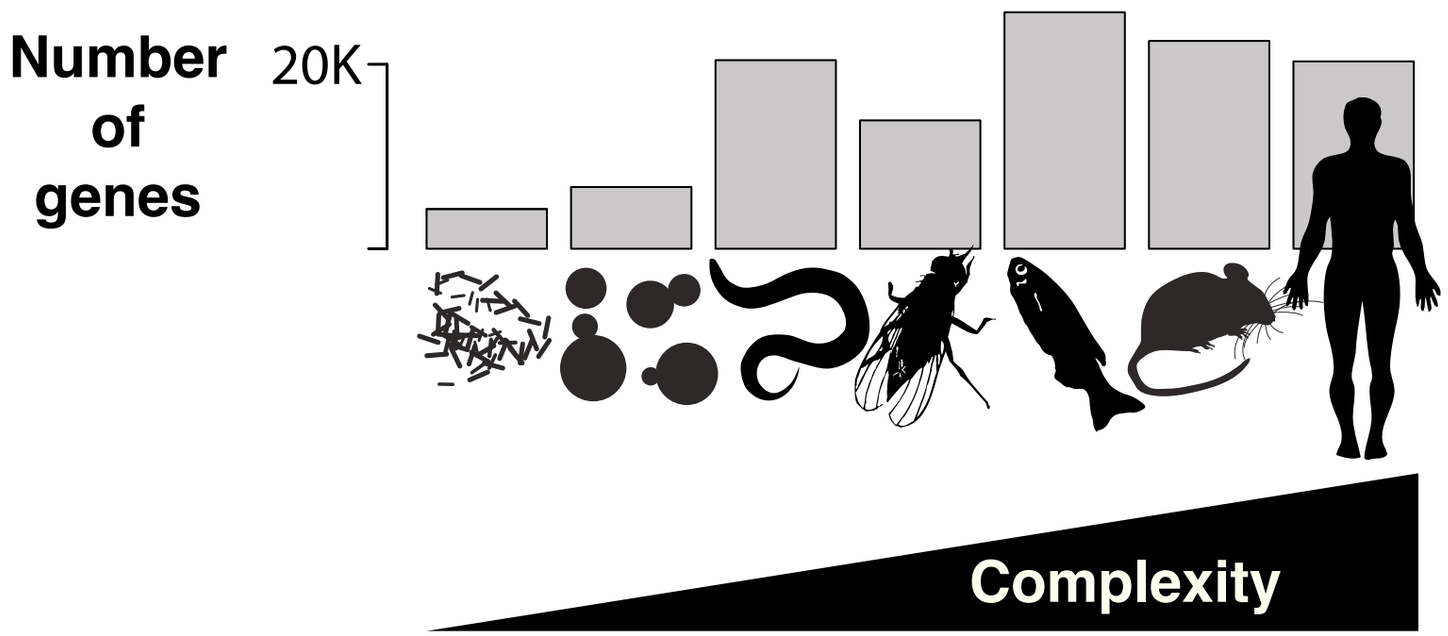
Tools and technologies to characterize isoforms at proteome-scale

Gloria Sheynkman
Marc Vidal Laboratory
Center for Cancer Systems Biology, Dana Faber Cancer Institute
Department of Genetics, Harvard University

SMRTLeiden
May 2nd, 2017

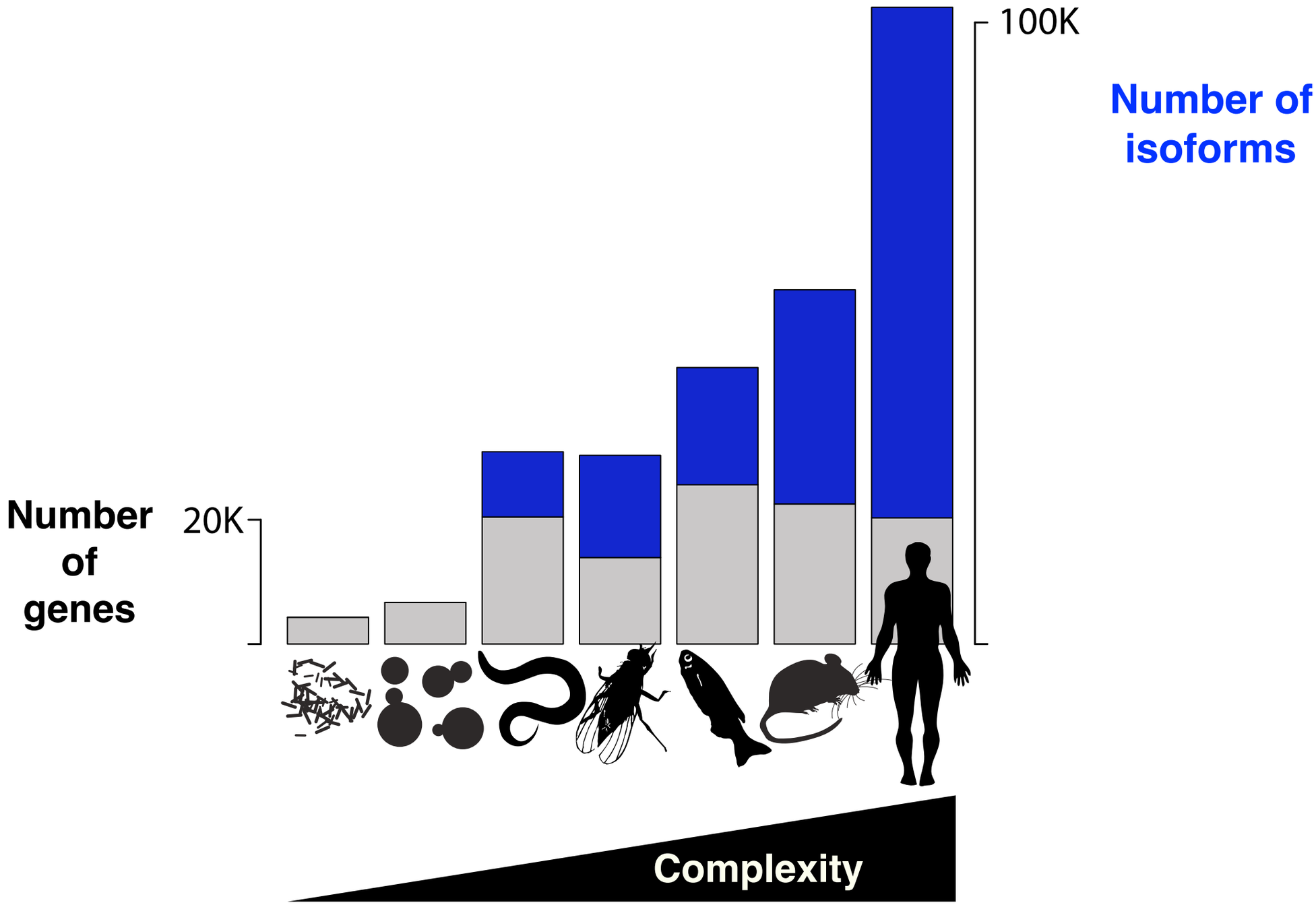


Gene numbers



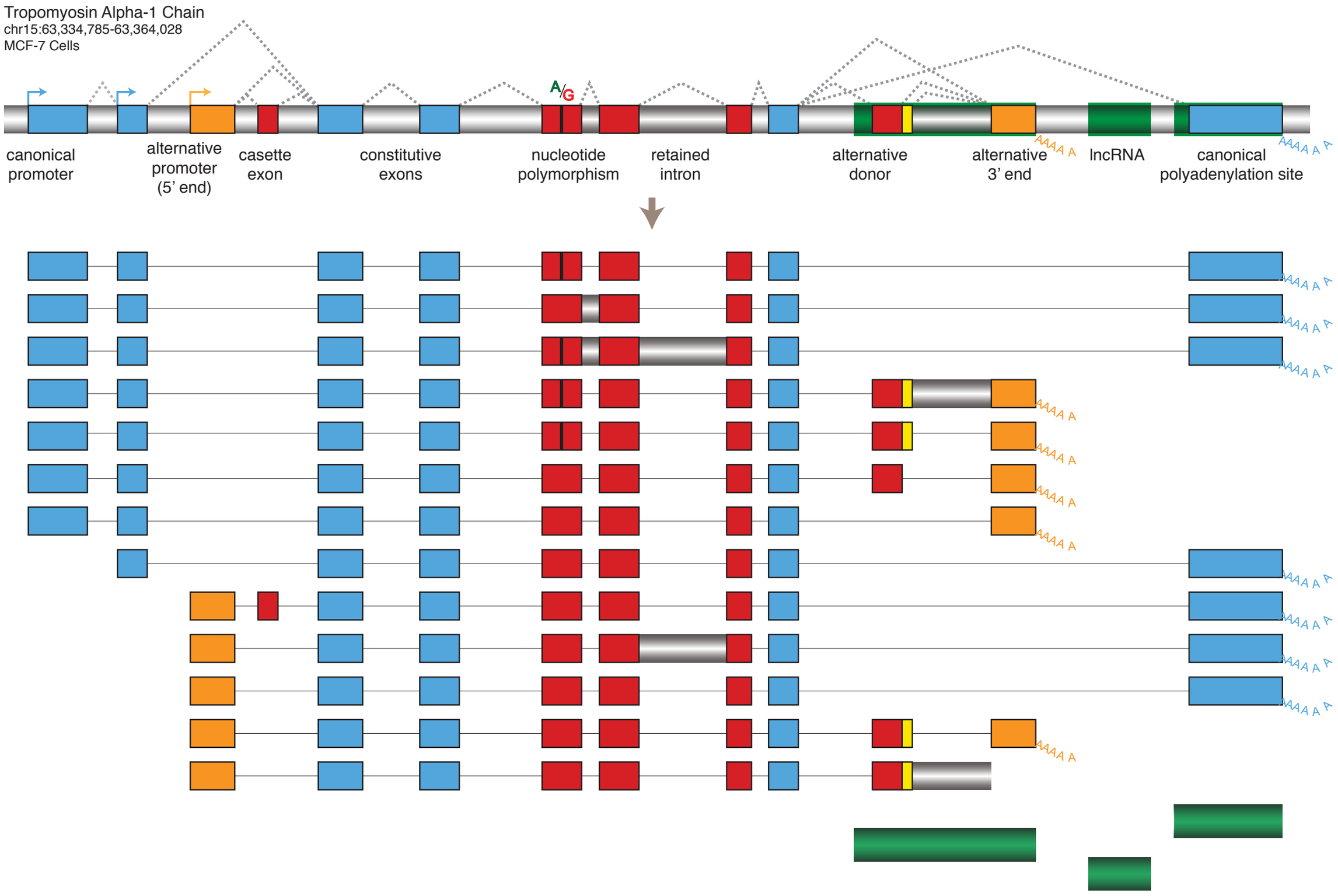
Ensembl Dec 2015 release

Isoform numbers

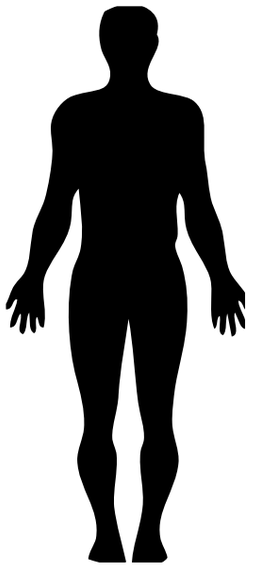


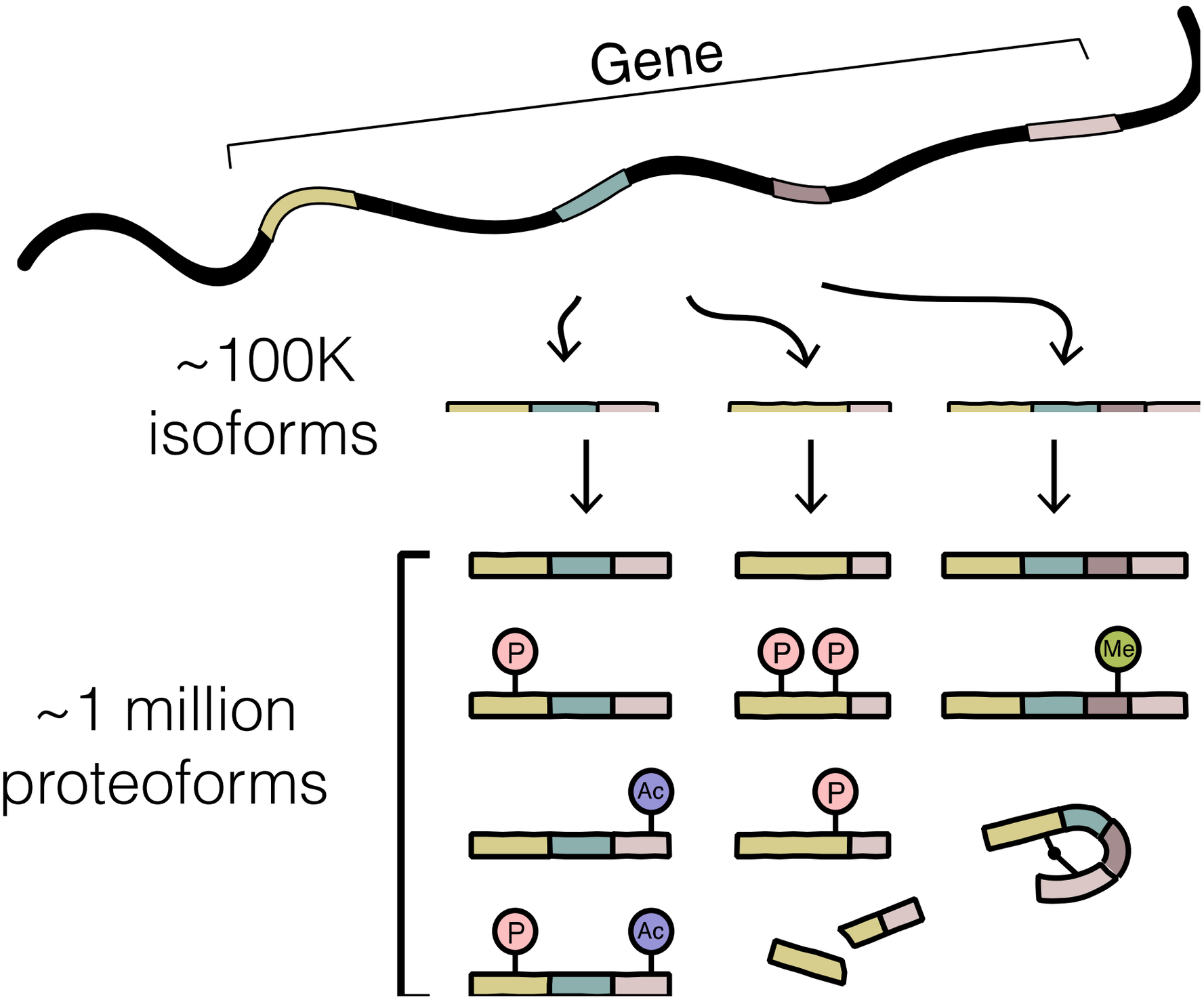
Ensembl Dec 2015 release

Combinations of splice sites produce diverse protein isoforms.



The proteoform hypothesis





Smith & Kelleher *Nat Methods* 2013

Splicing regulation and disease

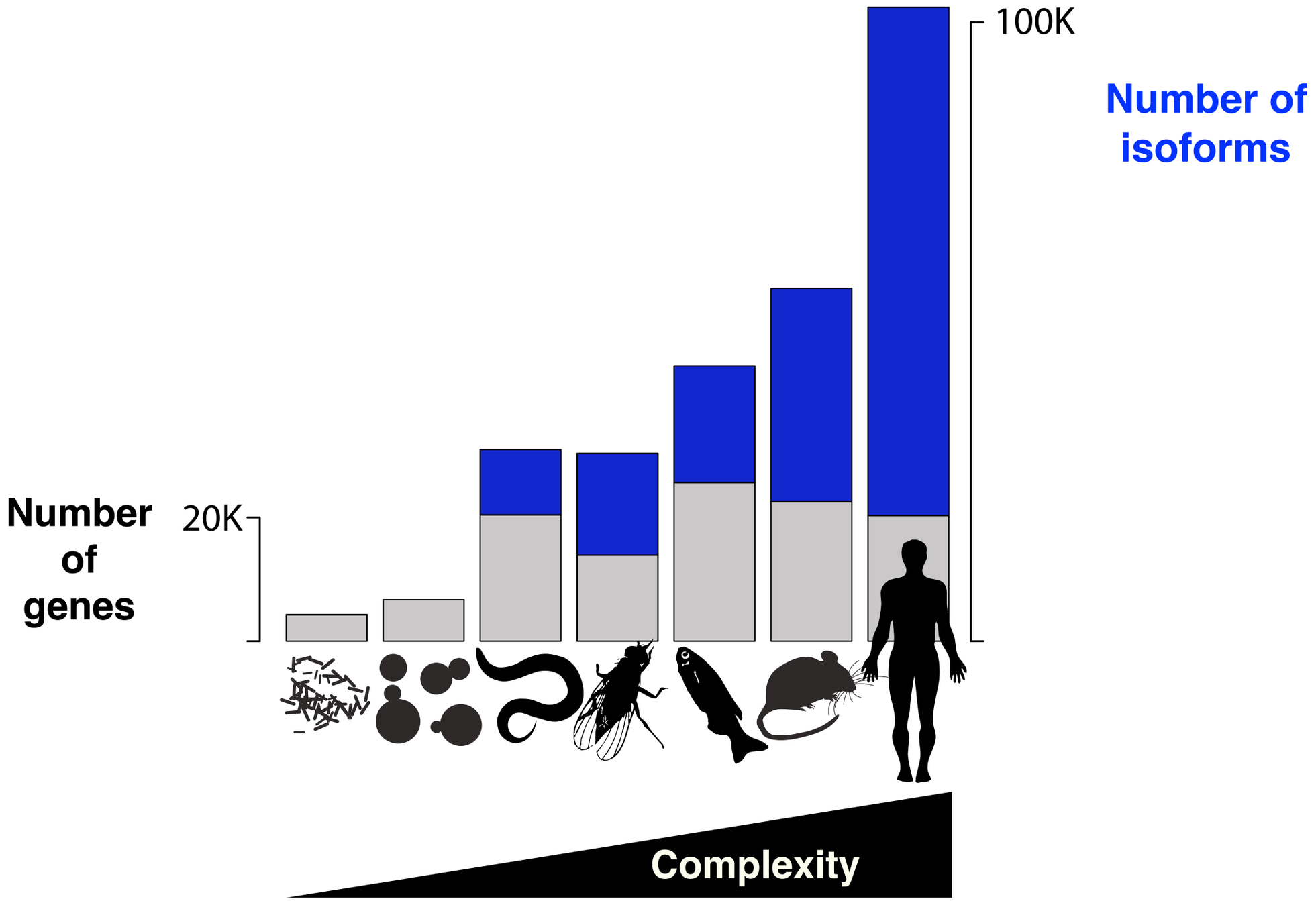
- splicing is pervasive, inherent to encoded products of the genome

- splicing is highly regulated in space and time
  - high tissue- and developmental- specificity
  - “splice code”

- splicing is dysregulated in many diseases, including cancer
  - estimates of 50% all disease variants affect splicing
  - splice-modulating therapies (e.g. antisense oligos)

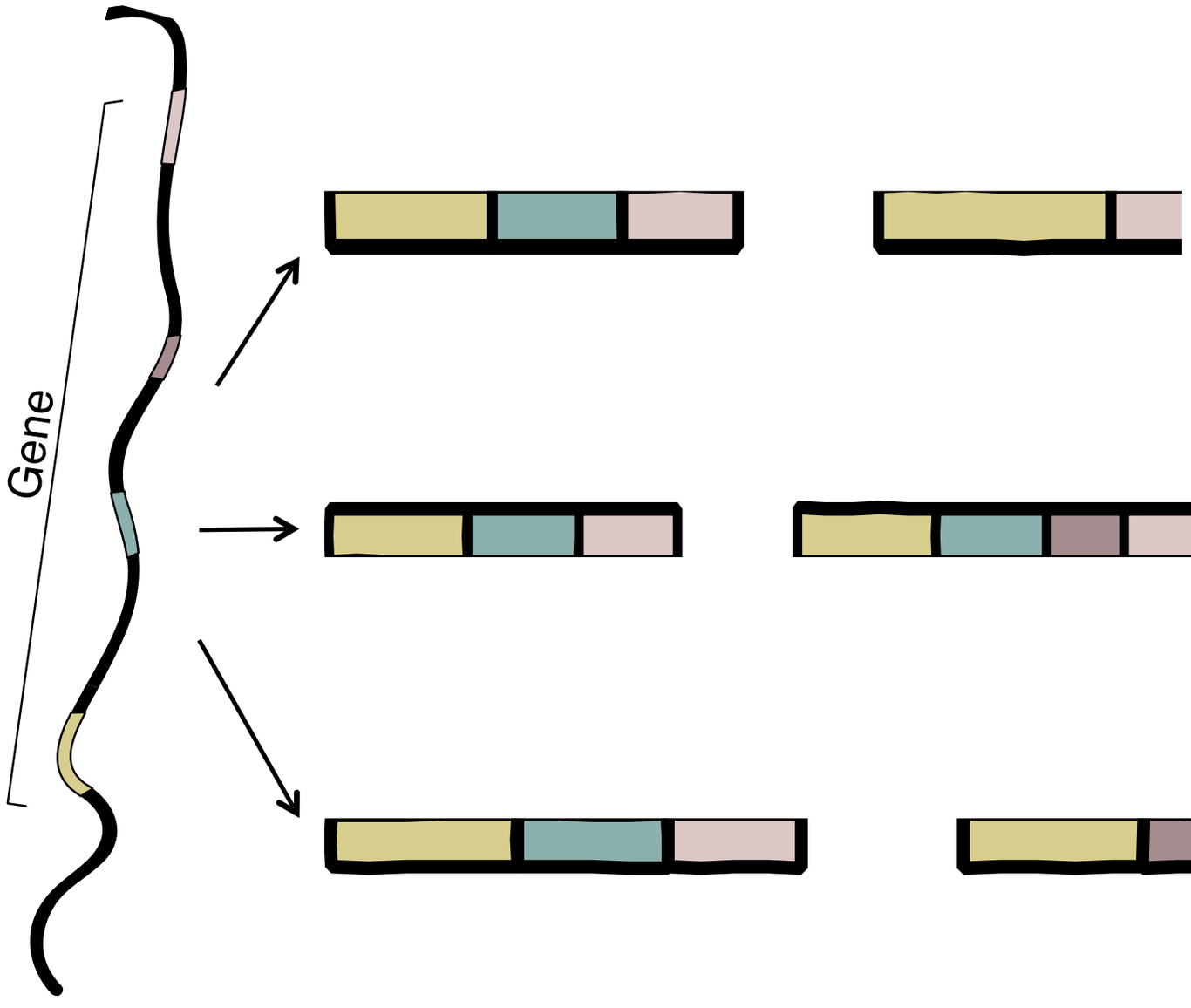
Wang et al *Nature Review Genetics* (2007),

Isoform function?



Ensembl Dec 2015 release

Isoforms and functional divergence



In vivo functions

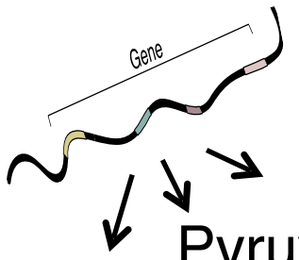
Identical "Isoforms"

Different "Alloforms"

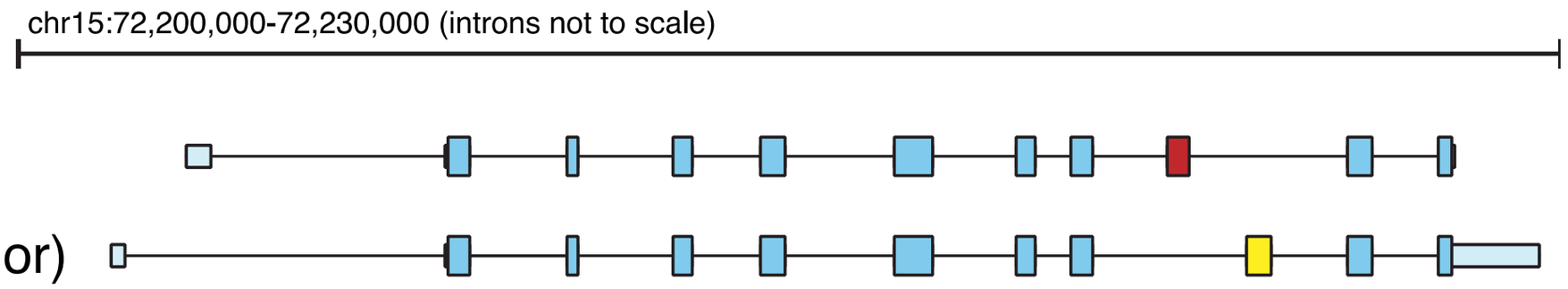
Opposite "Antiforms"

Examples of functionally divergent isoforms

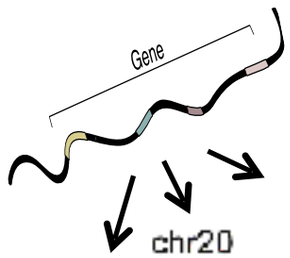
Alloforms

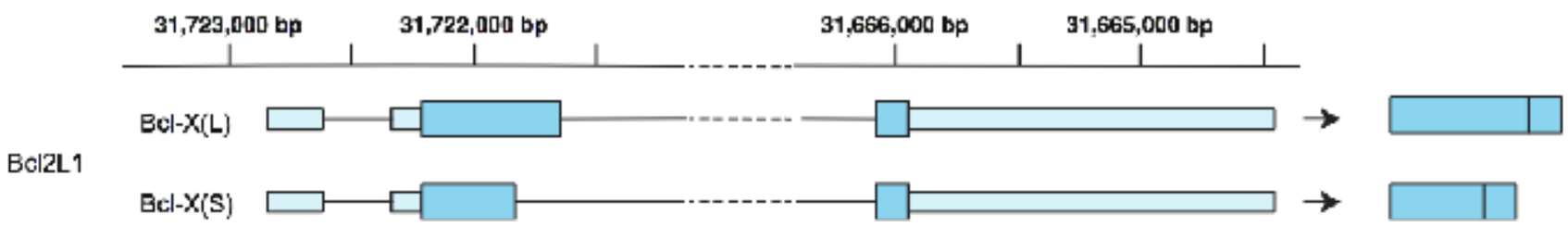


Pyruvate kinase



Antiforms





Bcl-X anti-apoptotic

Bcl-X pro-apoptotic

Christofk *et al Nature* 2008

Schwerk *et al Mol Cell* 2005

Divergent functional capabilities described in literature

Isoforms for a few hundred genes

↓

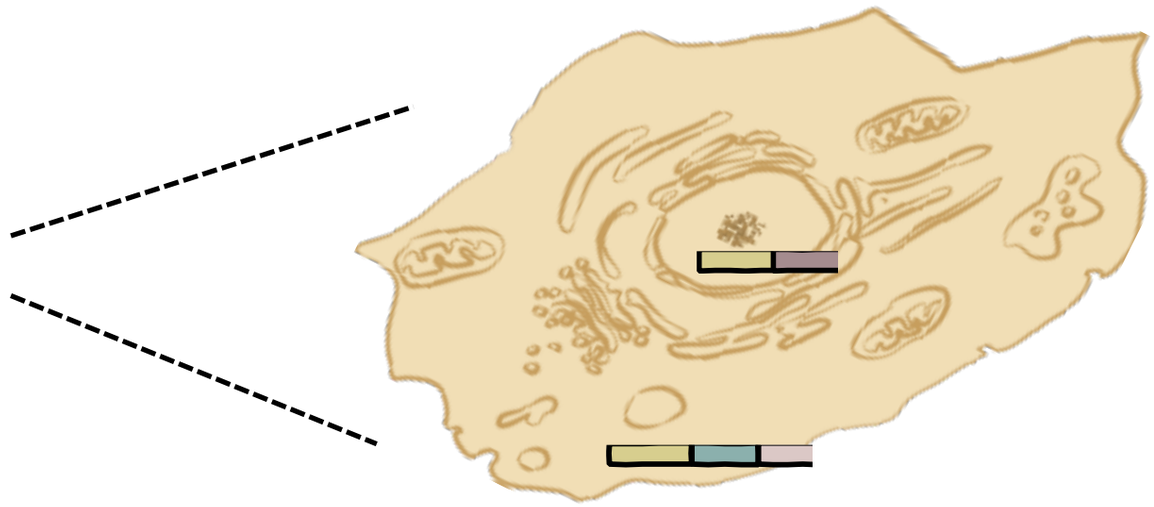
Physical interactions

Cellular localization

Enzymatic activities

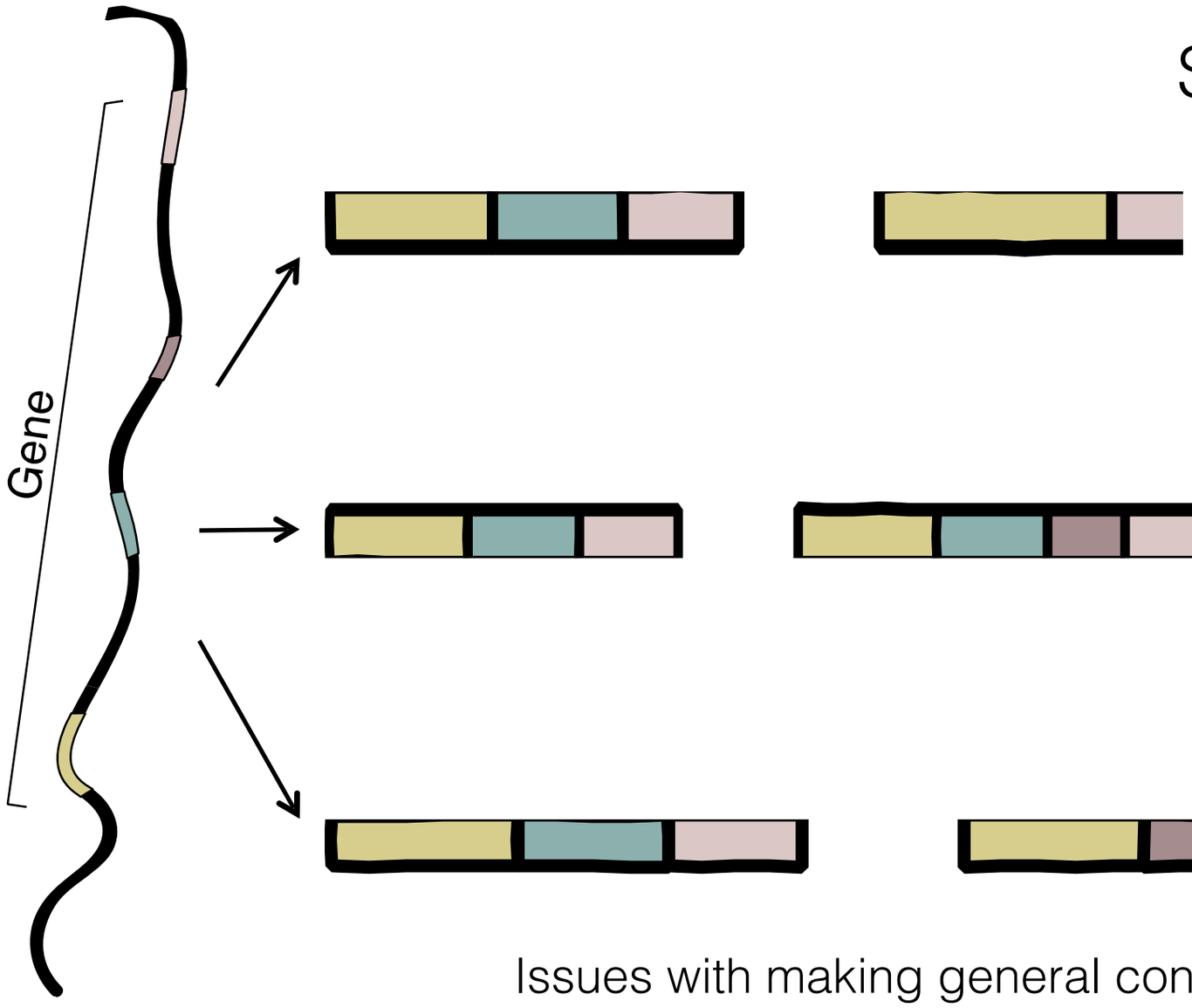
Stability

…..



Kelemen *et al Gene* 2013

Sociological biases in literature



	Publication		
	Submitted	Accepted	
	-	-	Isoforms
	+/-	?	Alloforms
	+	-	Antiforms

Issues with making general conclusions from literature:
-confirmation bias (sampling not random)
-experimental approaches for characterization heterogenous
-isoform identity unknown

See Rolland *et al Cell* 2014

How widespread is isoform functional divergence in the whole proteome?

Systematic identification
of large numbers of
isoform pairs
for large numbers
of human genes

↓

Physical interactions

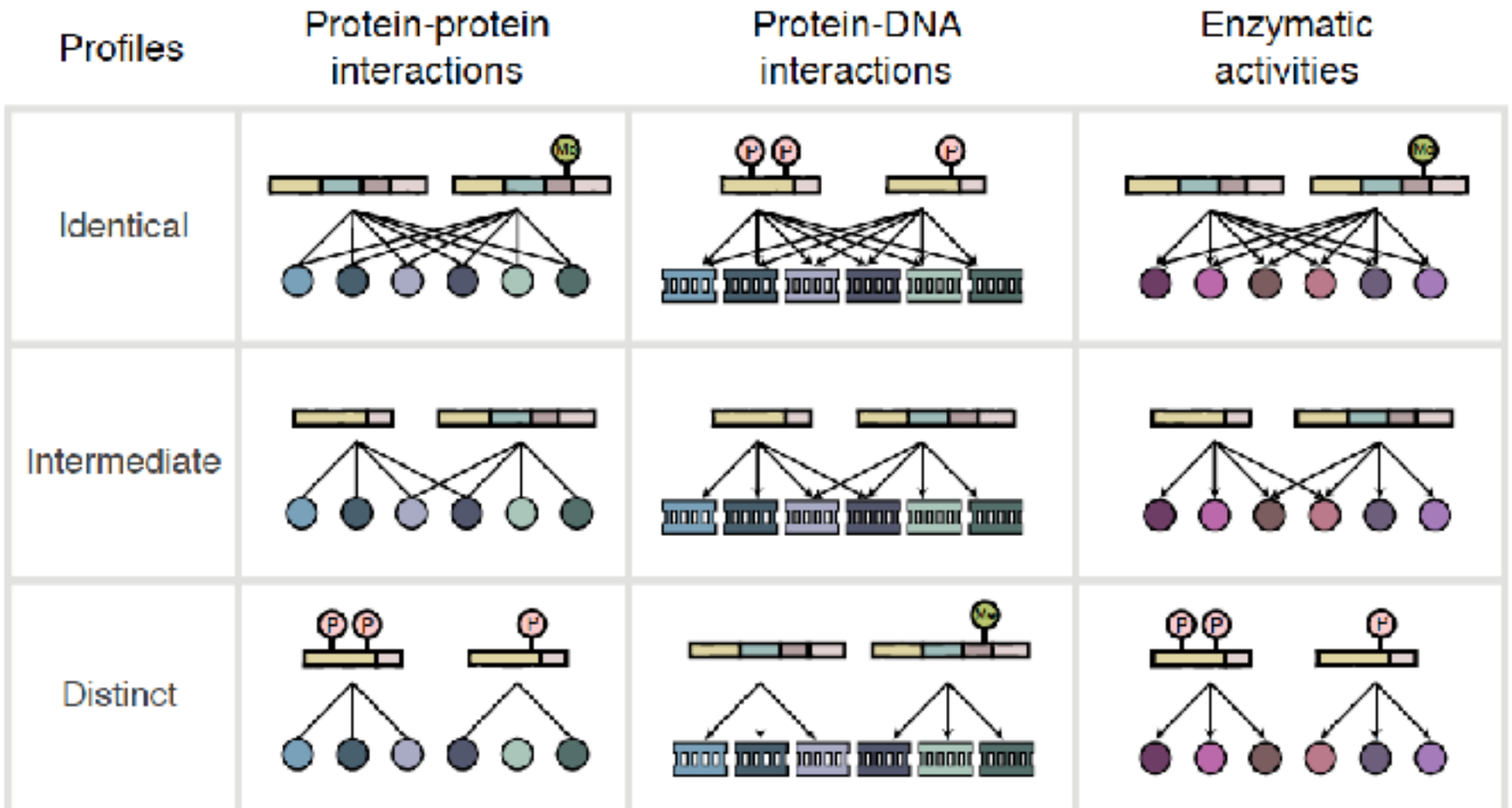
Enzymatic activities

Cellular localization

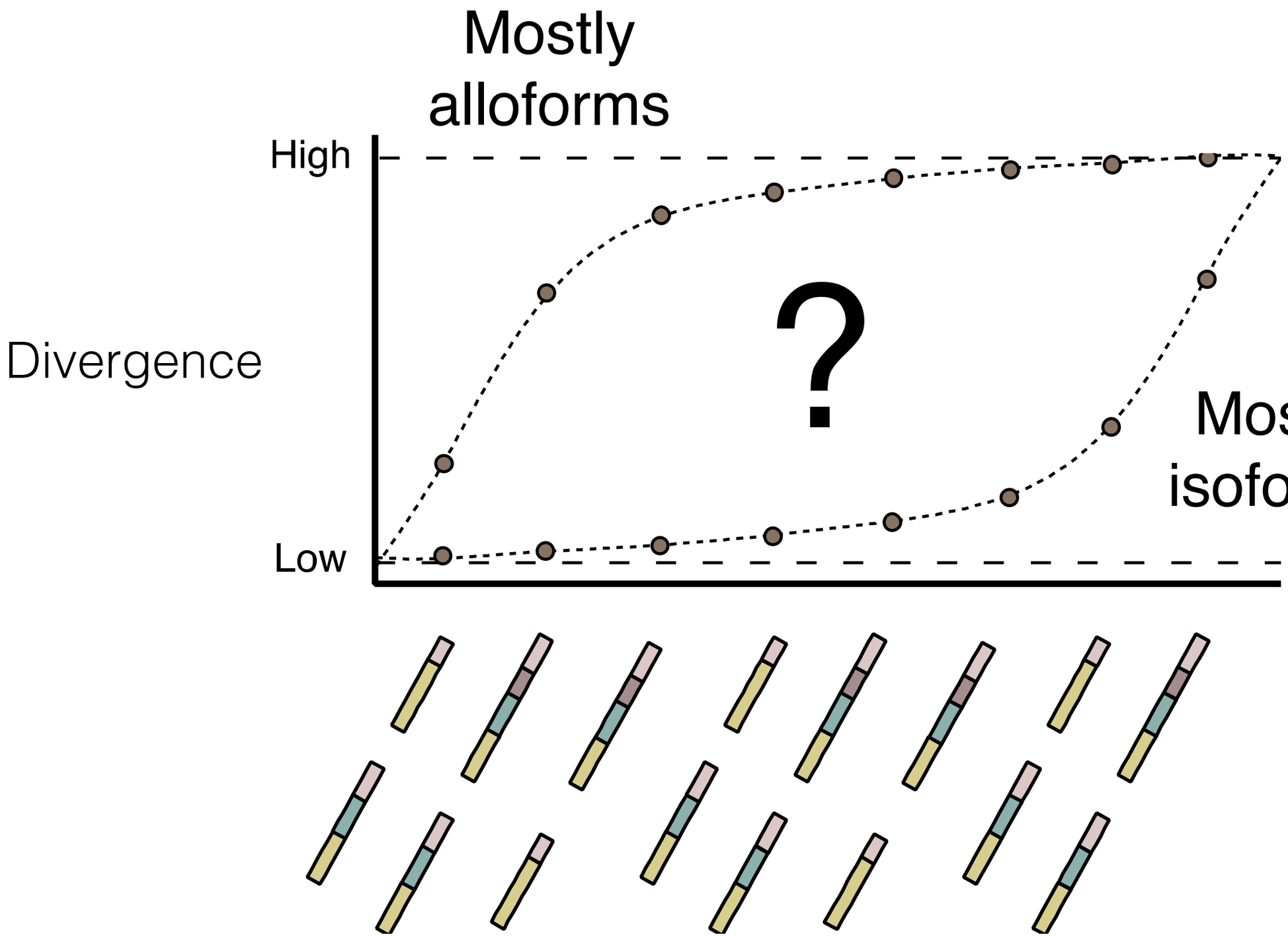
Stability

.....

Unbiased functional profiling



Landscape of protein isoform functional divergence



Large numbers of pairs of isoforms encoded by common genes

How widespread is isoform functional divergence in the whole proteome?

Systematic identification
of large numbers of
isoform pairs
for large numbers
of human genes

↓

Physical interactions

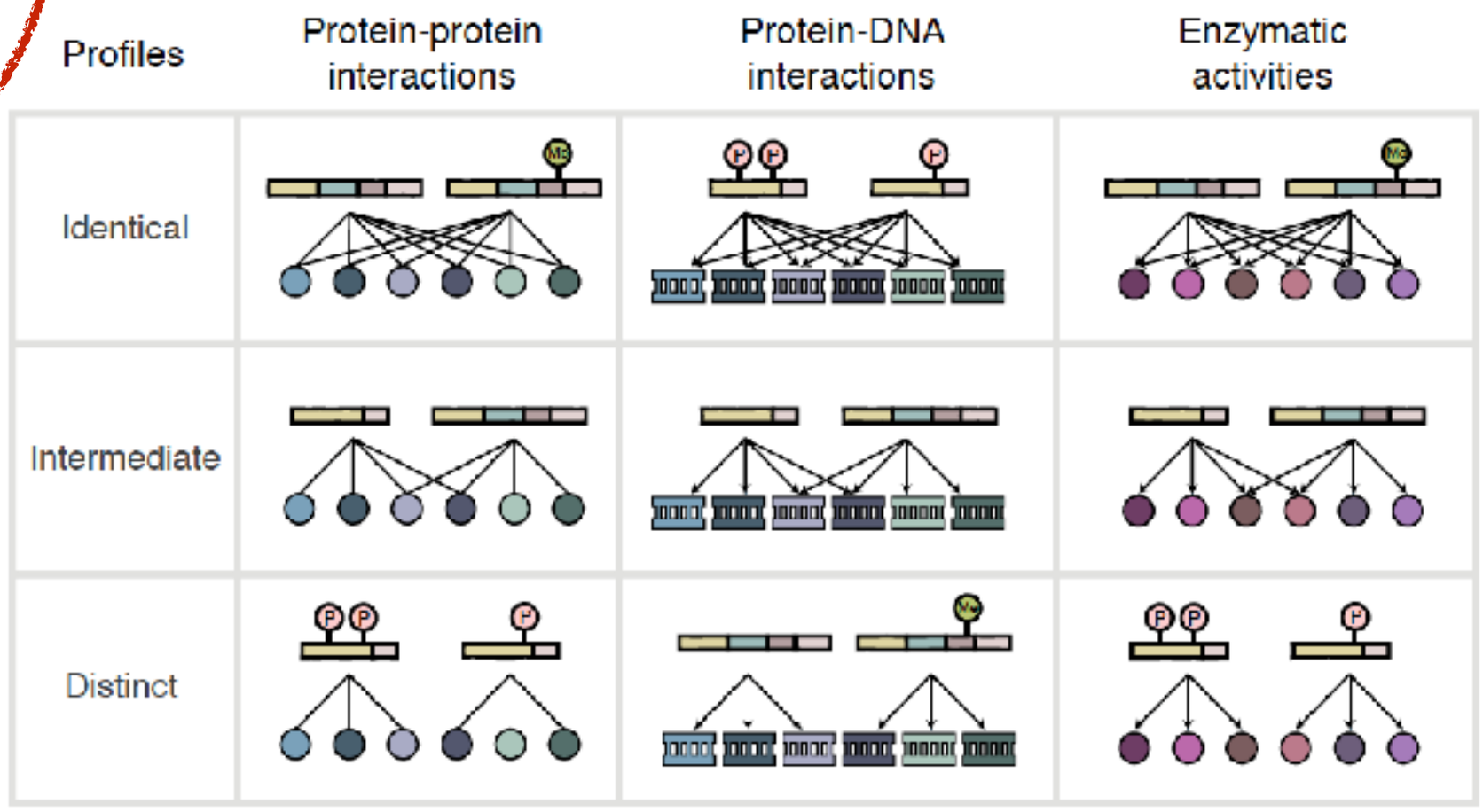
Enzymatic activities

Cellular localization

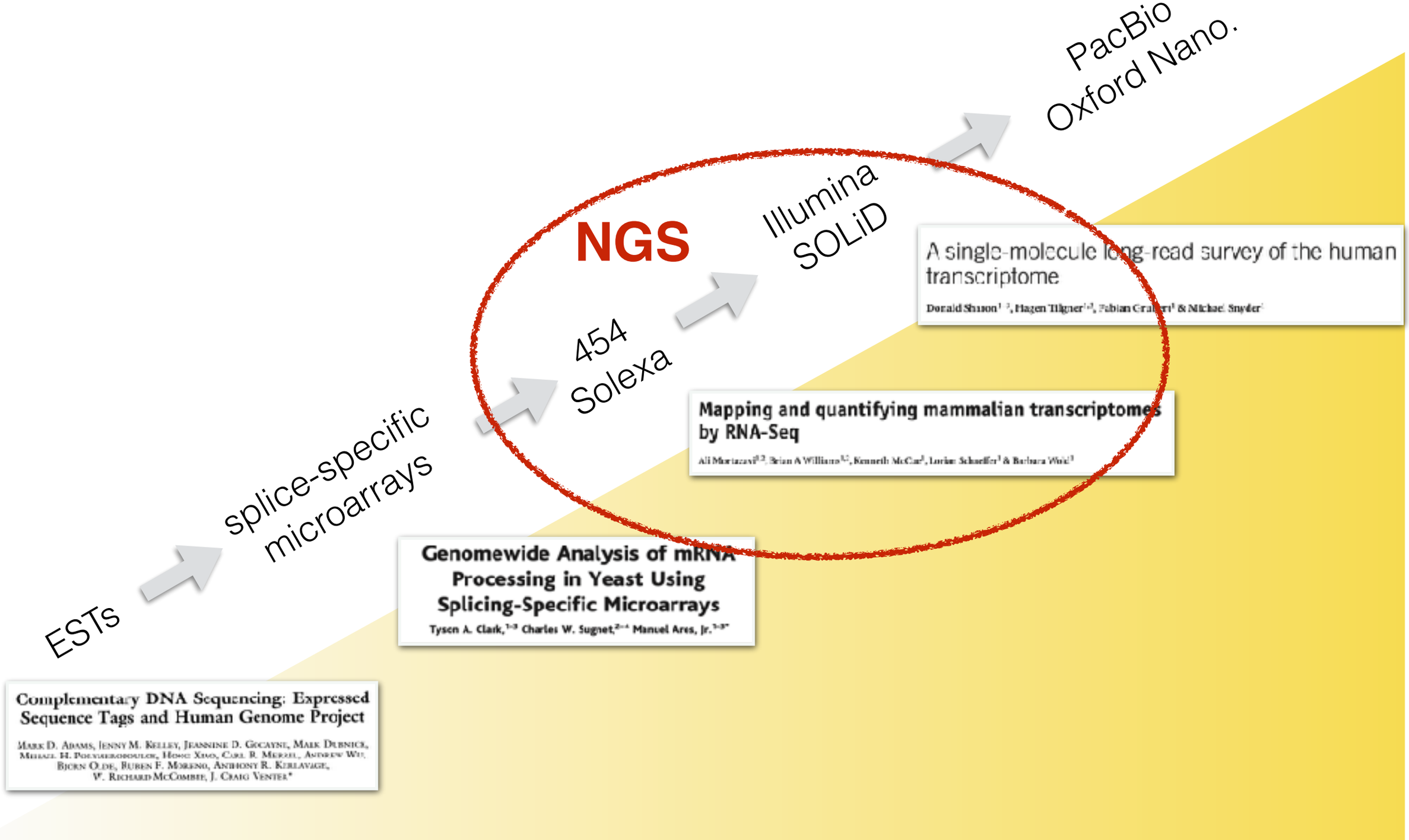
Stability

…..

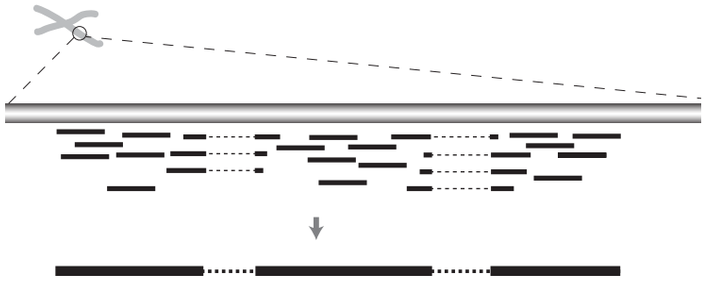
Unbiased functional profiling



RNA sequencing data has been the primary means to characterize isoforms



NGS (RNA-Seq) data can reveal the presence of exons and junctions, but fails to accurately reconstruct full-length isoforms.



transcript reconstruction

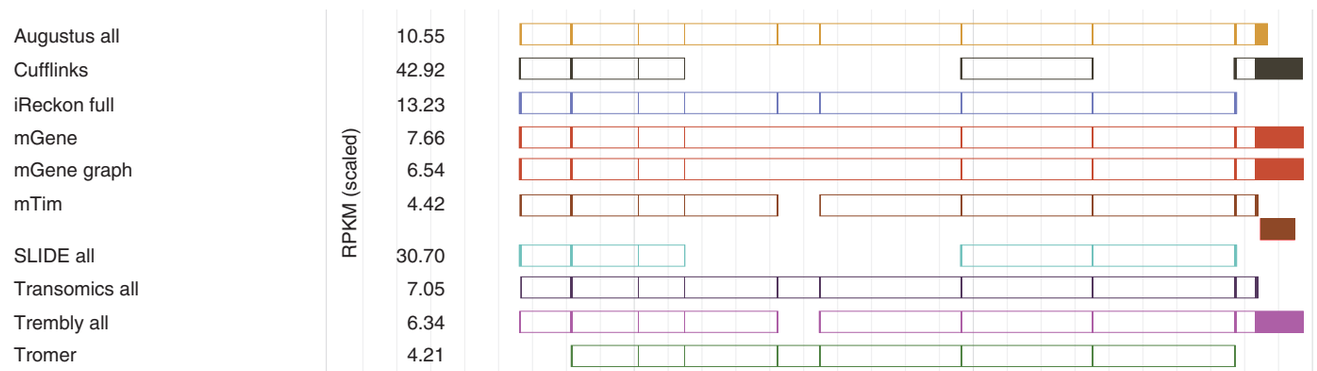
transcript assembly

Korf *Nature Methods* 2013

Assessment of transcript reconstruction methods for RNA-seq

Steijger et al *Nature Methods* 2013

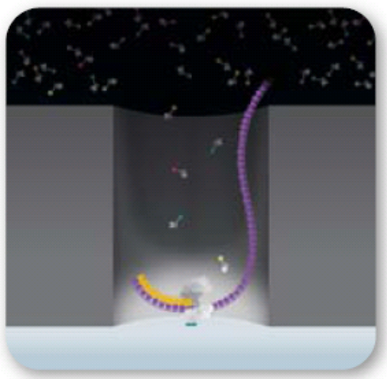
- 14 reconstruction/assembly methods evaluated
- high method variability
- poor performance on simulated data



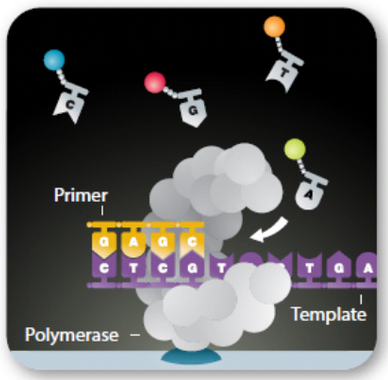
Iso-Seq enables direct sequencing of full-length isoforms and thus characterize transcriptome complexity

PacBio
detection: fluorescence

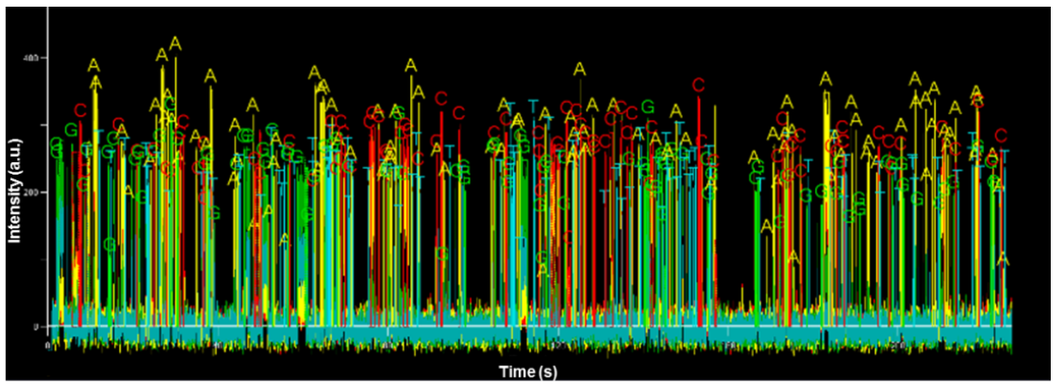
Zero-Mode
Waveguides

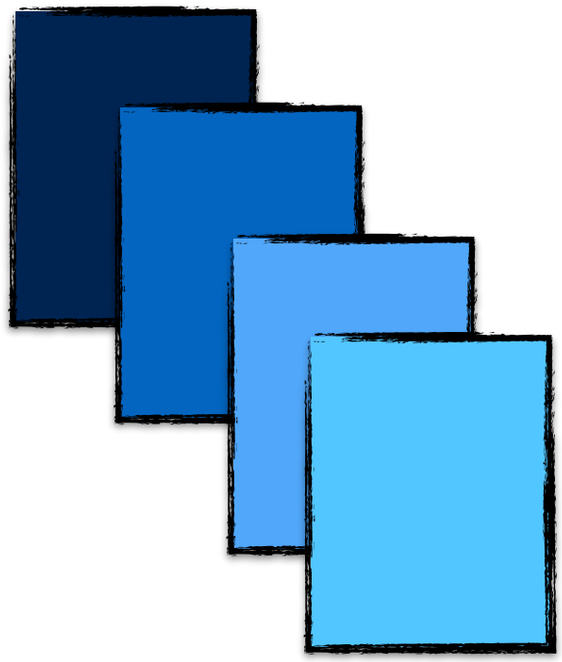


Phospholinked
Nucleotides



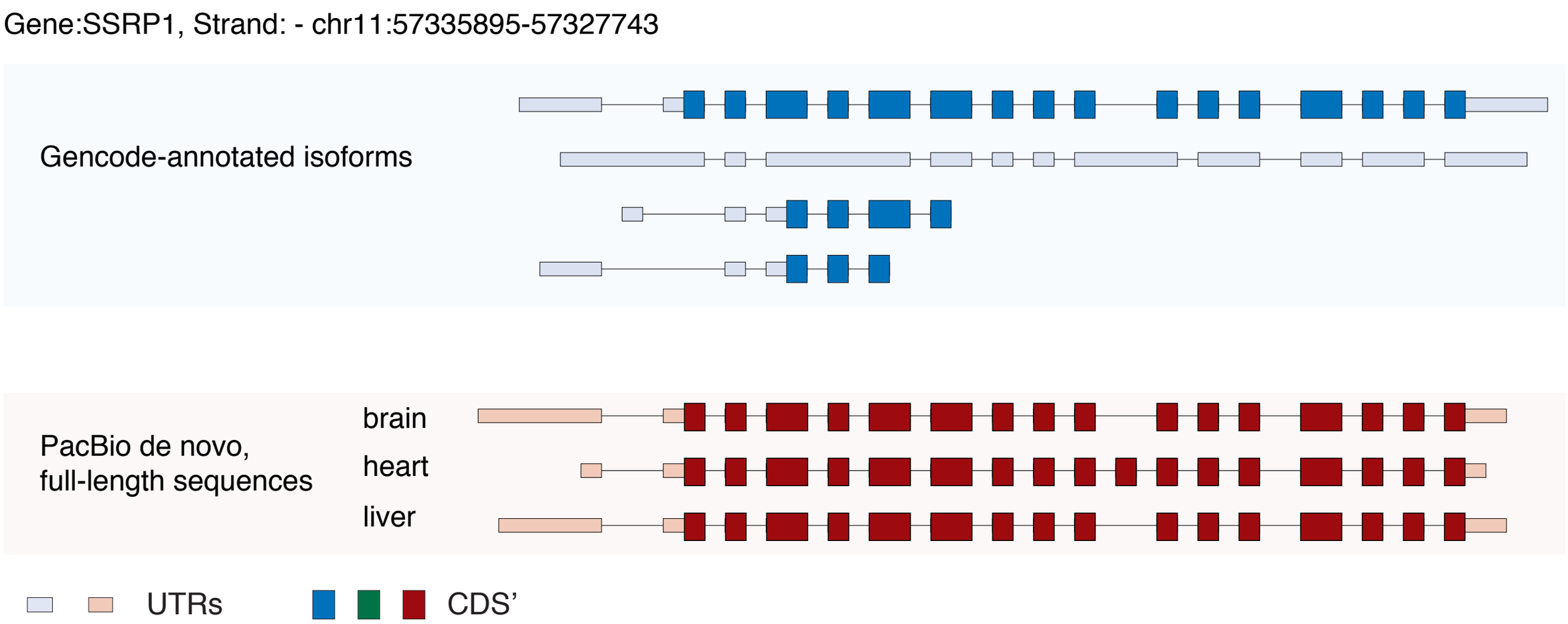
Trace





Eid et al. Science (2009)

Given a base mRNA sequence from Iso-Seq, one can predict open reading frames or “ORFs”
These become candidate protein isoforms for downstream functional analysis



The challenge of detecting low abundance isoforms

Wide dynamic range of human transcriptomes

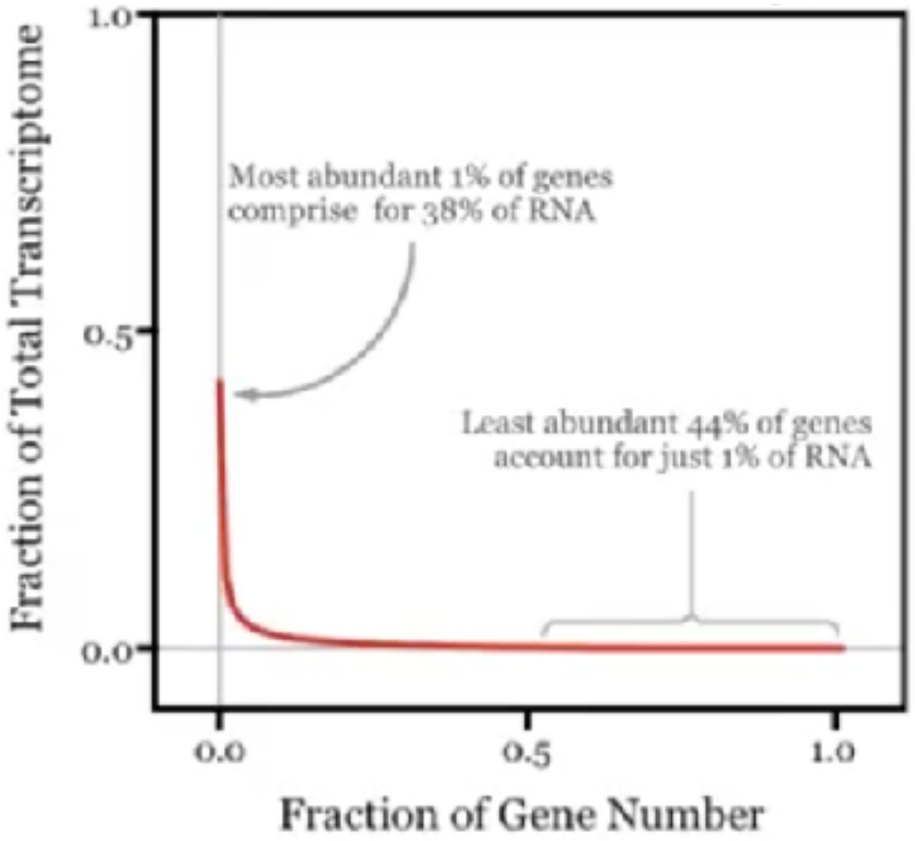
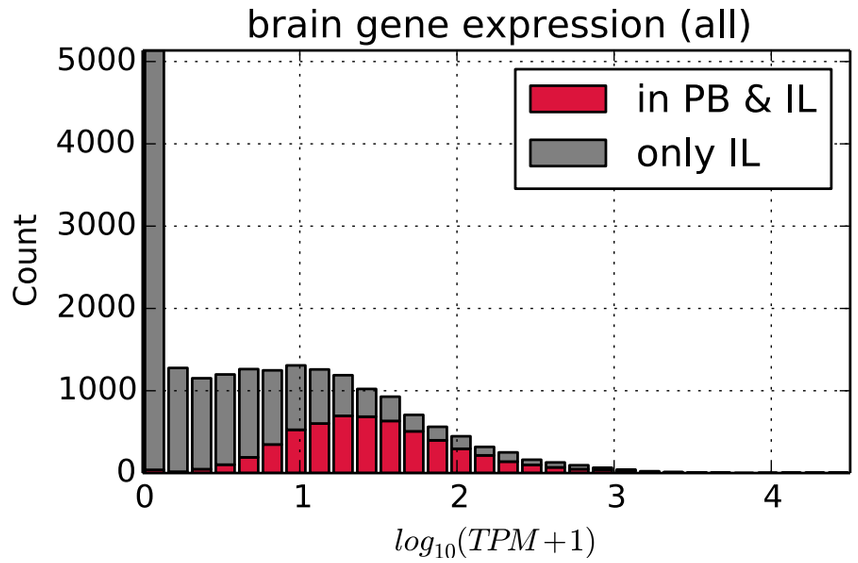


Figure from Tim Mercer (Capture-Seq)

Limited knowledge of isoform complexity
for low-abundance transcripts

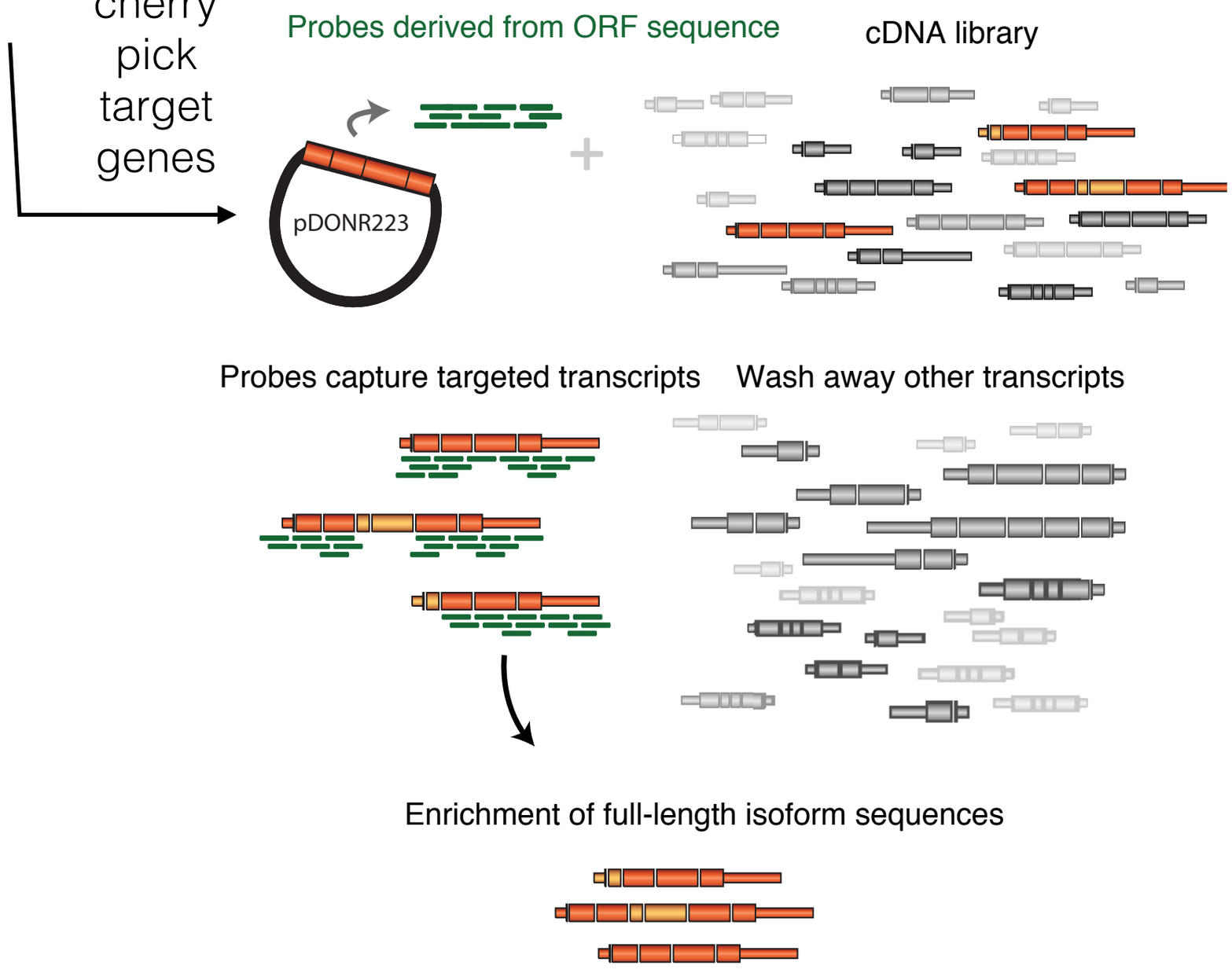


IL = Illumina data
PB = PacBio data

We developed a new method to generate probes en masse for use in hybridization-based target capture: "ORF frag capture" + Iso-Seq

18,000 human ORFs as clones*

cherry pick target genes



Pipeline

retrieve targets (baits)
cherry-pick entry clones from hORFeome

generate probes
biotin-spiked PCR
fragmentation

hyb. capture of targets from pond

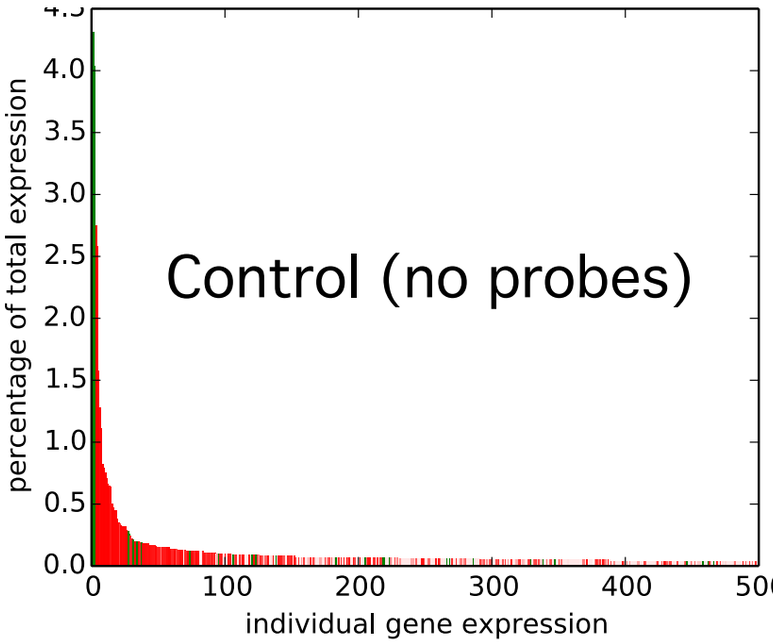
qPCR and NGS QC
quantify fold enrichment
on-target rates

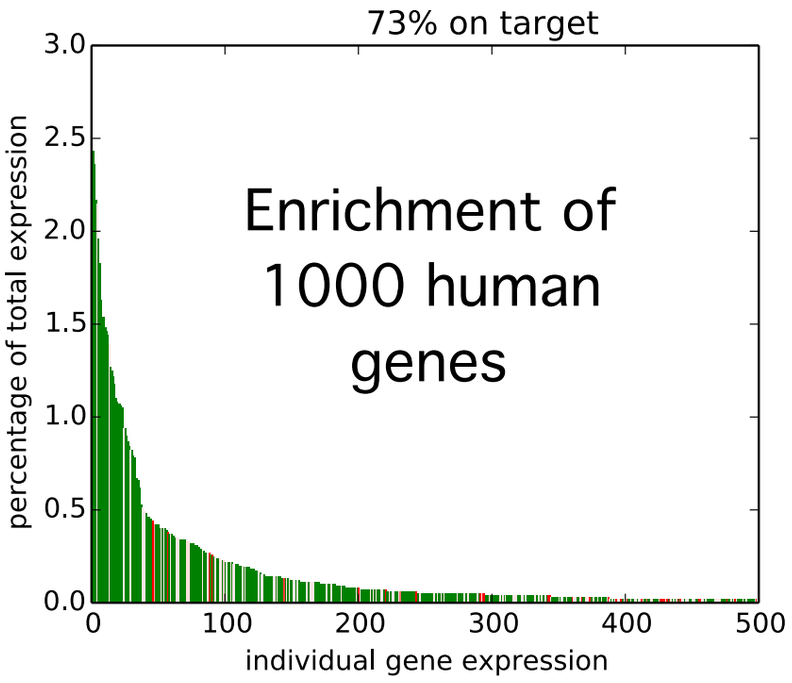
Iso-Seq
full-length sequencing and isoform discovery

*The ORFeome Collaboration *Nature Methods* (2016)

w Jason Underwood, Tyson Clark

“ORF frag capture” + Iso-Seq enables of high sensitivity discovery of full-length isoforms





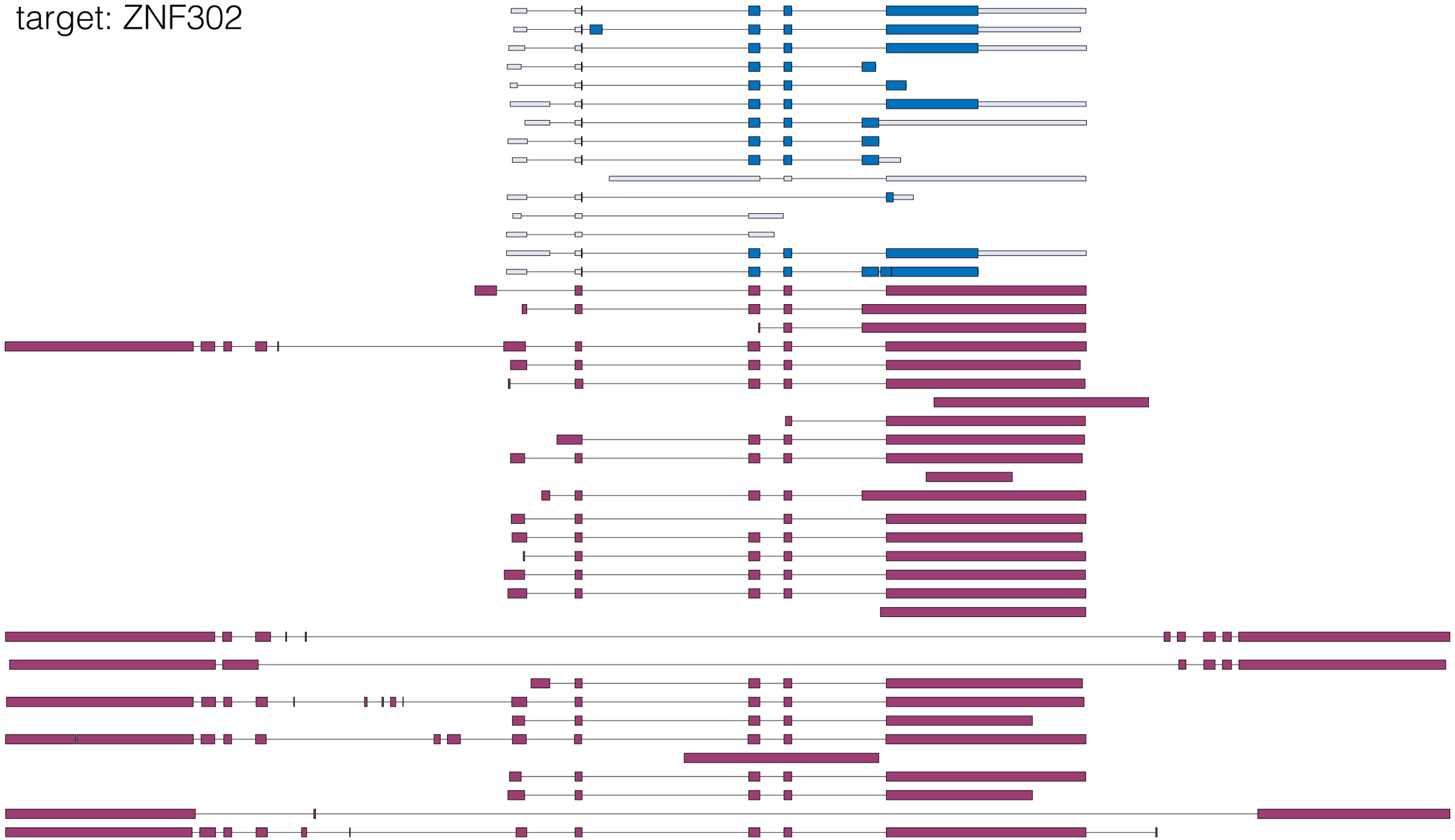
CREB3L4 isoforms detected*

FOXP1 isoforms detected*

*No isoforms detected in ultra-deep-coverage PacBio dataset of the same brain cDNA library.

“ORF frag-based” probe capture, sequencing, and discovery of full-length isoforms

target: ZNF302



How widespread is isoform functional divergence in the whole proteome?

Systematic identification
of large numbers of
isoform pairs
for large numbers
of human genes

↓

Physical interactions

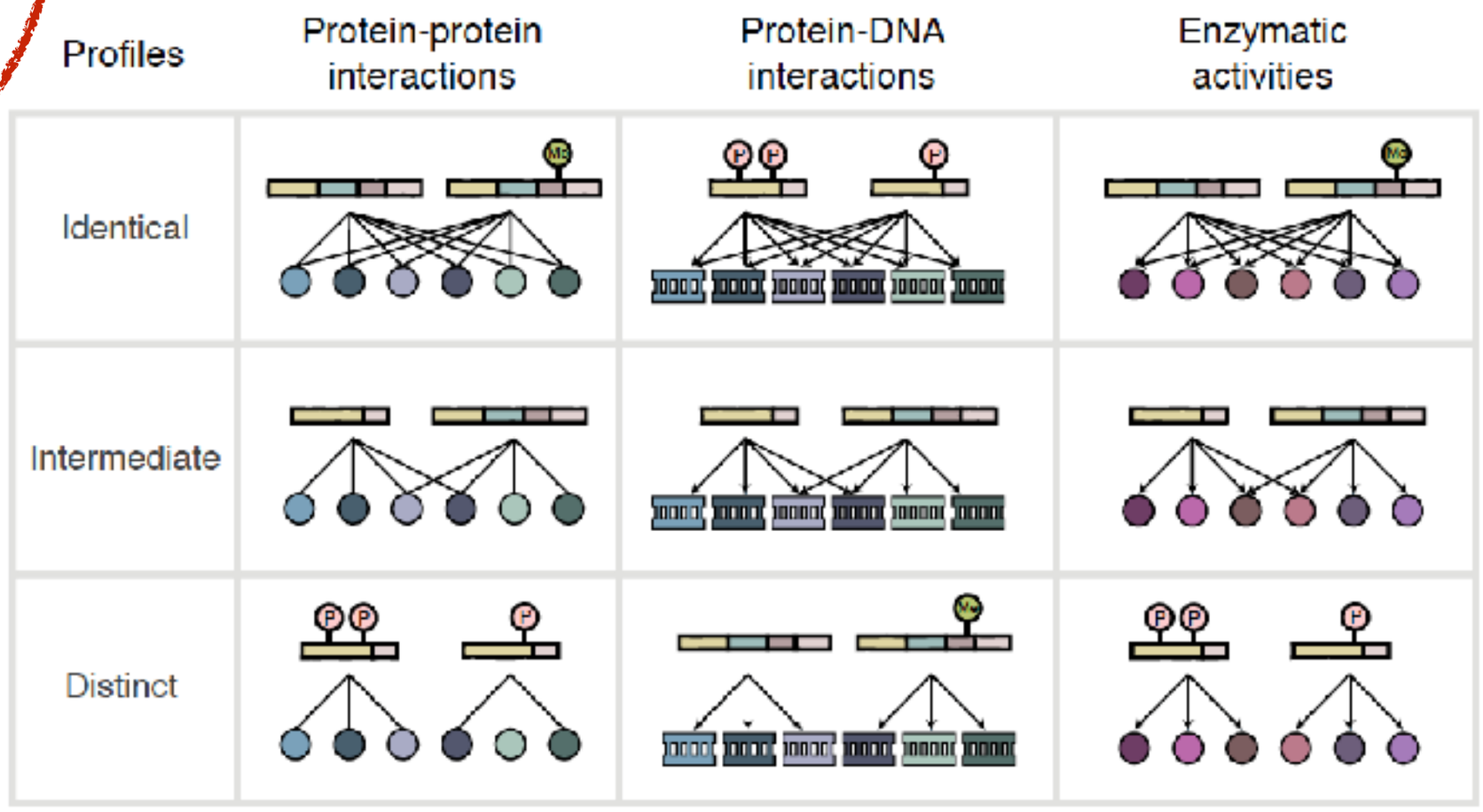
Enzymatic activities

Cellular localization

Stability

…..

Unbiased functional profiling

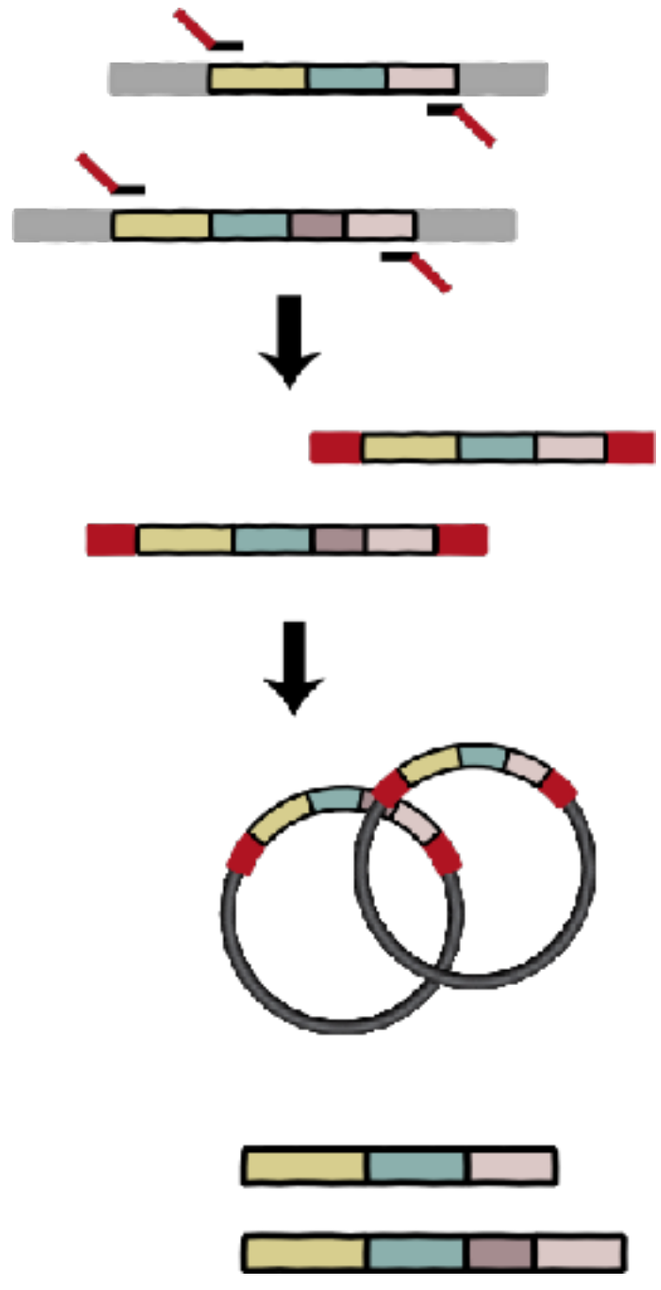


ORF-Seq:
Systematic full-length isoform sequencing and cloning

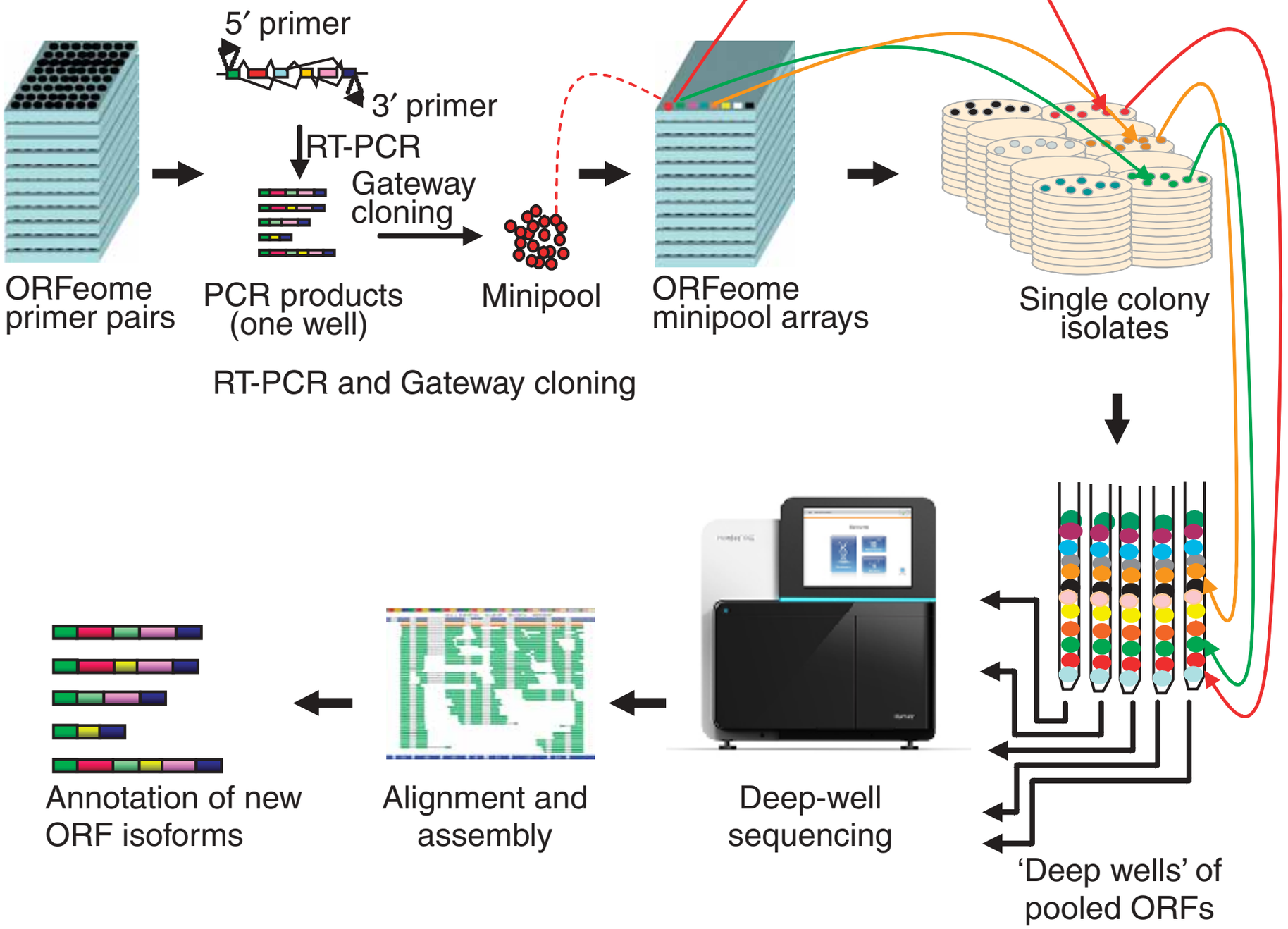
Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing

Kourosh Salehi-Ashtiani[1,2,5], Xinping Yang[1,2,5], Adnan Derti[1,3,5], Weidong Tian[1,3,5], Tong Hao[1,2,5], Chenwei Lin[1,2], Kathryn Makowski[4], Lei Shen[4], Ryan R Murray[1,2], David Szeto[1,2], Nadeem Tusneem[4], Douglas R Smith[4], Michael E Cusick[1,2], David E Hill[1,2], Frederick P Roth[1,3] & Marc Vidal[1,2]

NATURE METHODS | VOL.5 NO.7 | JULY 2008

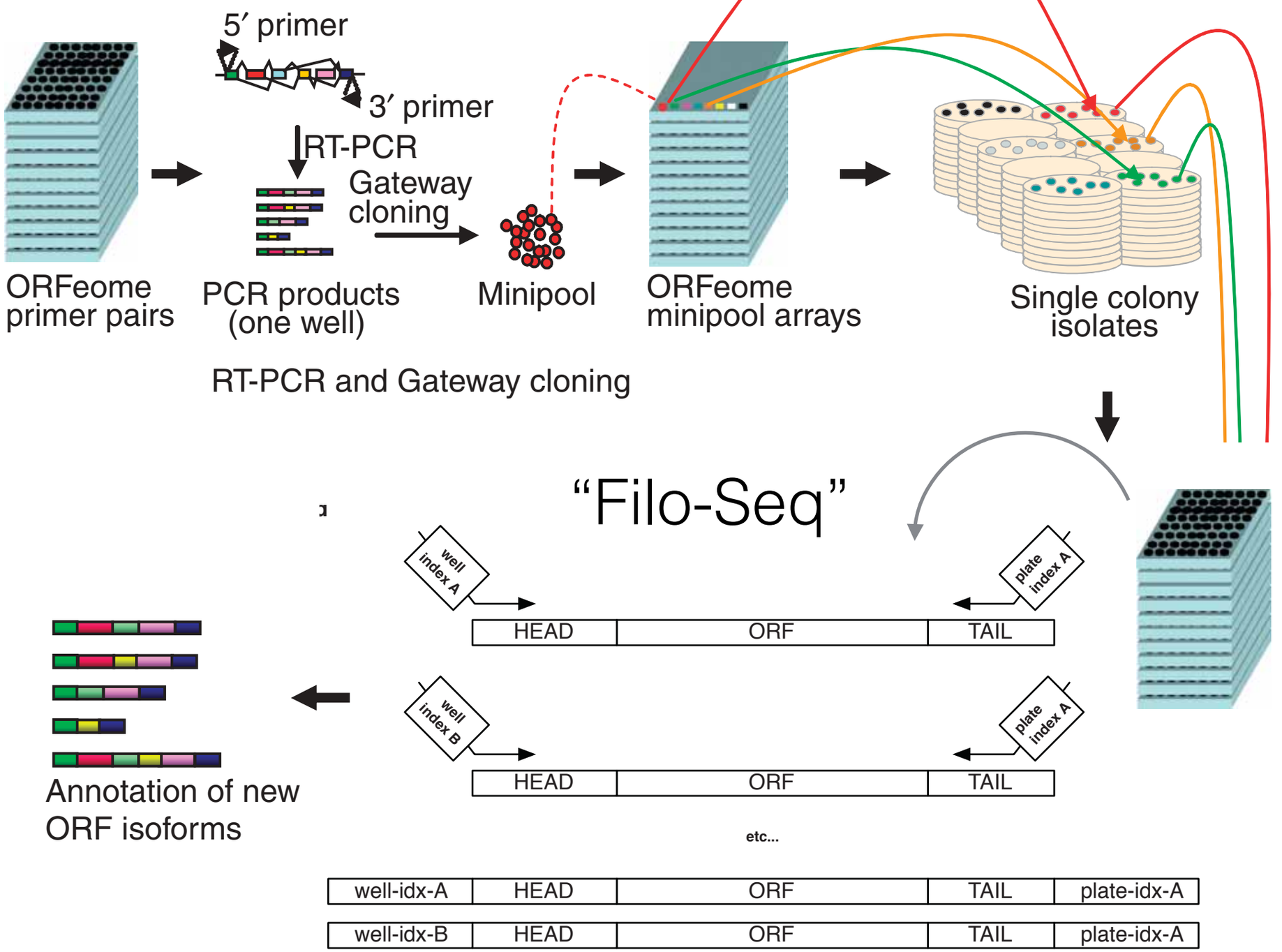


ORF-Seq:
Systematic full-length isoform sequencing and cloning

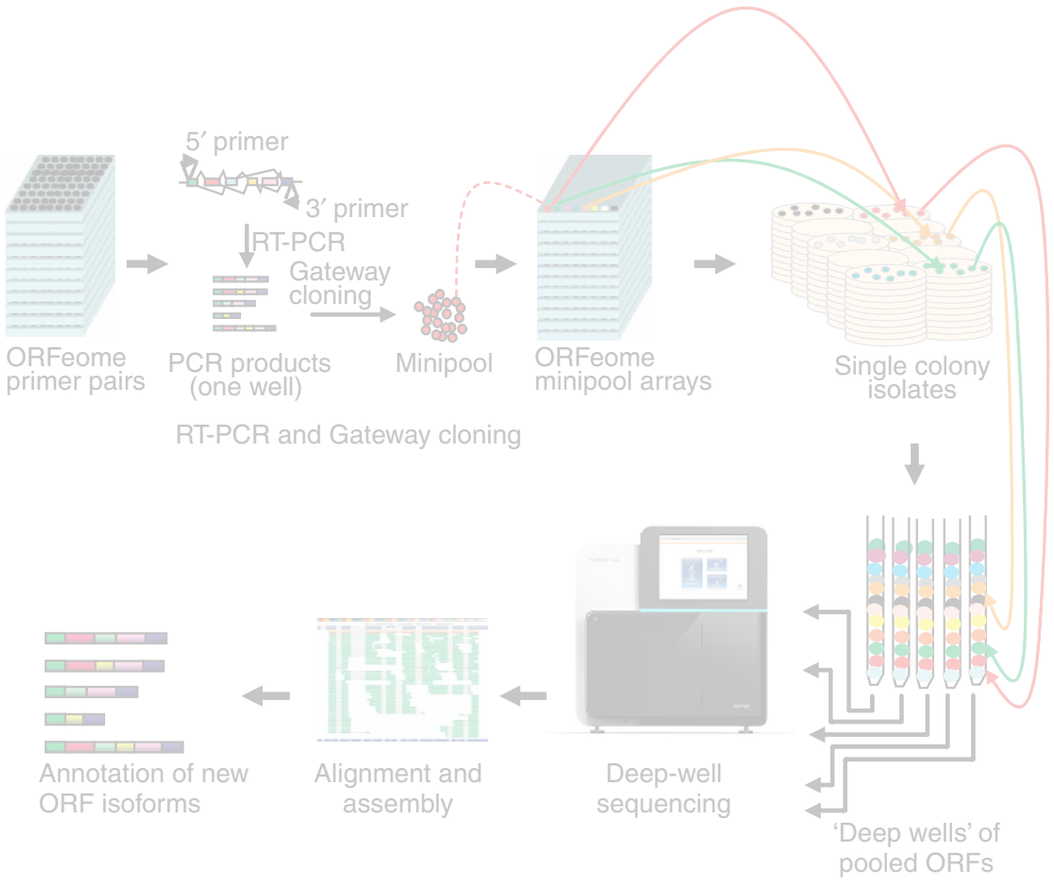


Original ORF-Seq: Salehi-Ashtiani *et al Nat Meth* 2008

ORF-Seq:
Systematic full-length isoform sequencing and cloning



ORF-Seq:
Systematic full-length isoform sequencing and cloning



~1,500 genes, 5 human tissues

1,423 full-length isoforms

506 genes with ≥ 2 isoforms

1677 pairs of alternative isoforms, each encoded by a common gene

Examples of isoforms

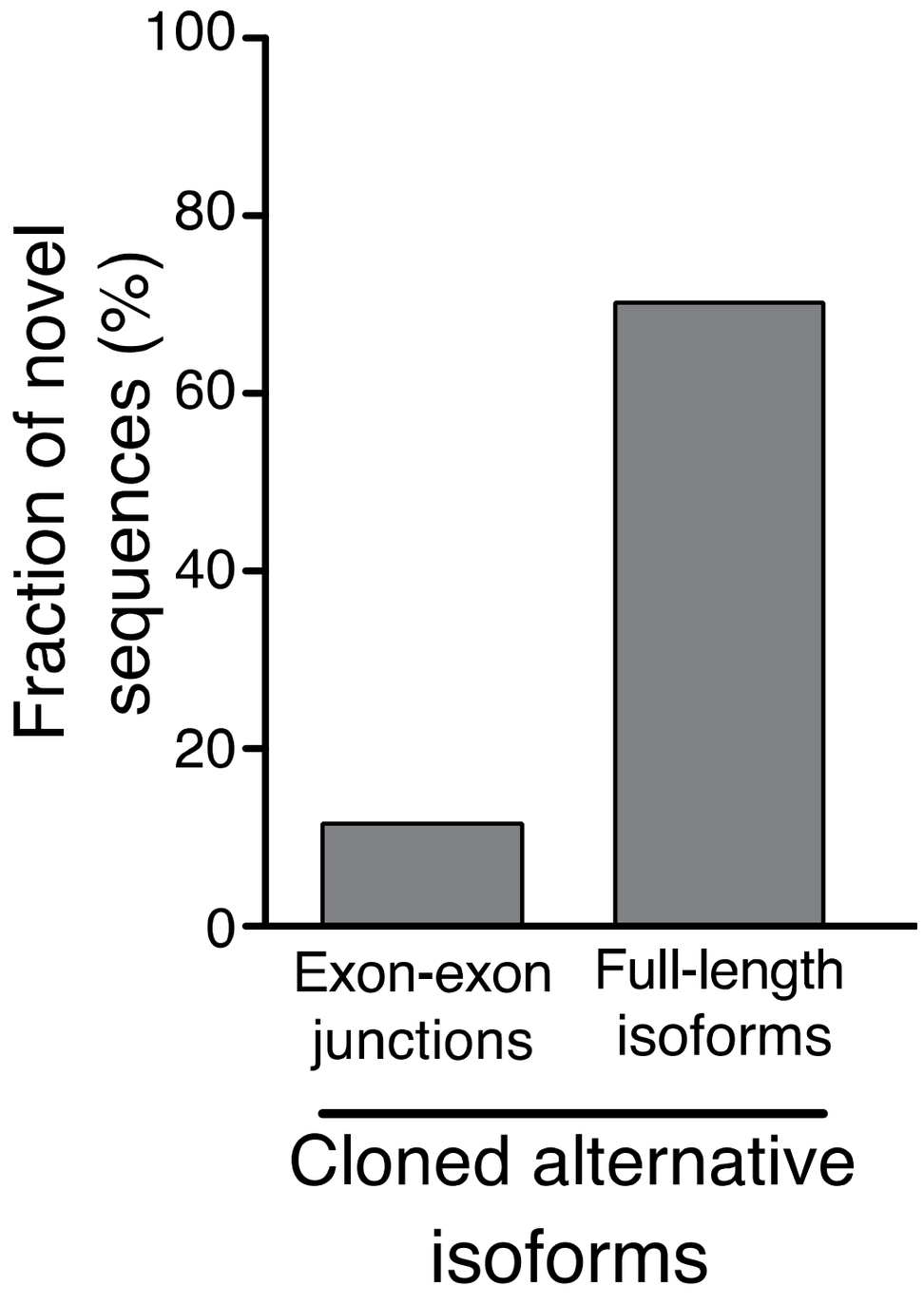
BTC

CD99L2

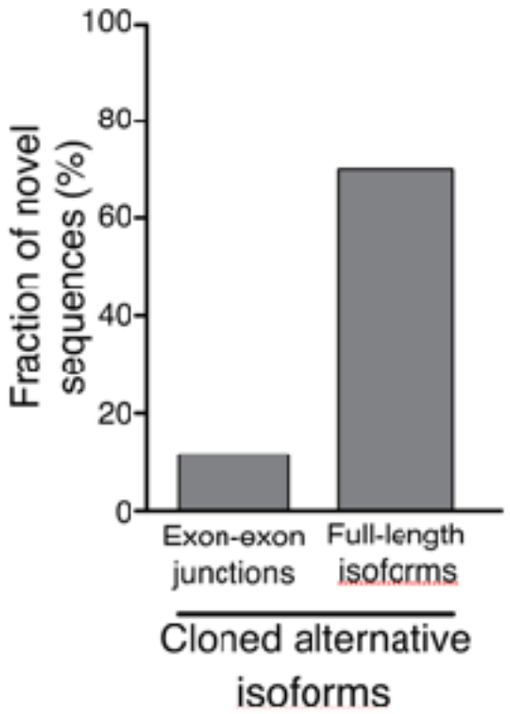
BANP

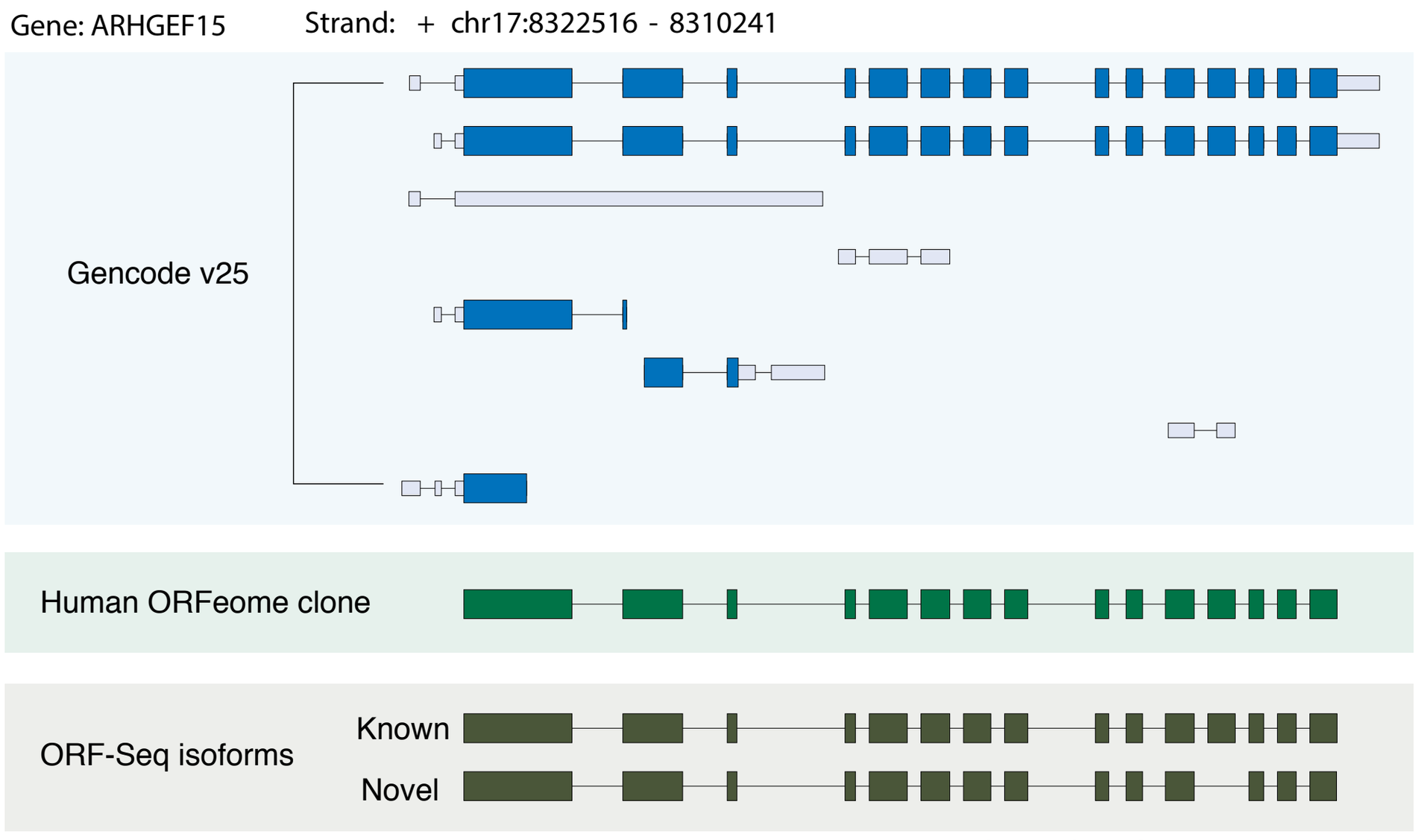
Novel full-length isoforms

506 genes with ≥ 2 isoforms



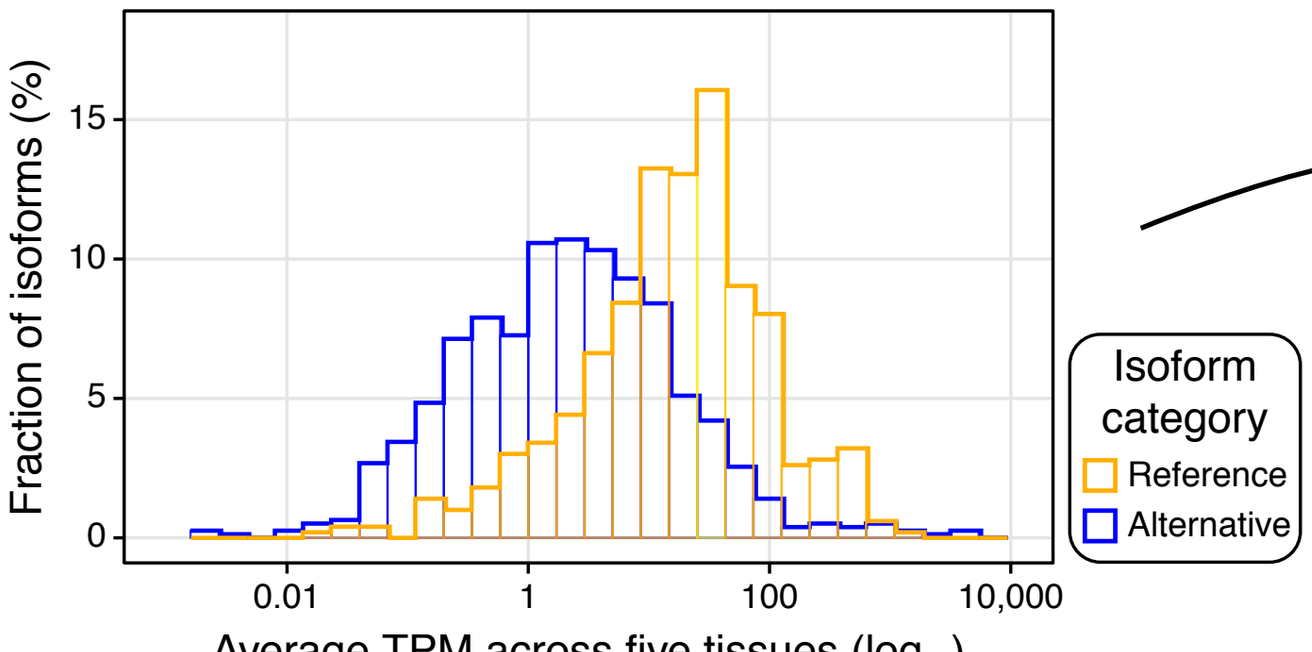
Novel full-length isoforms

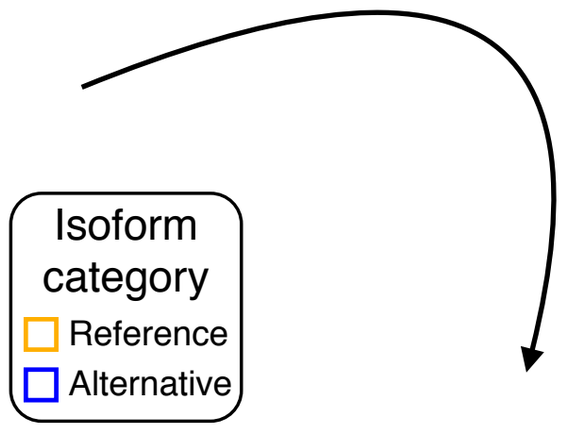




Intron lengths are not to scale.

Estimated isoform clone expression in the original, endogenous tissues





How widespread is isoform functional divergence in the whole proteome?

Systematic identification
of large numbers of
isoform pairs
for large numbers
of human genes

↓

Physical interactions

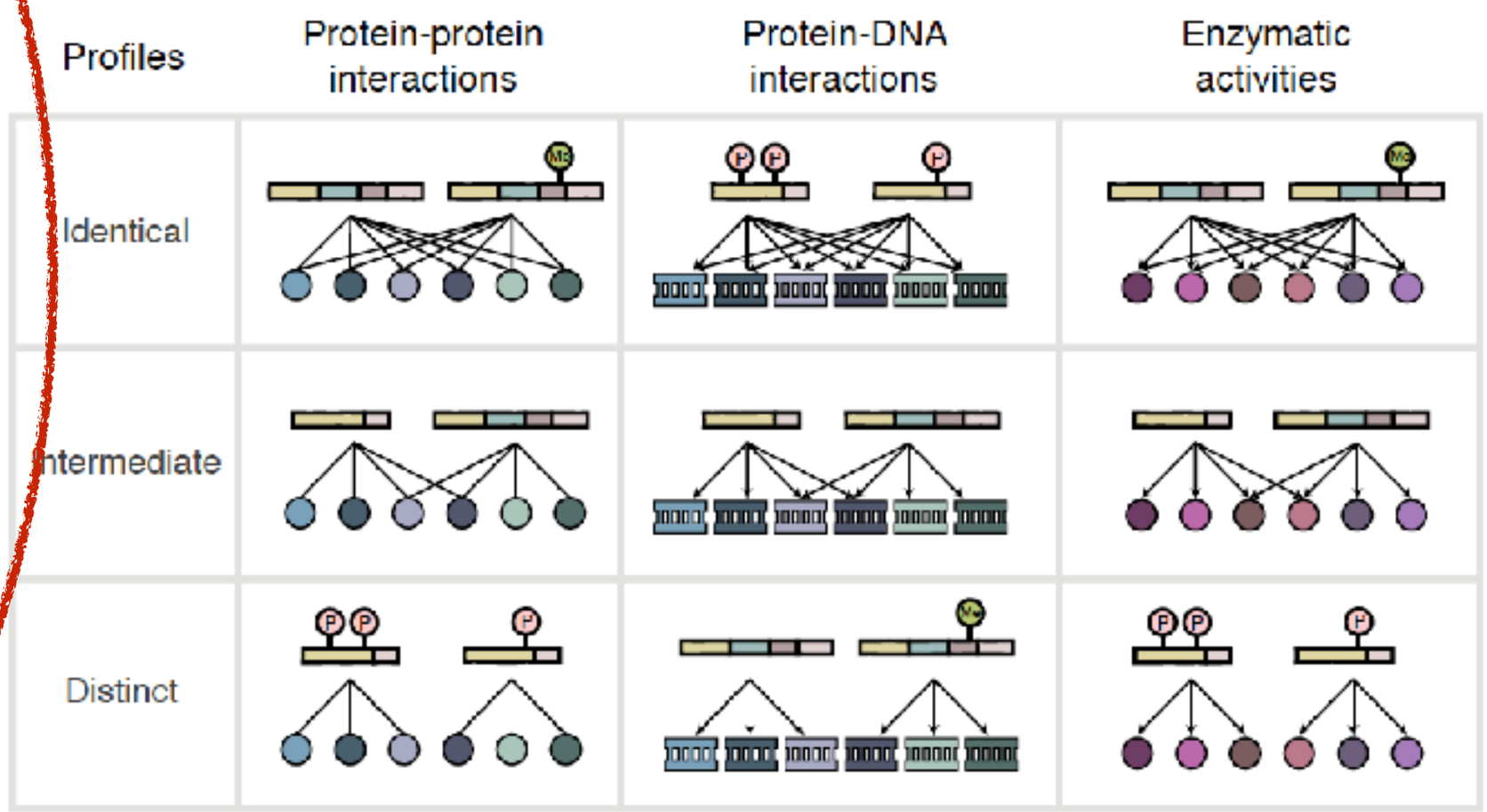
Enzymatic activities

Cellular localization

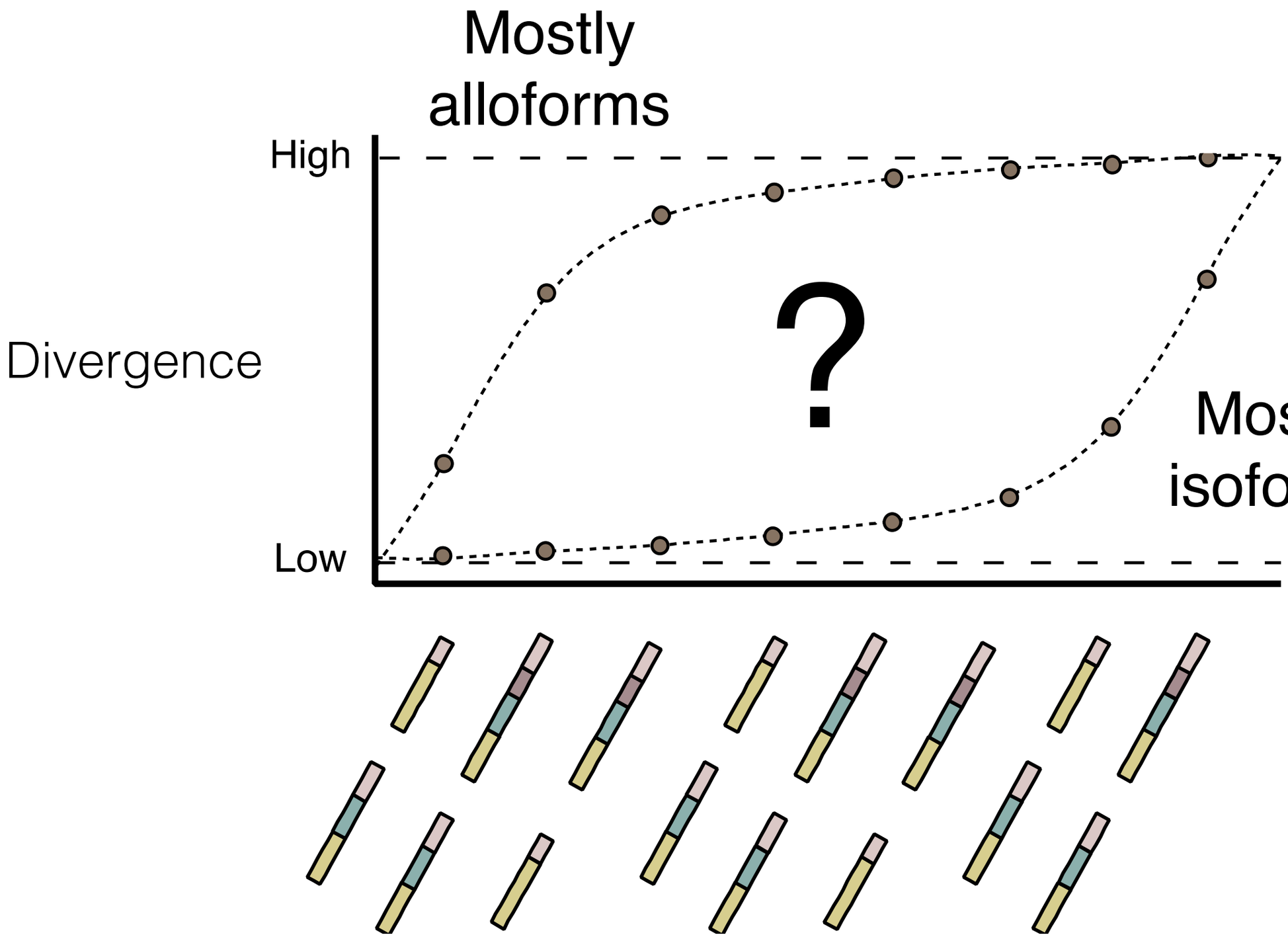
Stability

…..

Unbiased functional profiling

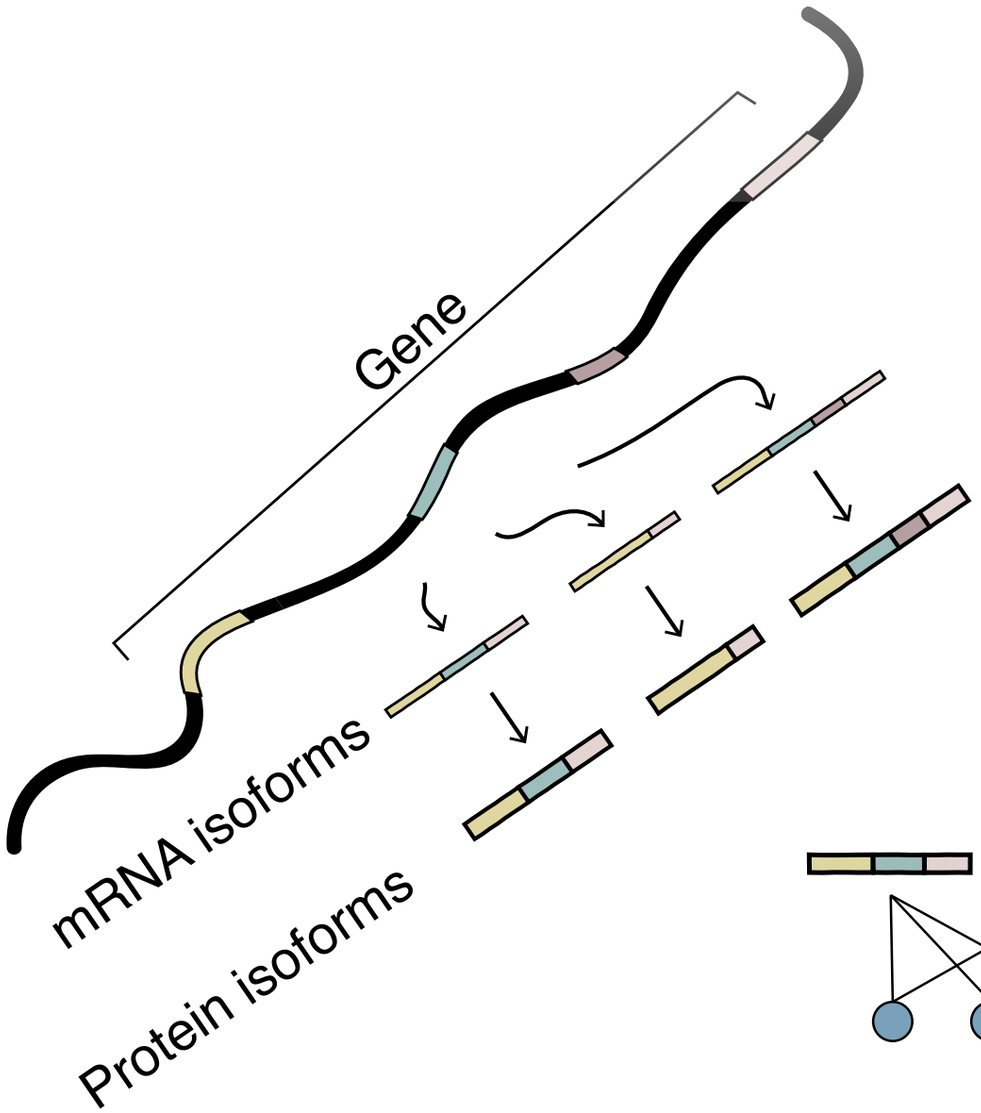


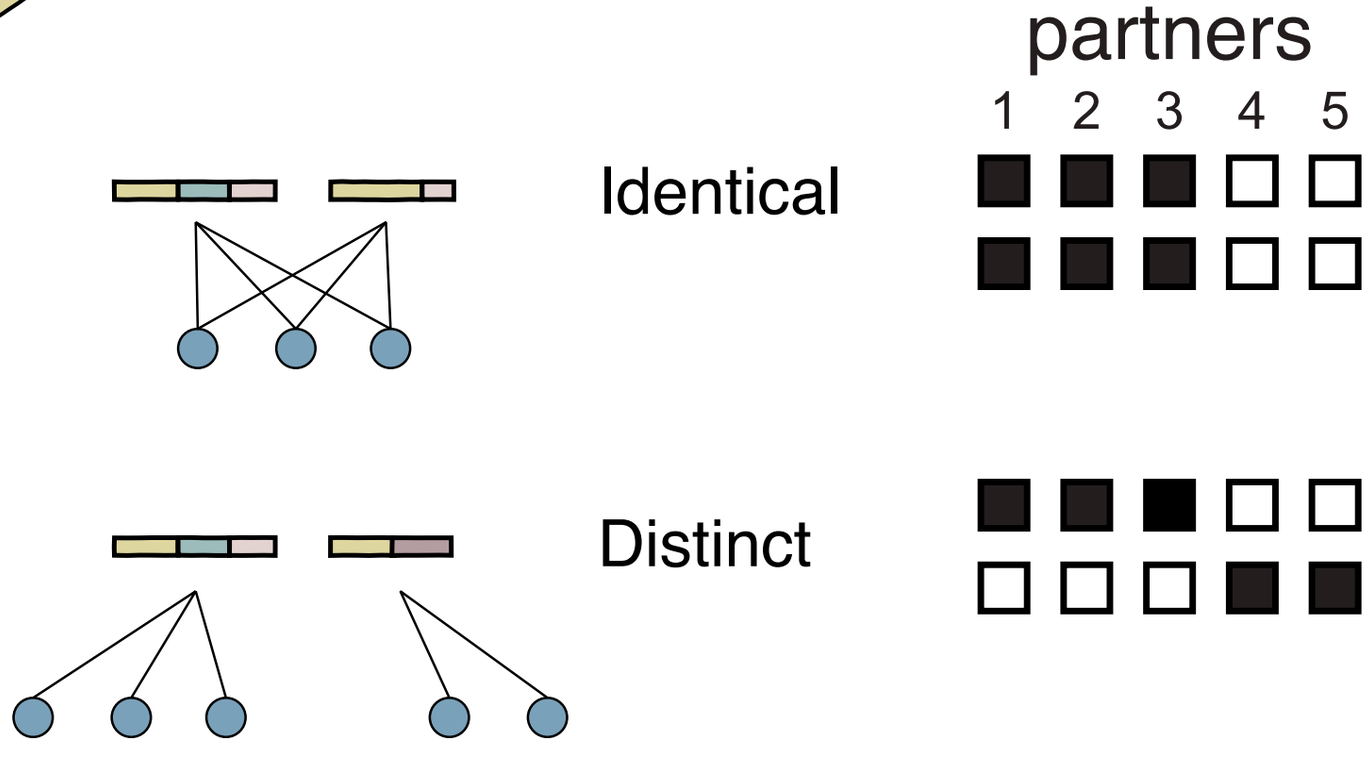
Landscape of protein isoform functional divergence



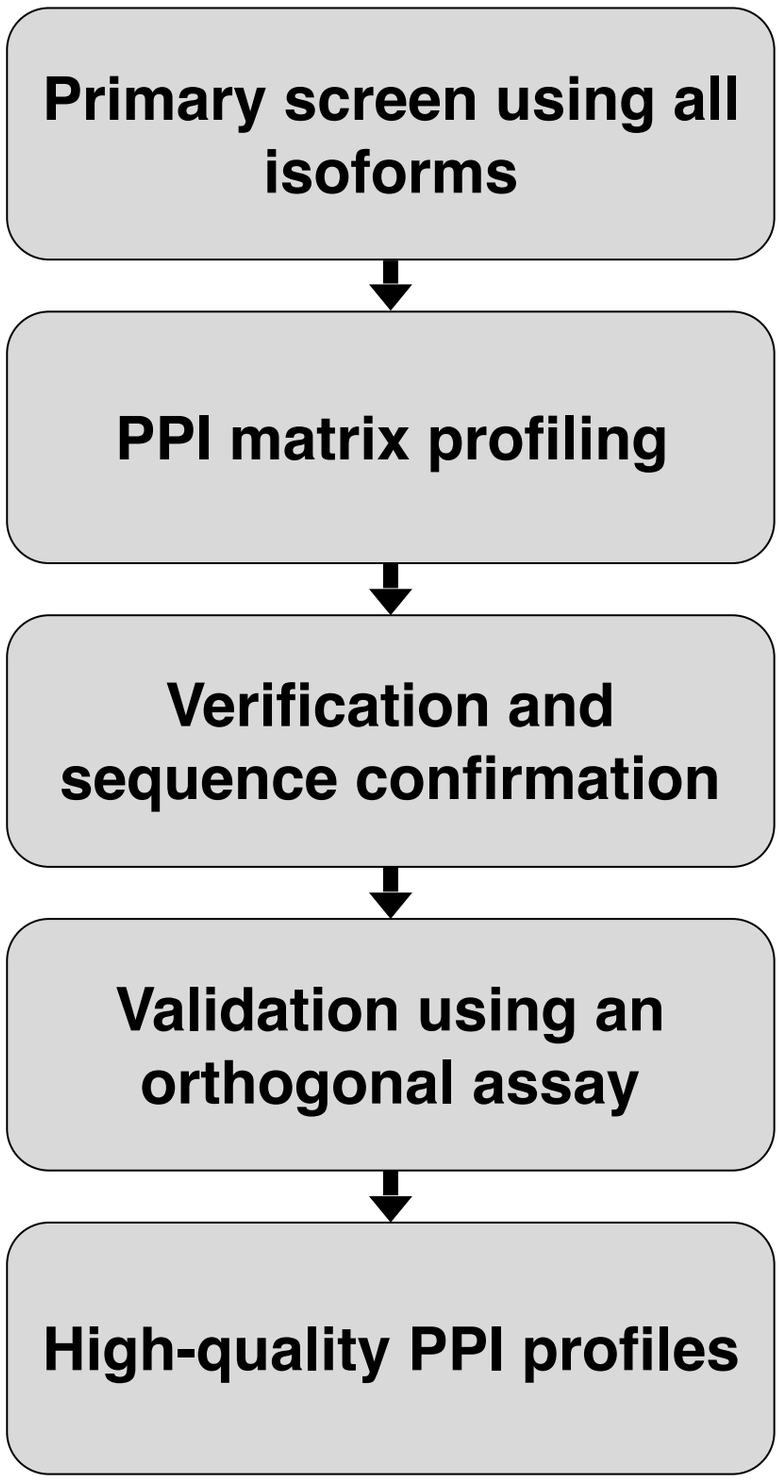
Large numbers of pairs of isoforms encoded by common genes

Comparative protein-protein interaction profiling for large numbers of isoform pairs





Comparative protein-protein interaction profiling for large numbers of isoform pairs



Comparative protein-protein interaction profiling for large numbers of isoform pairs

Primary screen using all isoforms

↓

PPI matrix profiling

↓

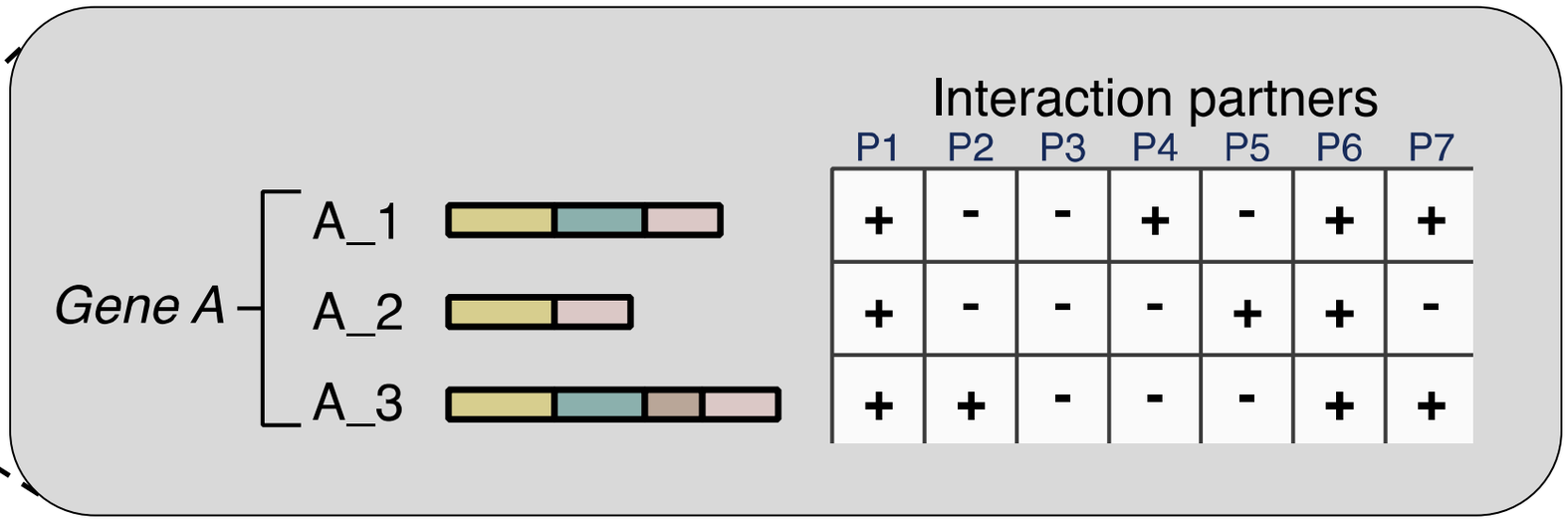
Verification and sequence confirmation

↓

Validation using an orthogonal assay

↓

High-quality PPI profiles



All first-pass pairs identified for any isoform were tested against all isoforms from a common gene.

Comparative protein-protein interaction profiling for large numbers of isoform pairs

Primary screen using all isoforms

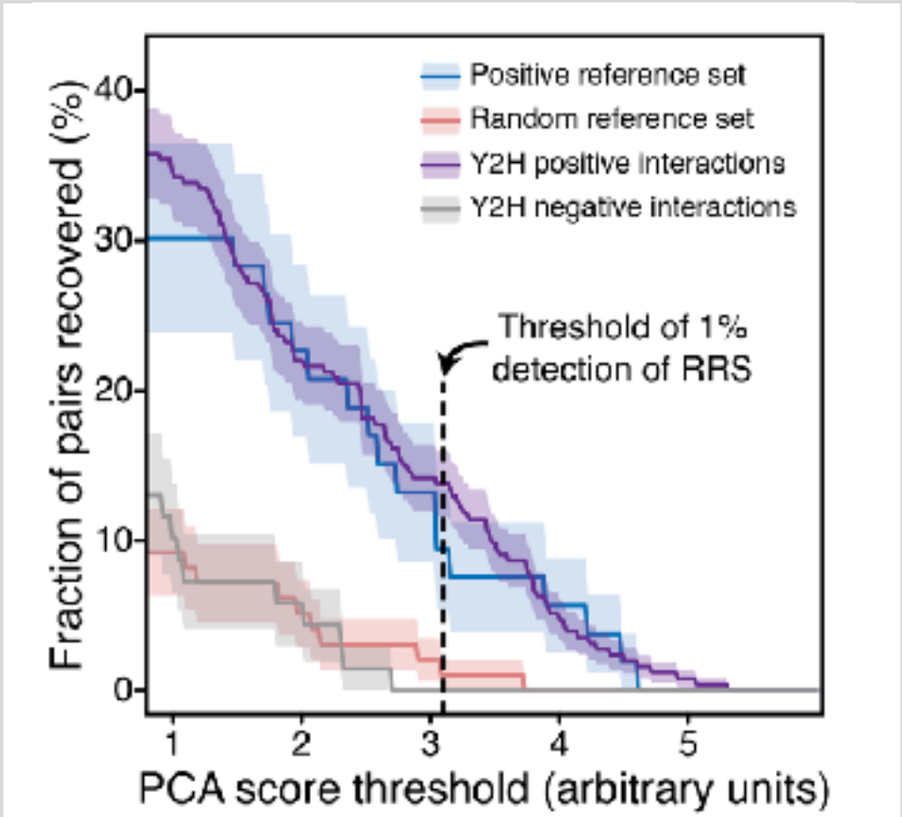
PPI matrix profiling

Verification and sequence confirmation

Validation using an orthogonal assay

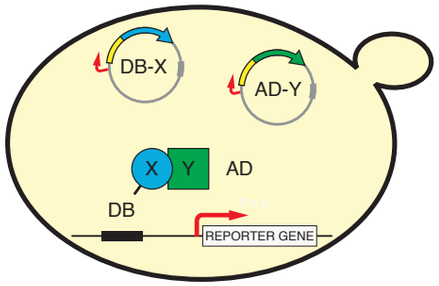
High-quality PPI profiles

Mammalian protein complementation assay

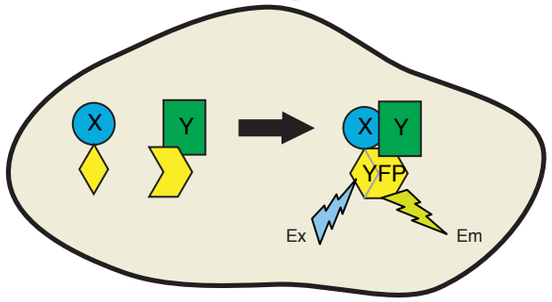


Validation

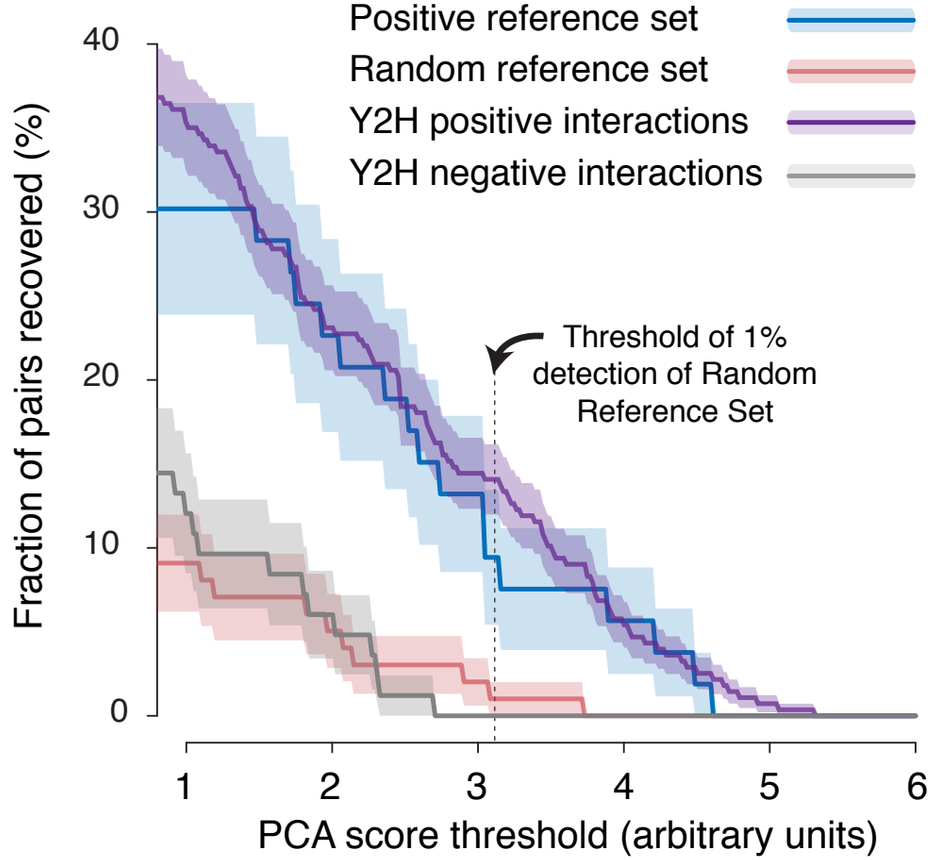
Validation using an orthogonal assay (PCA)



Two-hybrid
In yeast cells



PCA
In mammalian cells



Comparative protein-protein interaction profiling for large numbers of isoform pairs

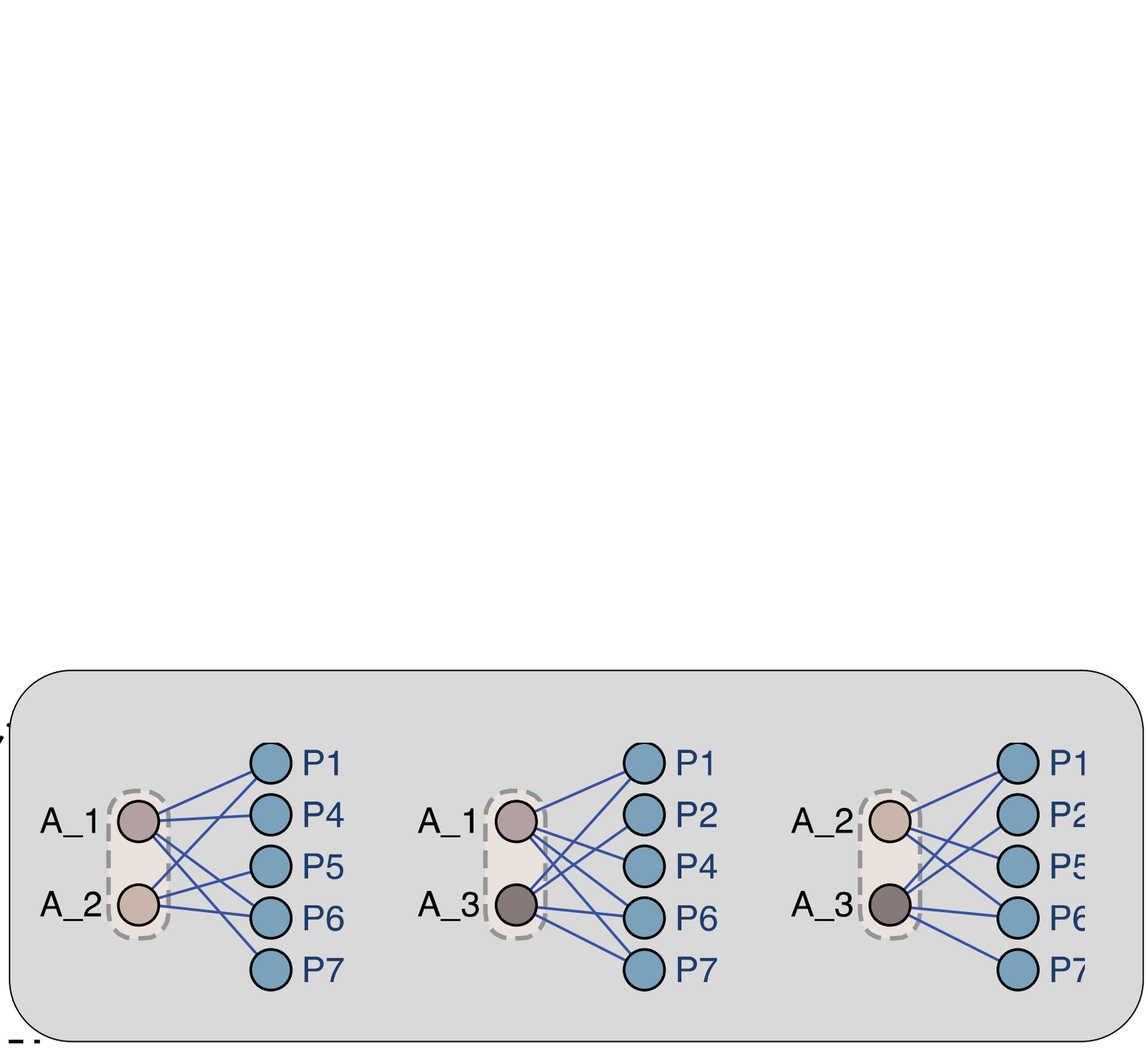
Primary screen using all isoforms

PPI matrix profiling

Verification and sequence confirmation

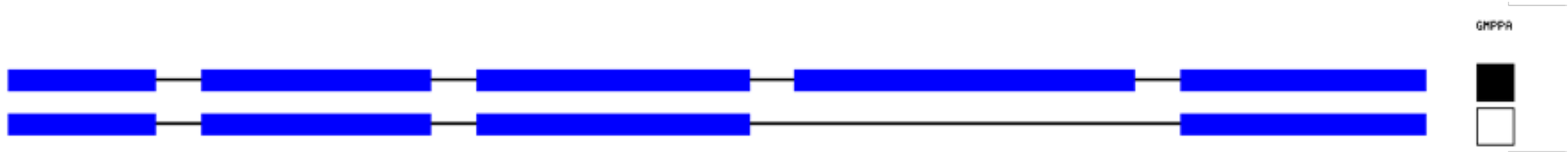
Validation using an orthogonal assay

High-quality PPI profiles



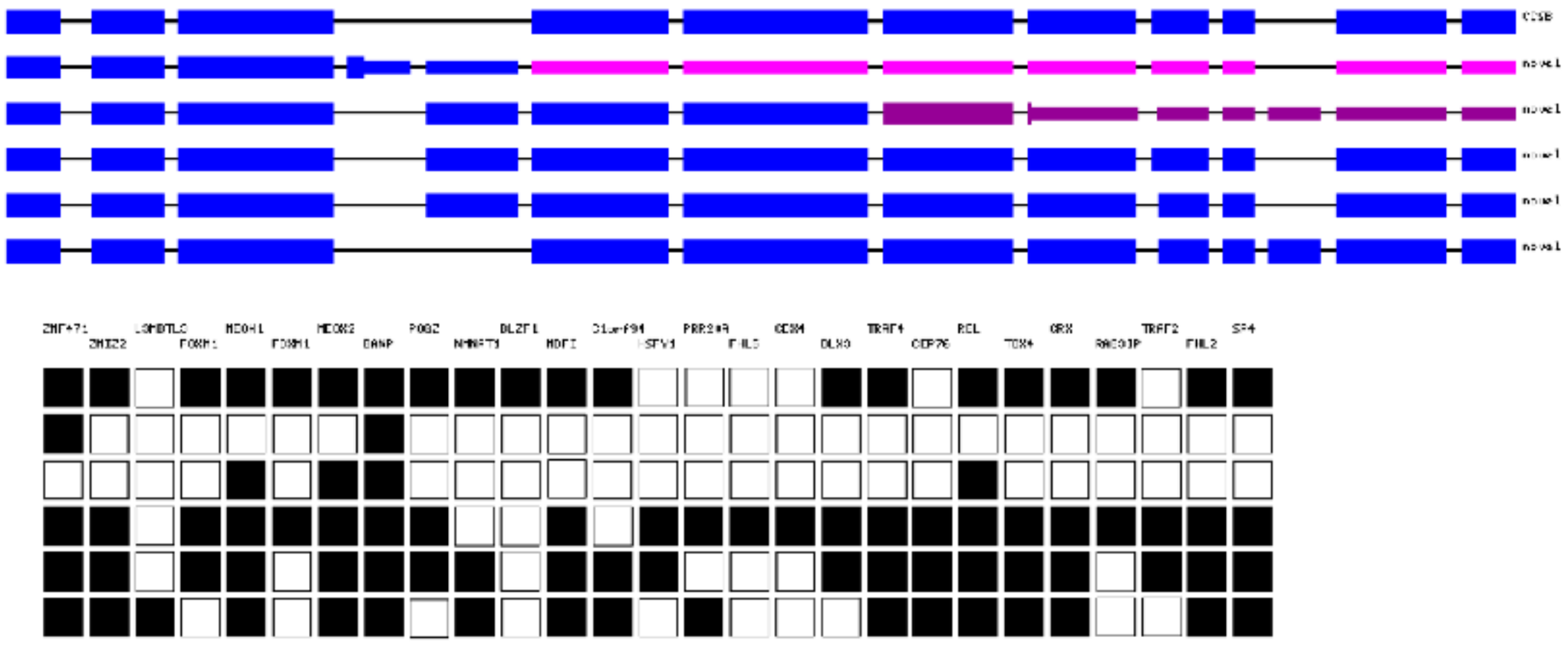
Isoform-specific interaction profiles

BTC



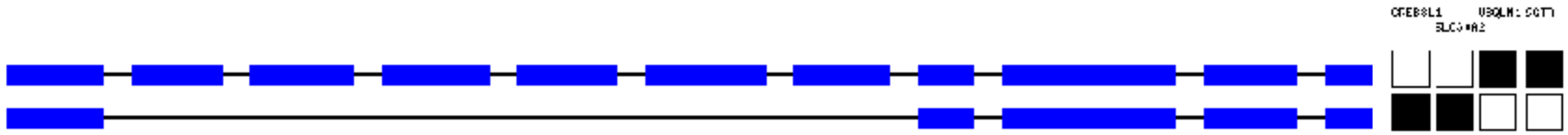
CD99L2

BANP

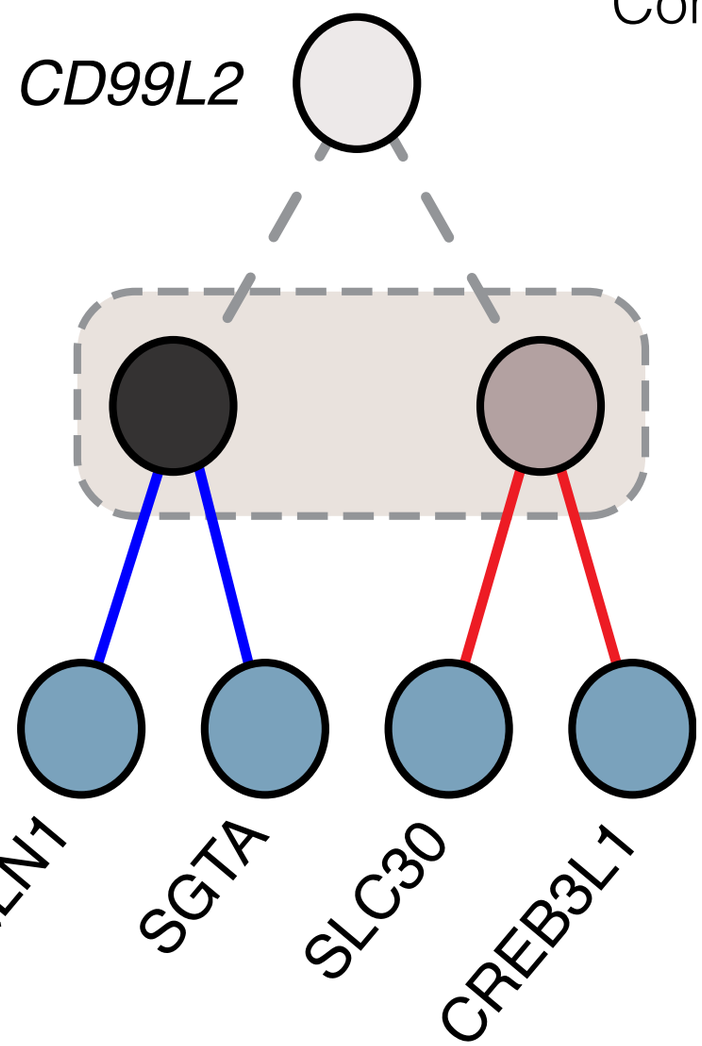


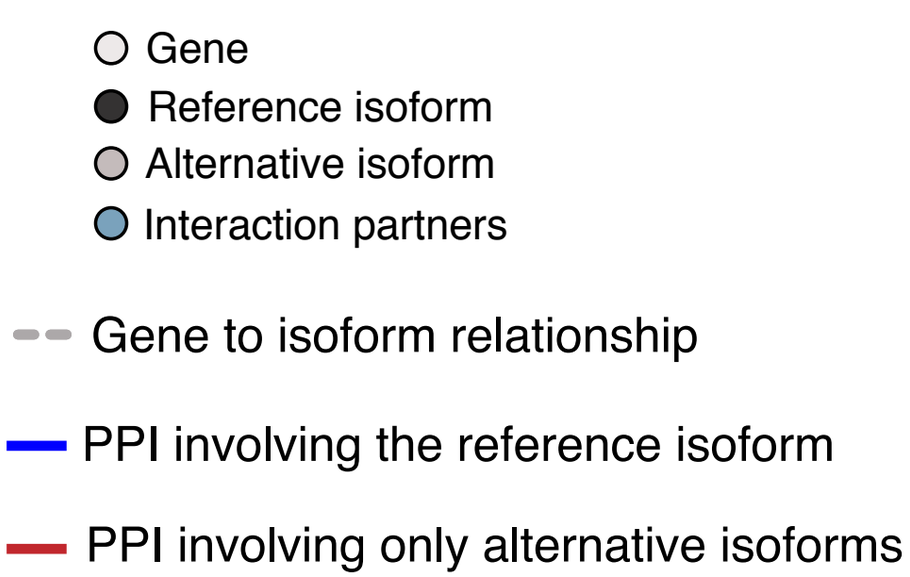
Isoform-specific interaction network

CD99L2 isoforms

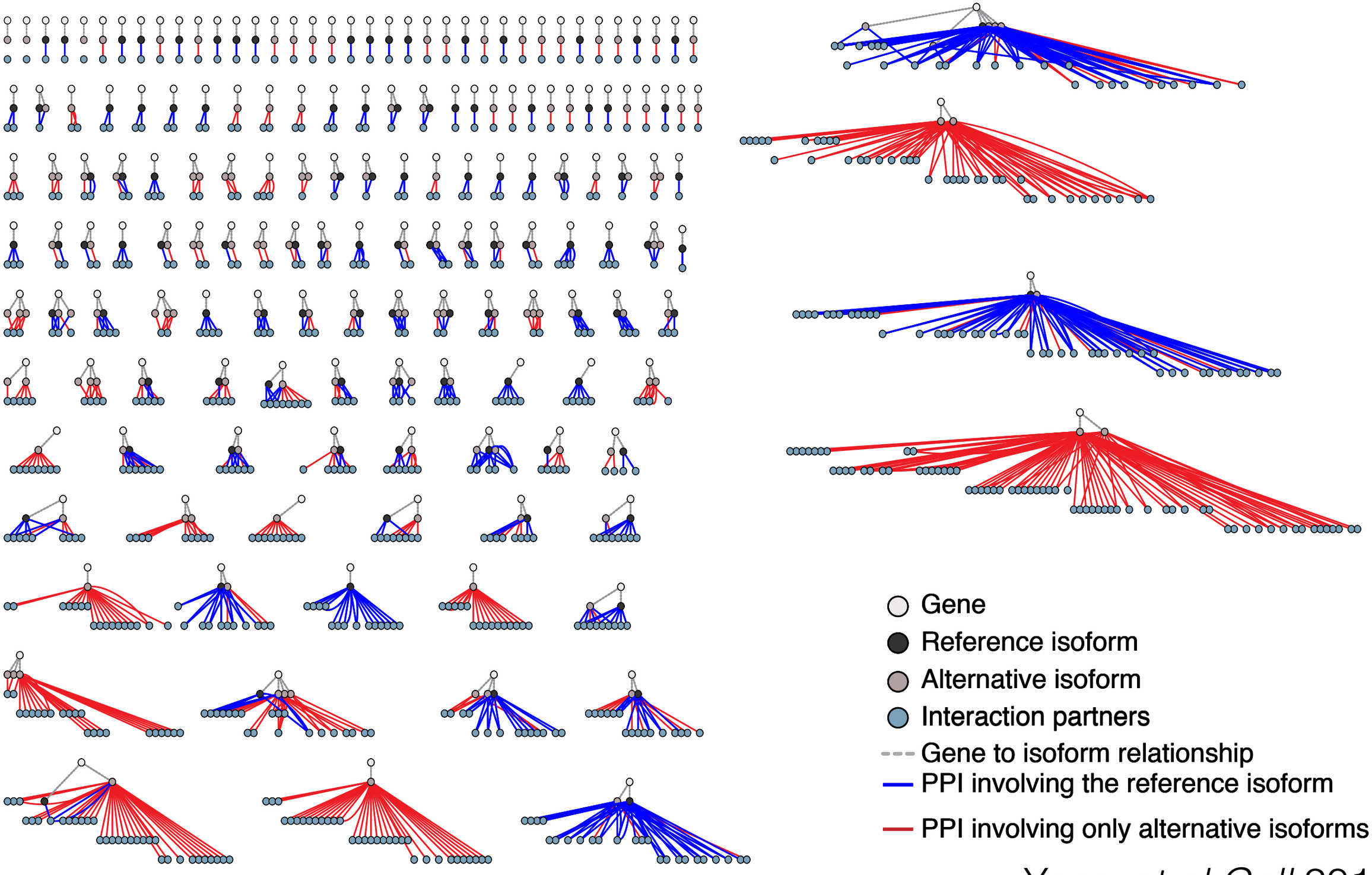


Convert into nodes and edges





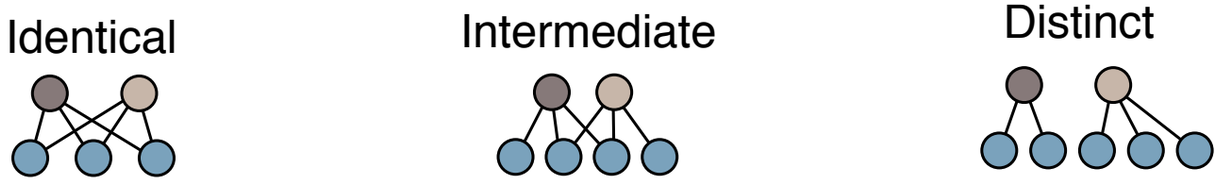
Isoform-specific interaction network



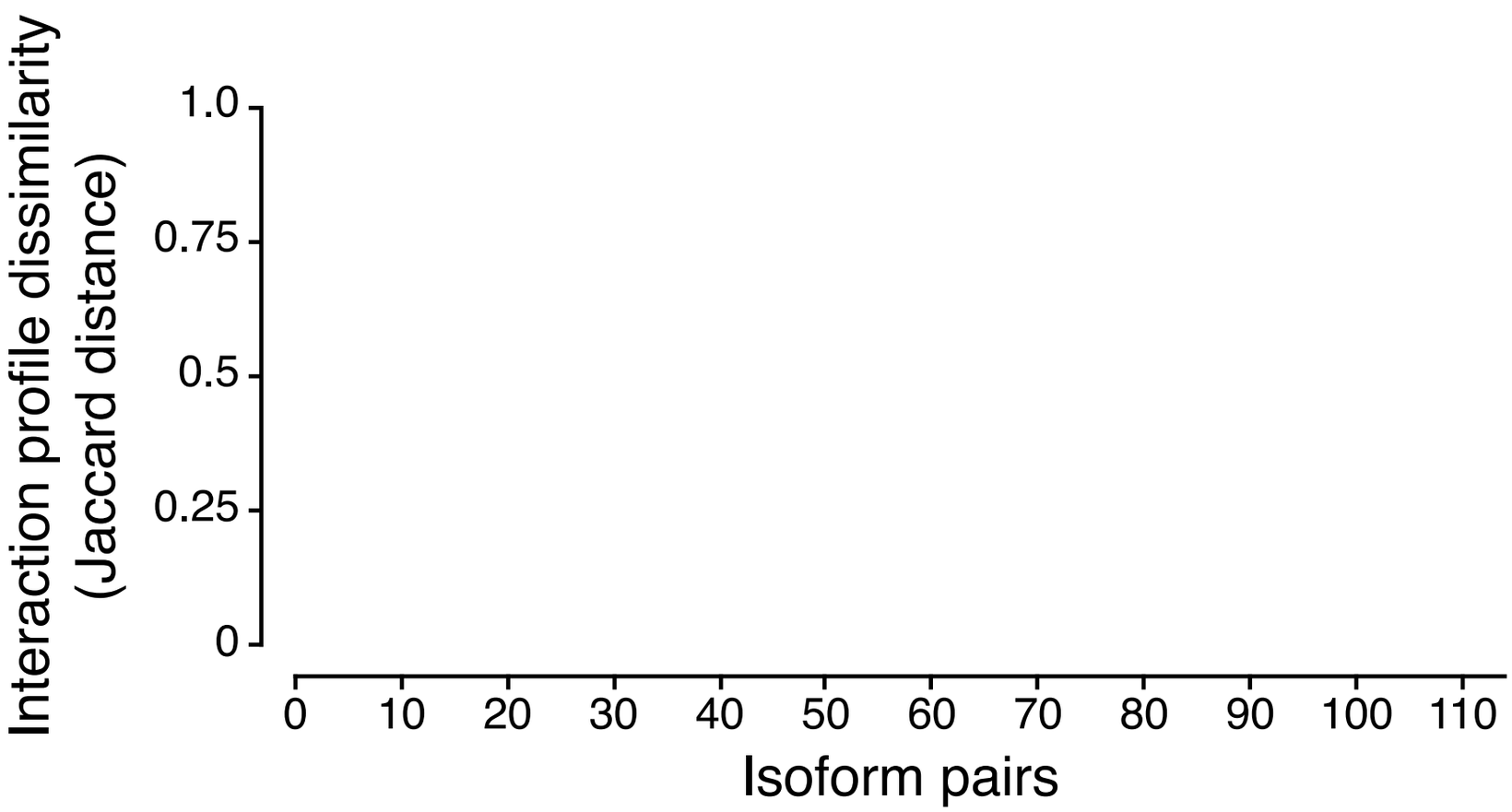
Yang *et al Cell* 2016

Interaction profile dissimilarity

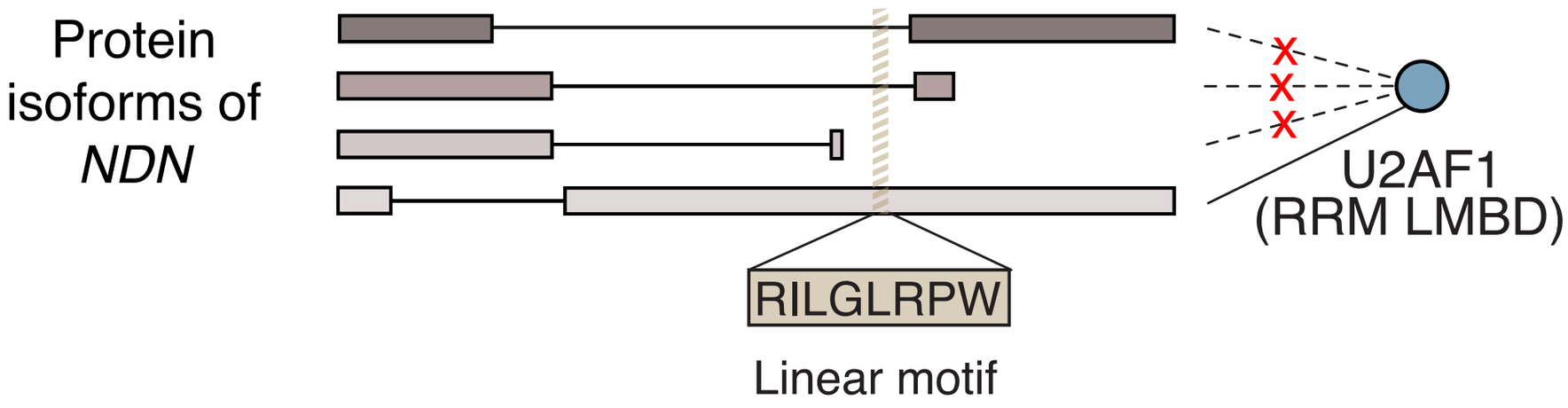
$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$



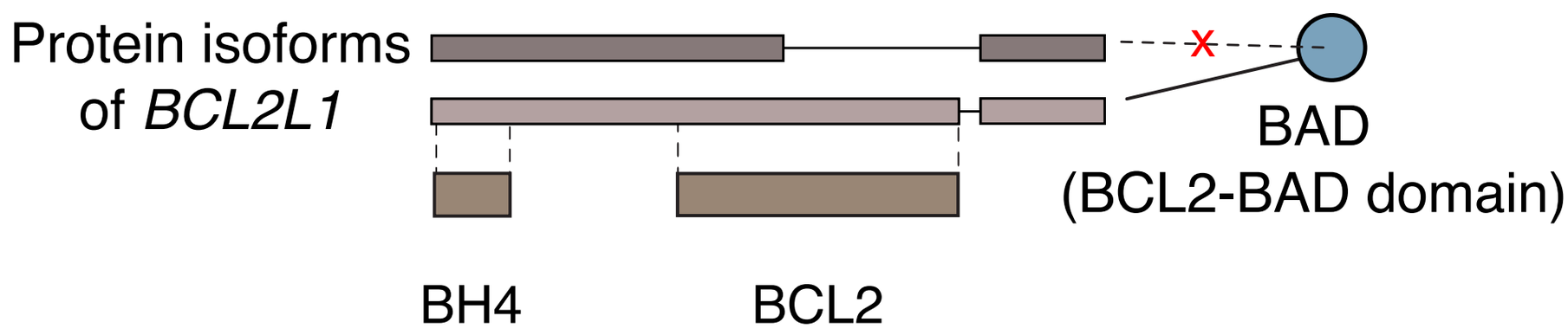
Jaccard distance:	0	0.5	1.0



Sequence features in isoforms underlie interaction perturbations



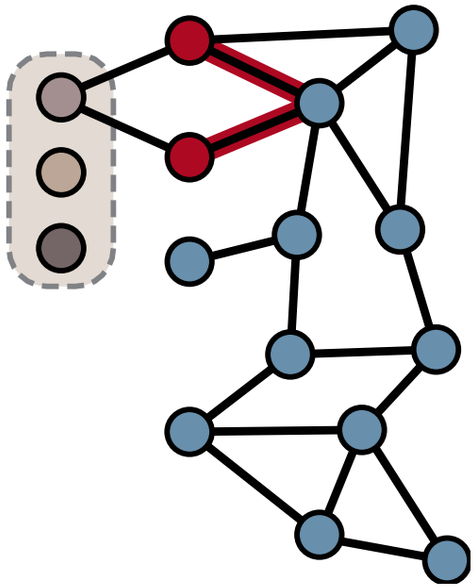
Isoforms with a high density of linear motifs tend to promote interactions (p-value = 5.7×10^{-4})



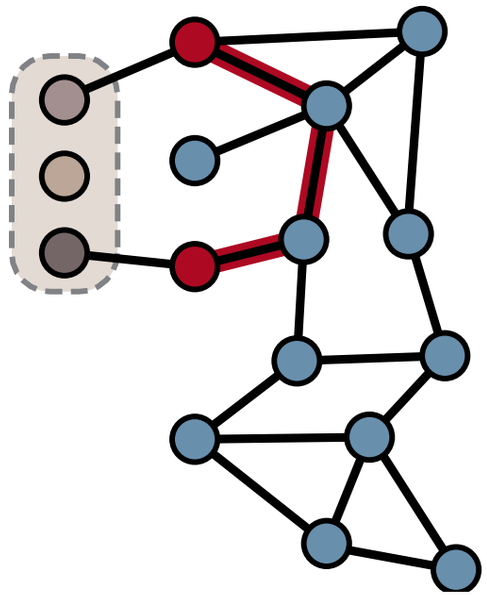
Truncated domains tend to correspond to a loss of interaction (p-value = 6.4×10^{-5})

How functionally divergent are isoform-specific interactors?

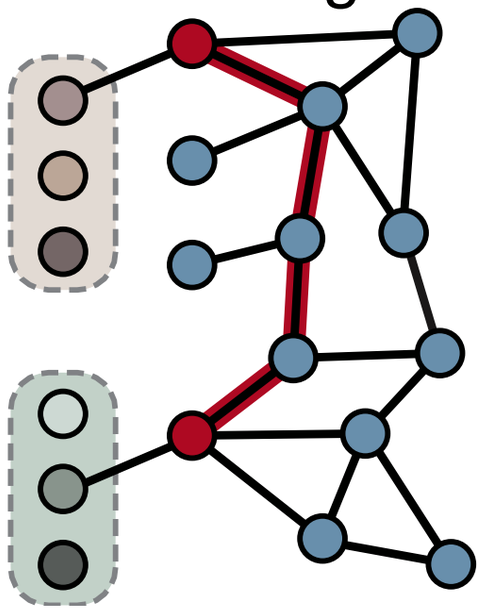
Single protein



Alternative isoforms

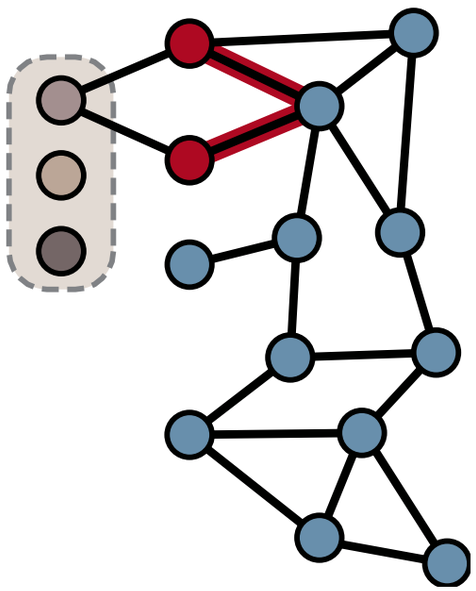


Proteins from different genes

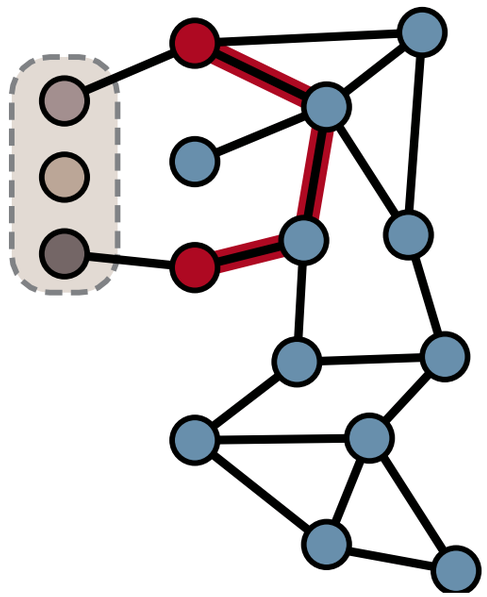


How functionally divergent are isoform-specific interactors?

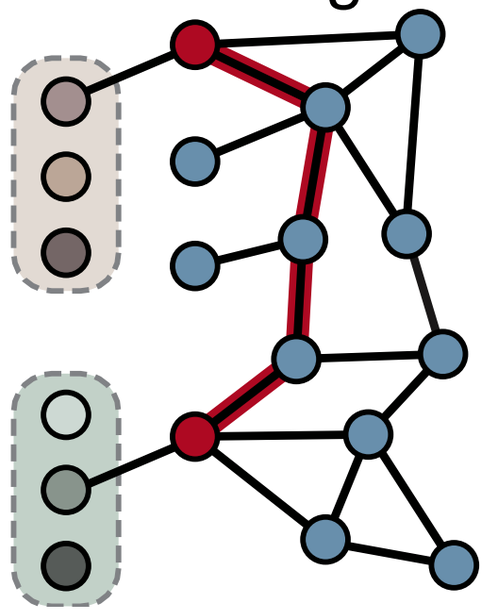
Single protein



Alternative isoforms



Proteins from different genes

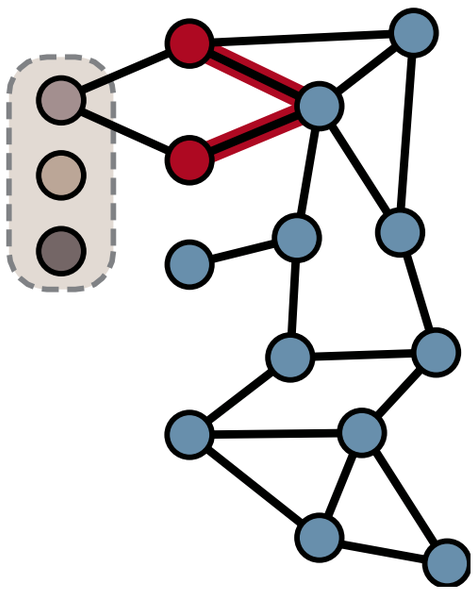


A Proteome-Scale Map of the Human Interactome Network

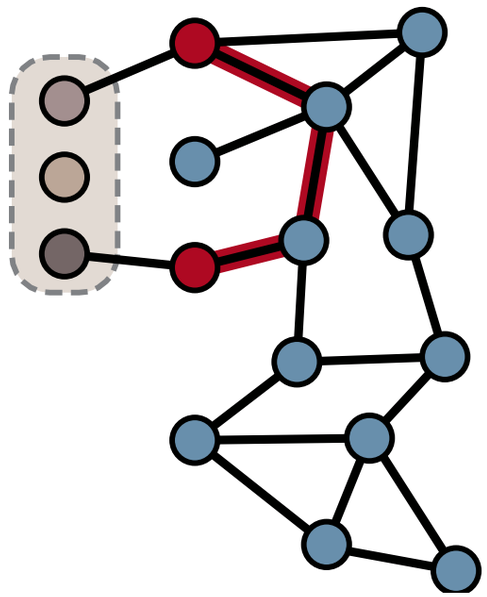
Rolland *et al Cell* 2014

How functionally divergent are isoform-specific interactors?

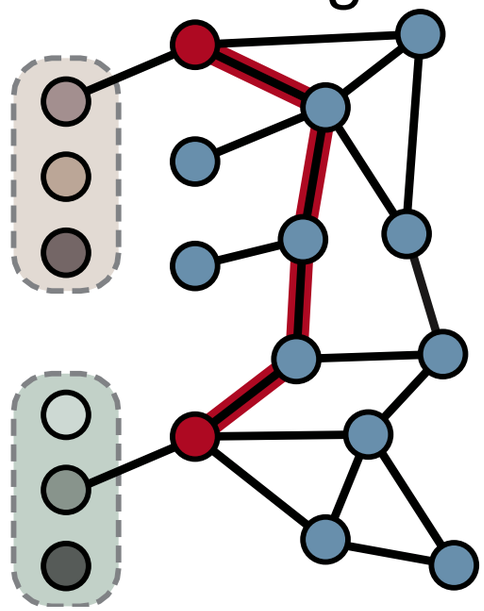
Single protein



Alternative isoforms



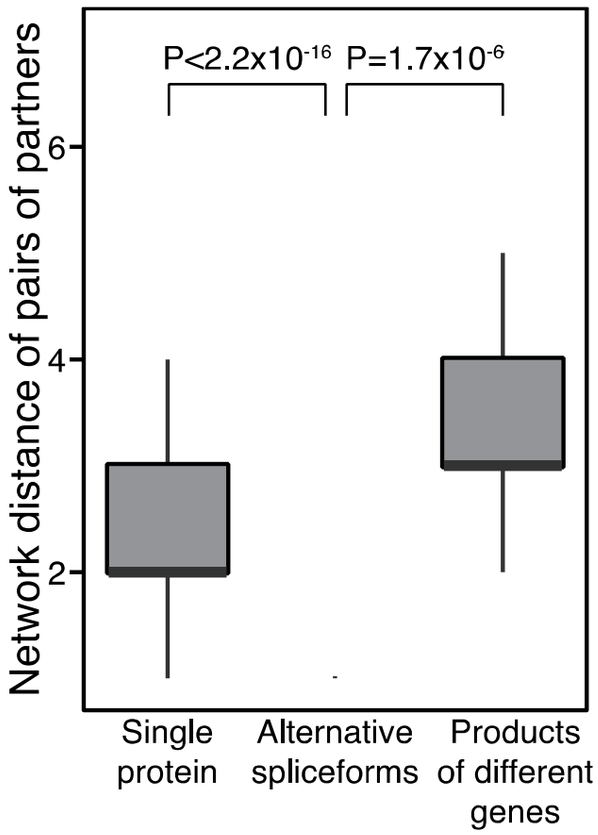
Proteins from different genes



A Proteome-Scale Map of the Human Interactome Network

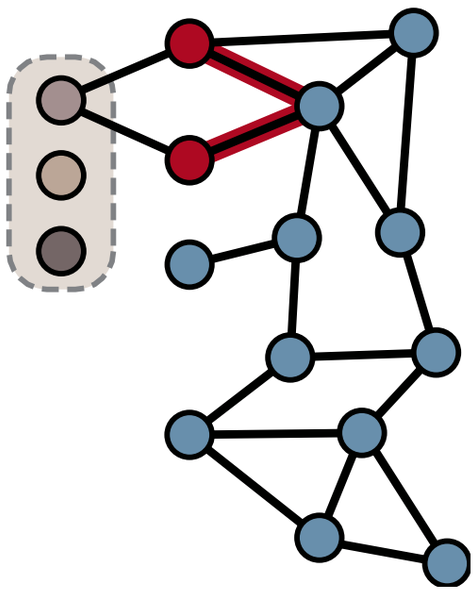
Rolland *et al Cell* 2014

How functionally divergent are isoform-specific interactors?

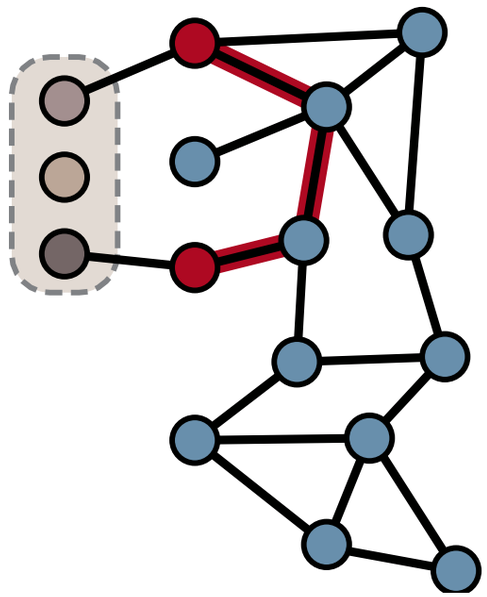


How functionally divergent are isoform-specific interactors?

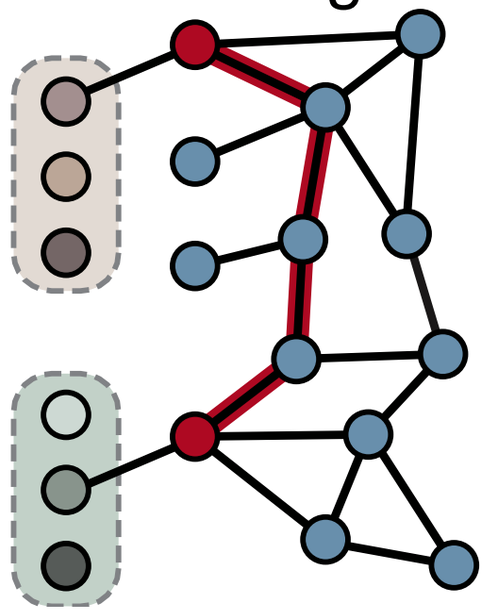
Single protein



Alternative isoforms



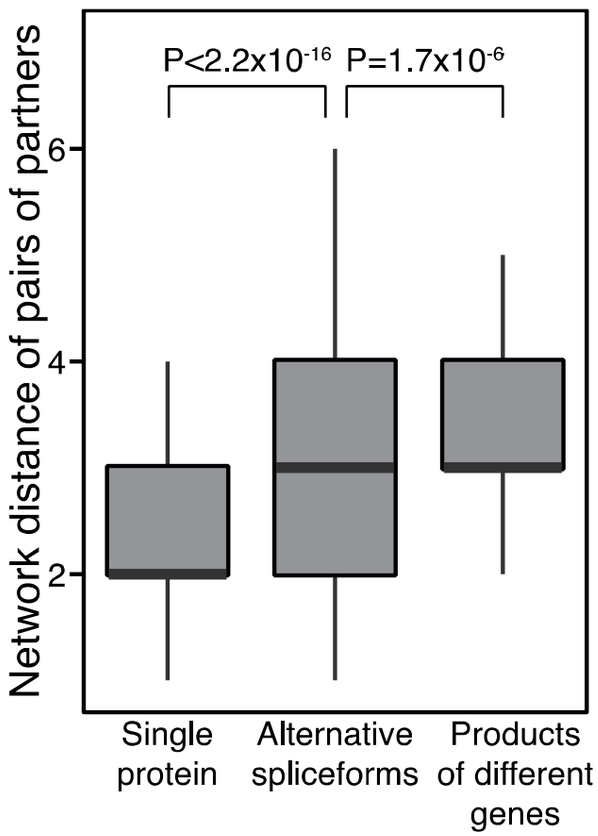
Proteins from different genes

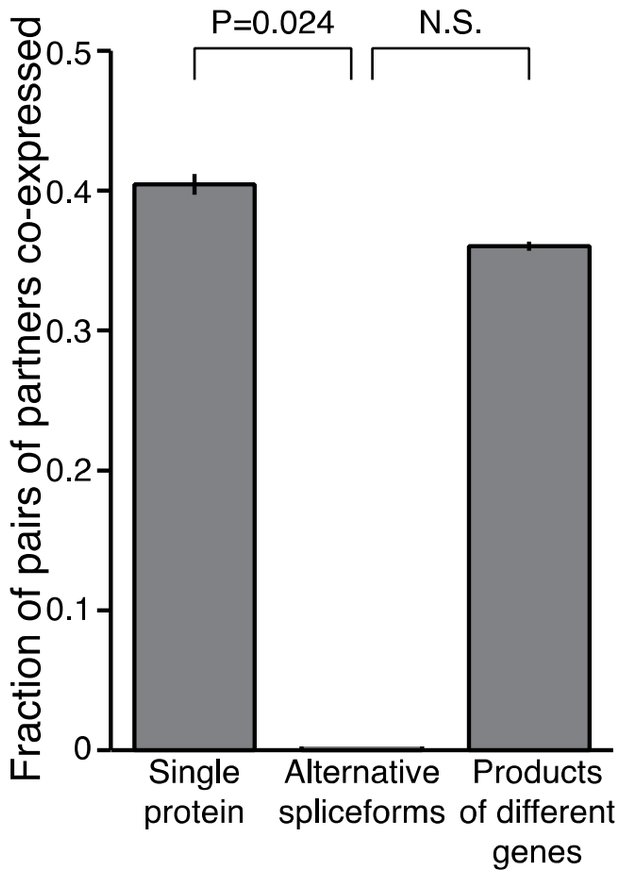


Illumina BodyMap 2.0
Gene expression for 16 human tissues

Reference

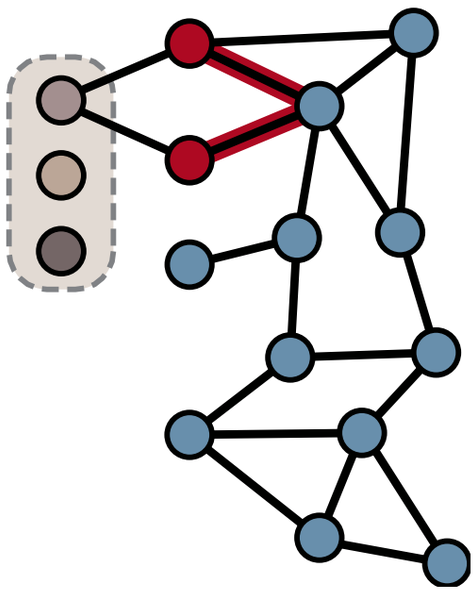
How functionally divergent are isoform-specific interactors?



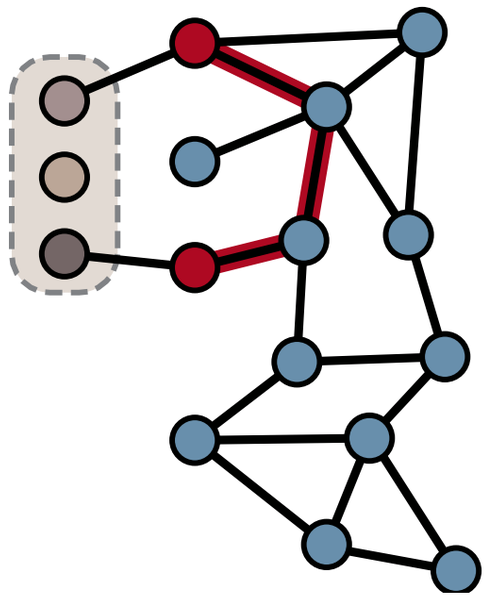


How functionally divergent are isoform-specific interactors?

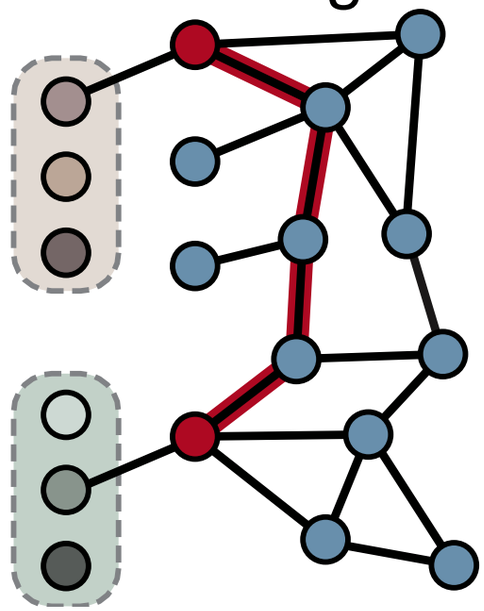
Single protein



Alternative isoforms



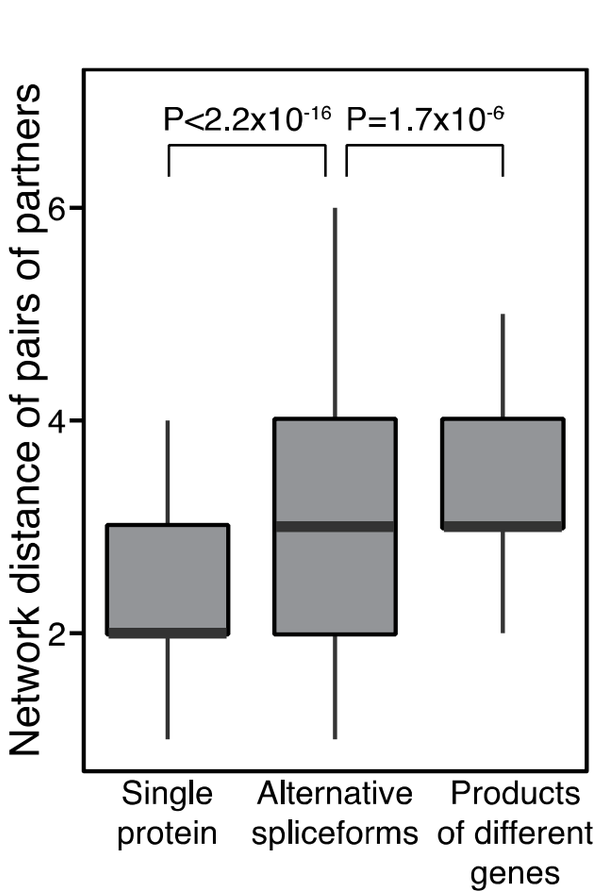
Proteins from different genes

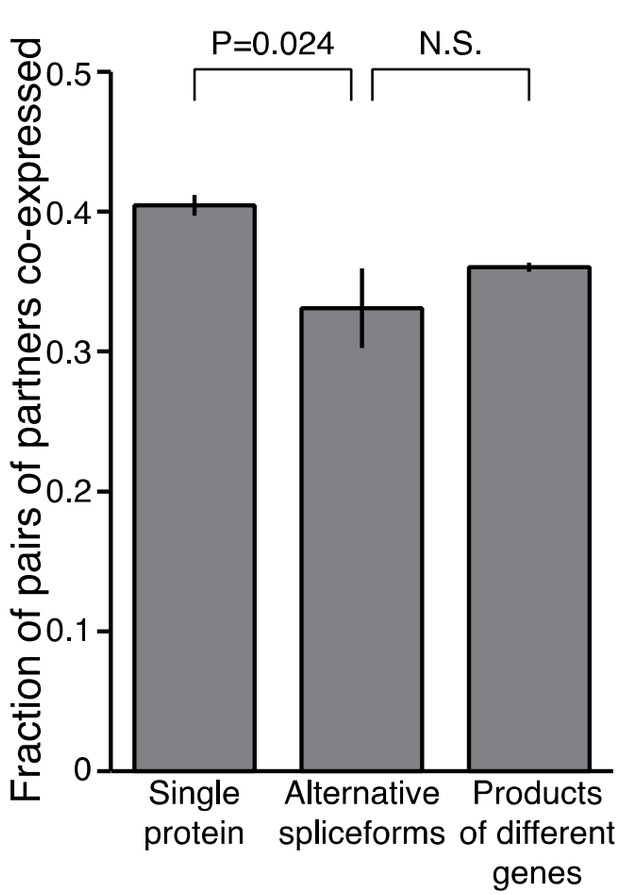


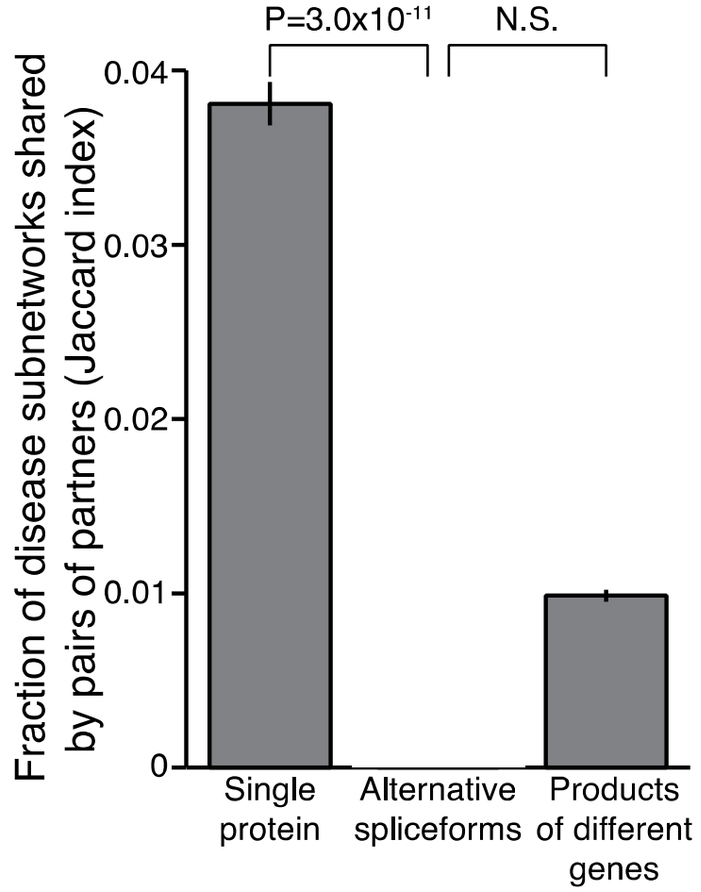
Disease subnetworks derived from GeneCards database

Safran *et al Database* 2010

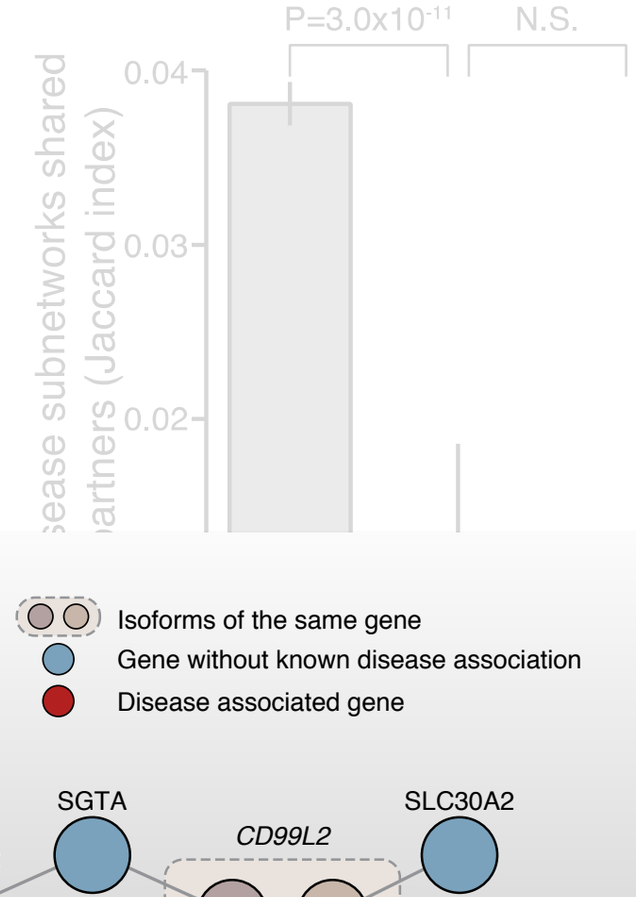
How functionally divergent are isoform-specific interactors?

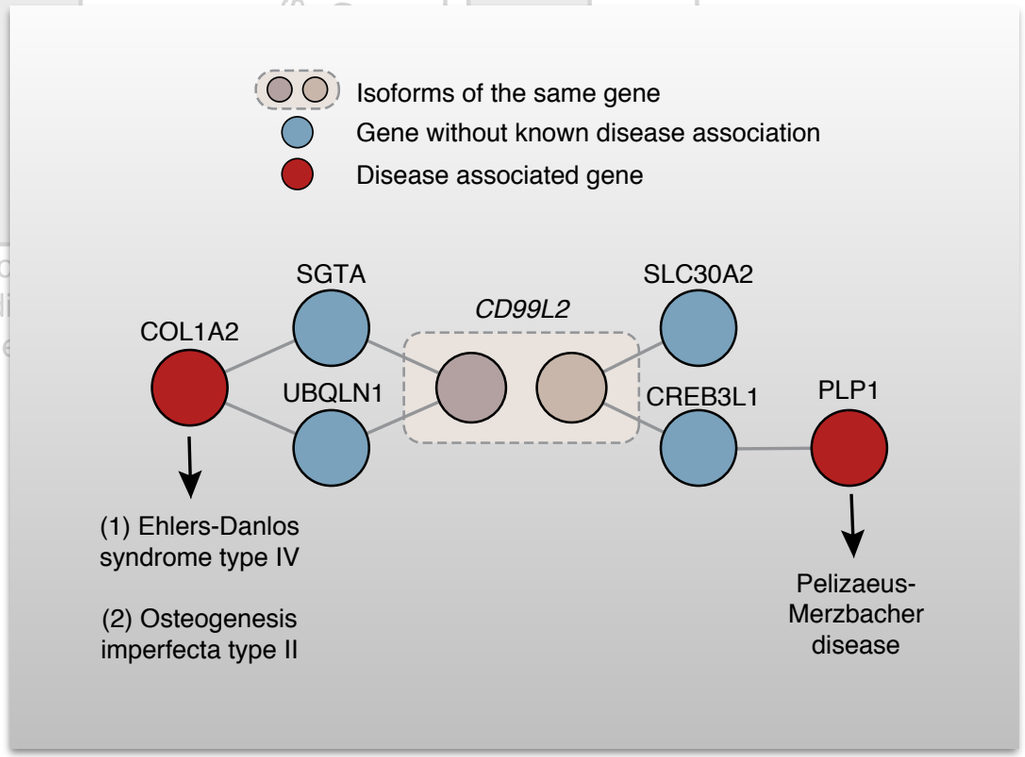






How functionally divergent are isoform-specific interactors?





How widespread is isoform functional divergence in the whole proteome?

Systematic identification
of large numbers of
isoforms pairs
for large numbers
of human genes

↓

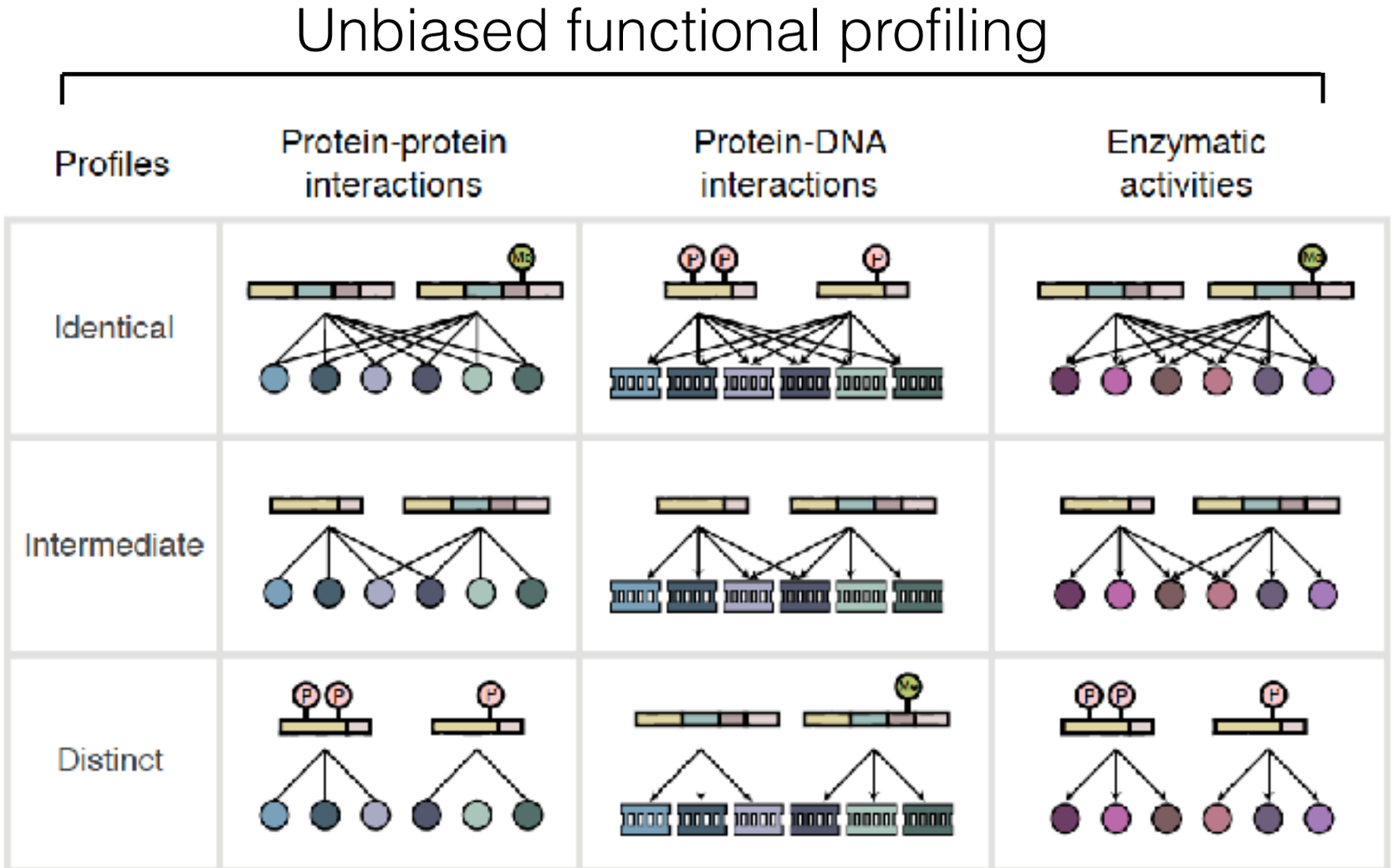
Physical interactions

Enzymatic activities

Cellular localization

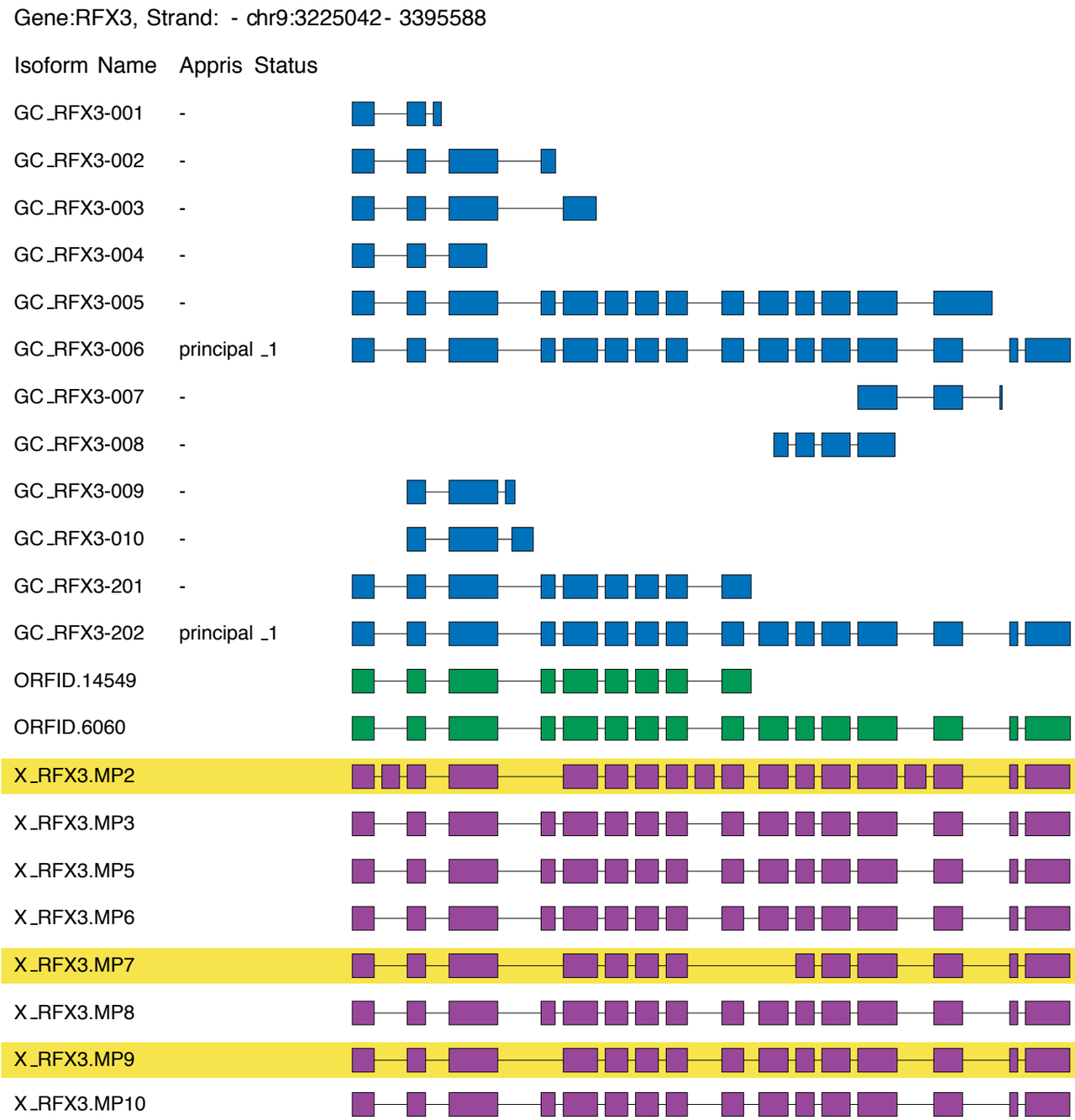
Stability

.....



Integration of ORF-Seq with third generation sequencing technologies

Pipeline in progress for 800 human TFs



Isoform clones that are not present in Gencode v25, and thus represent novel forms.

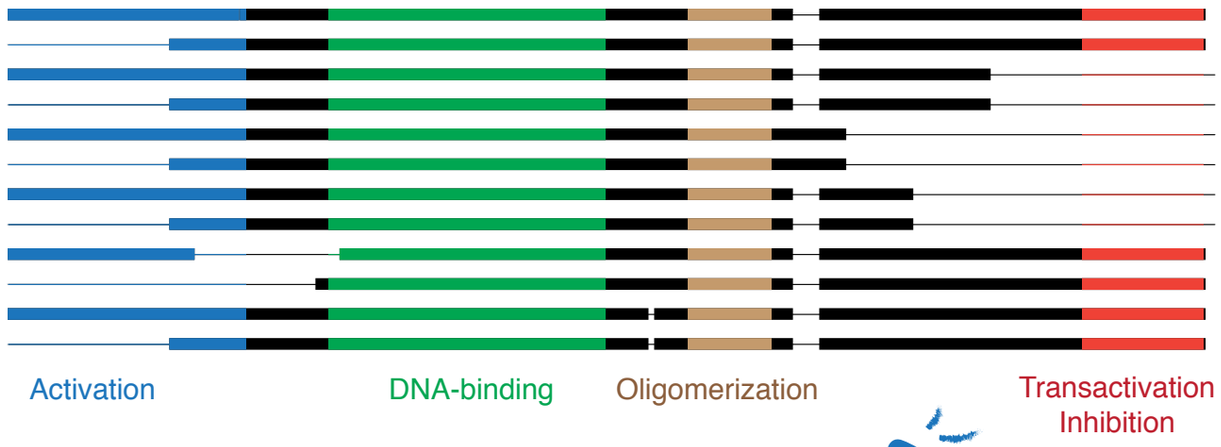
Multi-parameter comparative profiling of isoforms

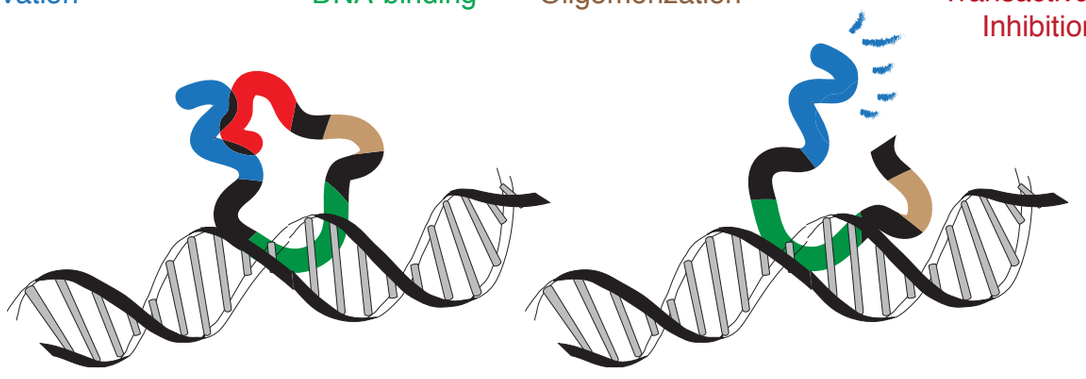
Variations in TF sequence perturb protein and/or DNA interactions

Sahni *et al Cell* 2016

Similar perturbations likely for isoforms

TP63 isoforms

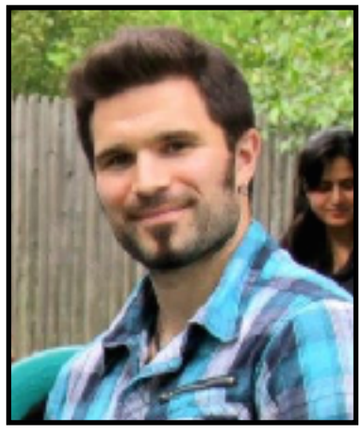




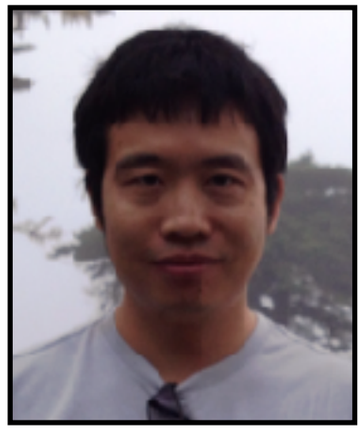
Will conduct screens for differential protein-protein and protein-DNA interactions for TF isoforms

Acknowledgements

Xinping Yang



Jasmin Coulombe-Huntington



Shuli Kang

Tong Hao

DFCI, Boston
Marc Vidal
Dawit Balcha
Wenting Bian
Tiziana Cafarelli
Michael Calderwood
Soon Gang Choi
Meaghan Daley
David DeRidder
Alice Desbuleux
Tong Hao
David Hill
Katja Luck
Dylan Markey
Julien Olivet
Carl Pollis
Aaron Richardson
Sadie Schlabach
Kerstin Spirohn
Yang Wang

Donnelly Center, Toronto
Fritz Roth

UCSD, San Diego
Lilia Iakoucheva
Shuli Kang

McGill, Montreal
Yu Xia
Jasmin Coulombe-Huntington

DANA-FARBER CANCER INSTITUTE

CCSB Center for Cancer Systems Biology

HARVARD MEDICAL SCHOOL