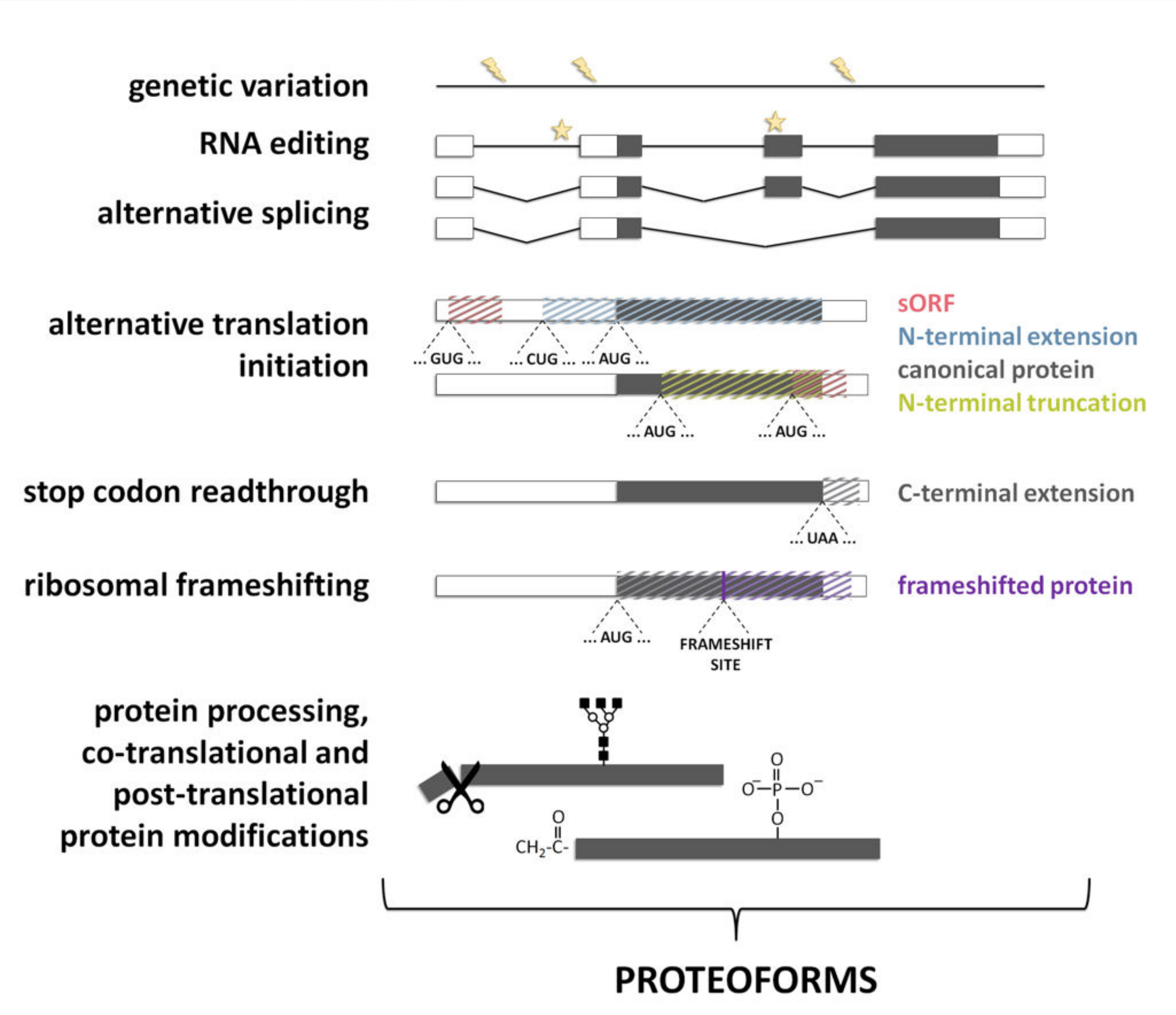
The quest for novel proteoforms: integration of proteomics and ribosome profiling based translatomics.

Gerben Menschaert

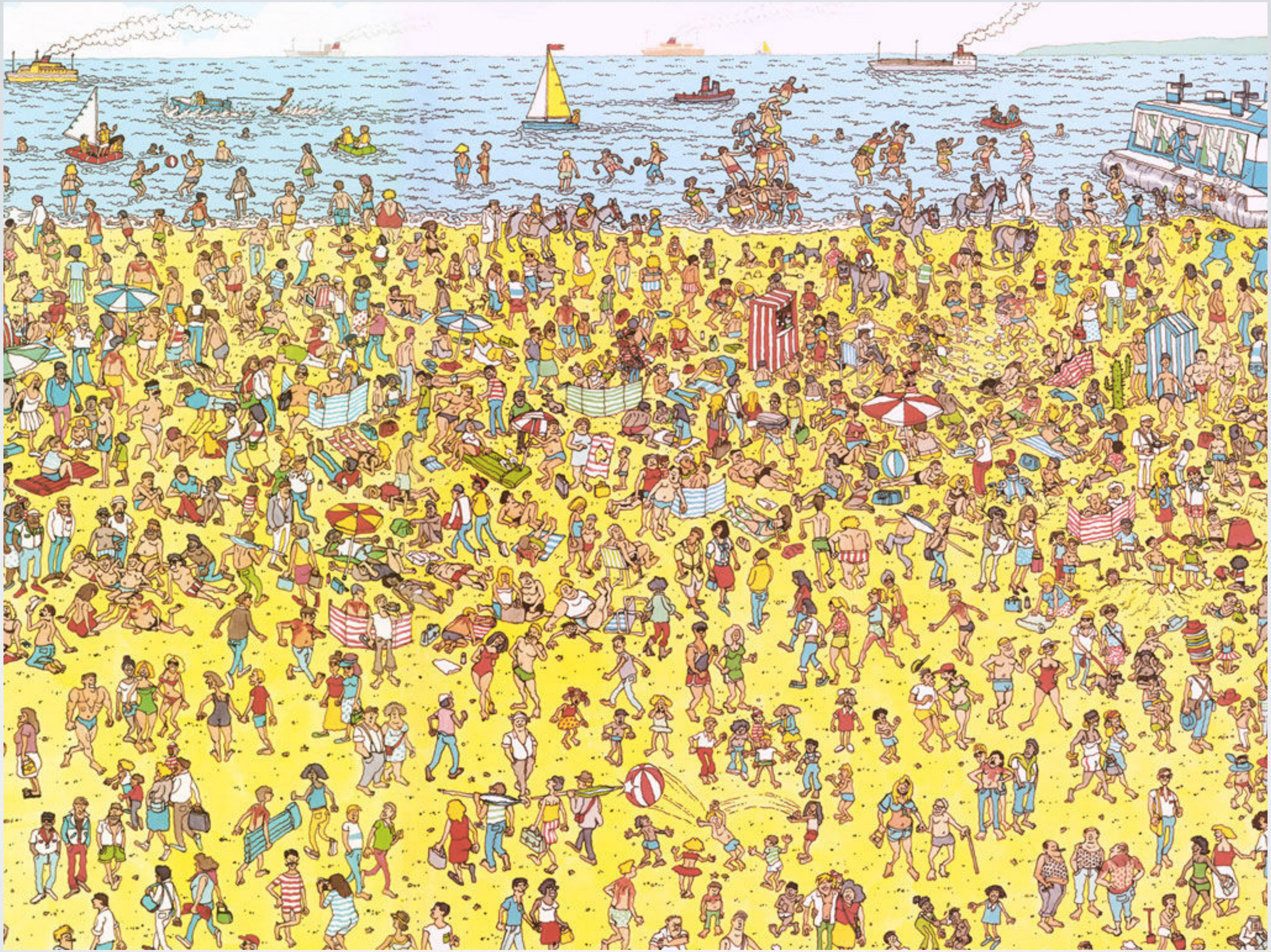
Biobix,
Lab of Bioinformatics and Computational Genomics,
Faculty of Bioscience Engineering,
Ghent University

PROTEOFORMS

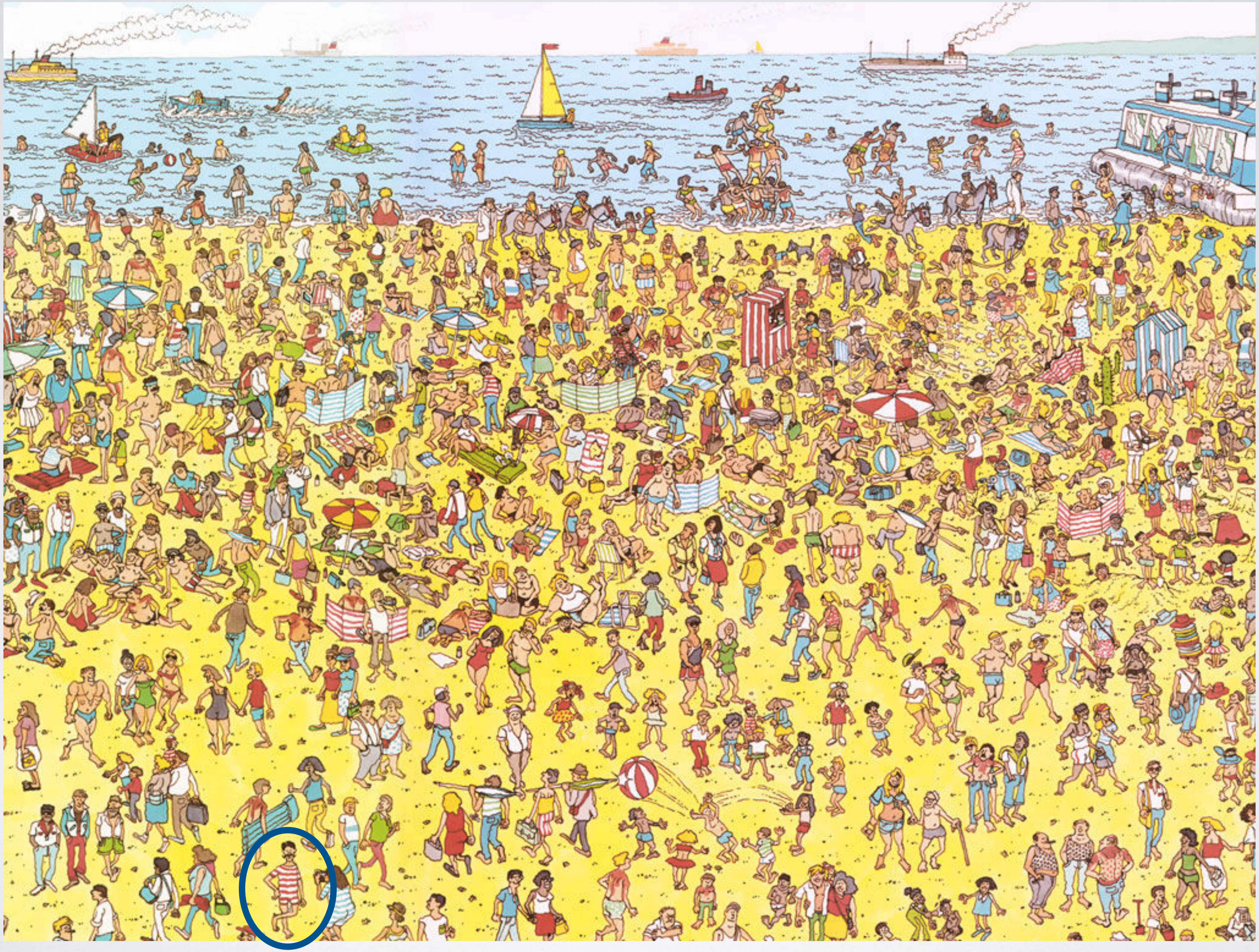


A single term – proteoforms – may be used to describe the molecular forms of proteins derived from individual genes, thus capturing the complete biological variability and all possible modifications of protein primary structure

Gawron D. et al., 2014, Proteomics; Smith L.M. et al., 2013, Nat Methods



Wally et al., Children's Literature, 1987



Wally et al., Children's Literature, 1987

Canonical protein

Wally et al., Children's Literature, 1987

Canonical protein

Other proteoform:
- extension

Wally et al., Children's Literature, 1987

Canonical protein

Other proteoform:

- extension
- truncation

Wally et al., Children's Literature, 1987

Canonical protein

Other proteoform:
- extension
- truncation
- SAV

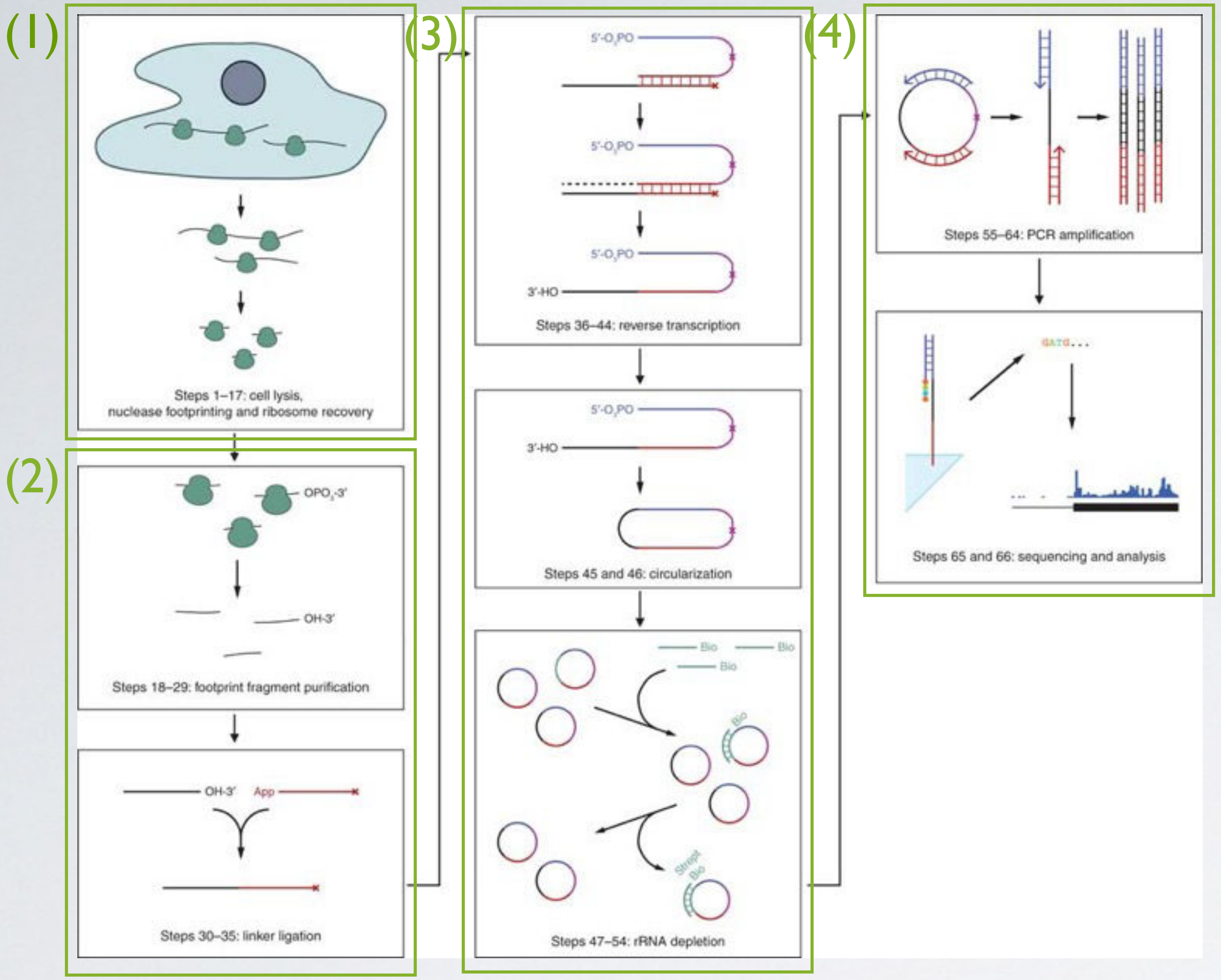
Wally et al., Children's Literature, 1987

Canonical protein

Other proteoform:
- extension
- truncation
- SAV
- sORF

Wally et al., Children's Literature, 1987

RIBOSOME PROFILING



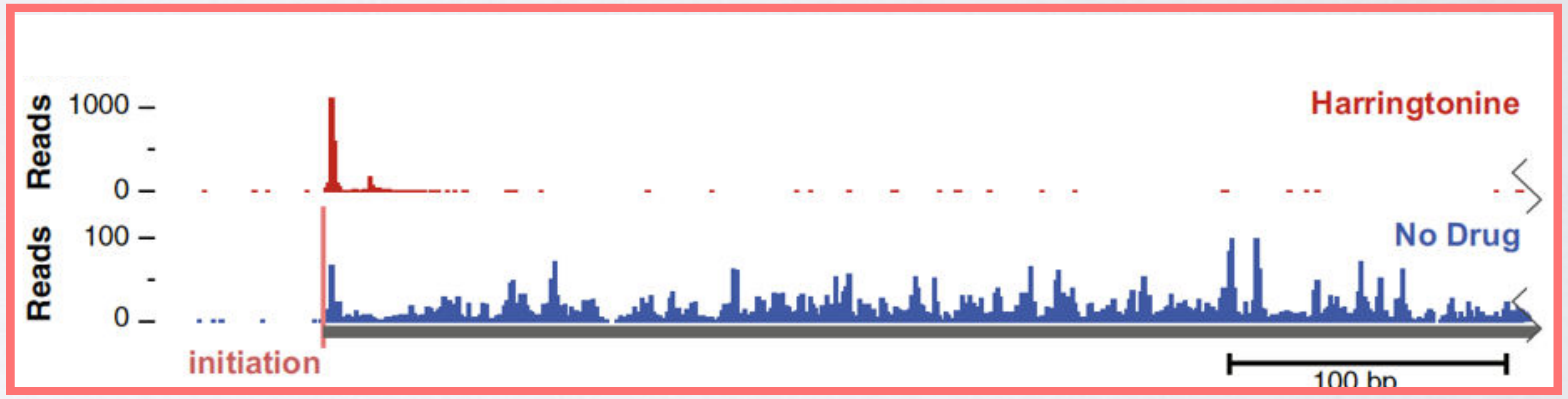
Ingolia N. et al., Nature Protocols, 2012

(1) Generation of cell extracts in which ribosomes have been faithfully halted along the mRNA they are translating in vivo

(2) Nuclease digestion of RNAs that are not protected by the ribosome followed by recovery of the ribosome-protected mRNA fragments

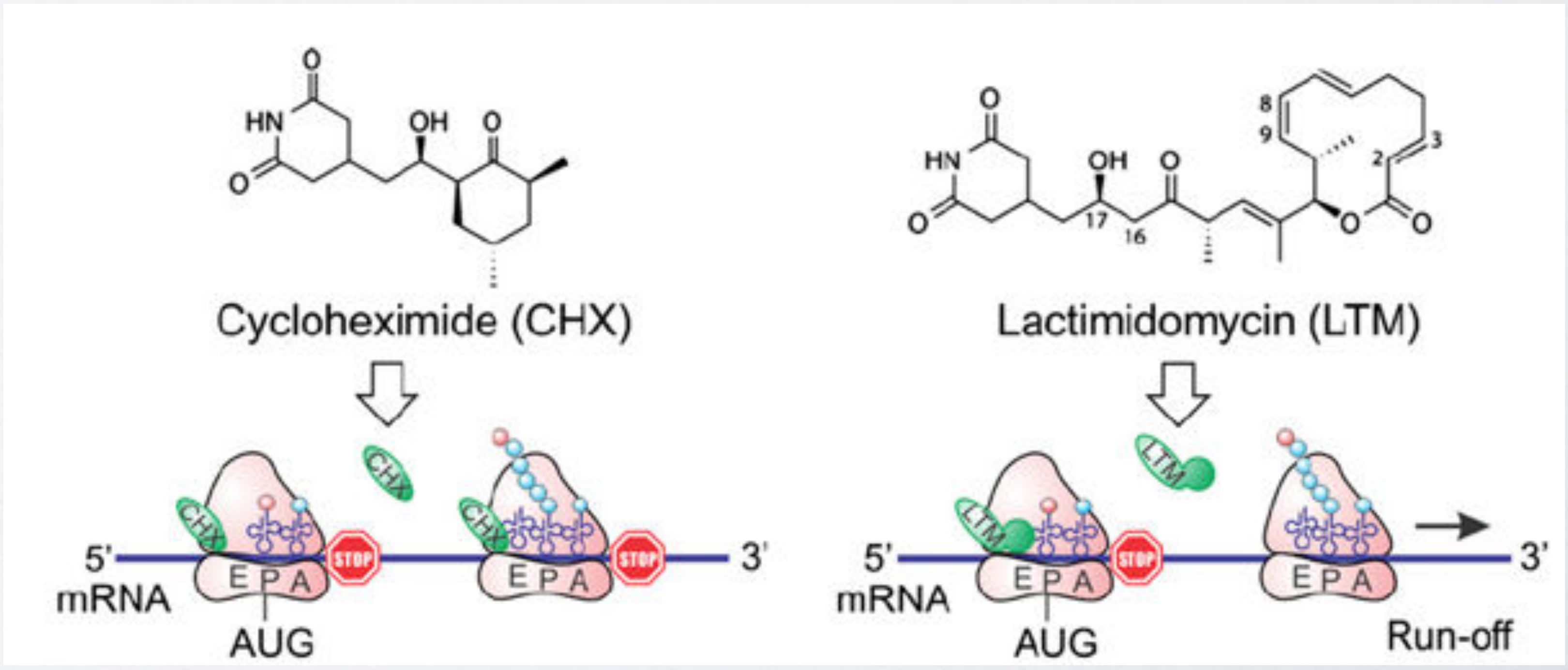
(3) Quantitative conversion of the protected RNA fragments into a DNA library

(4) That can be analyzed by deep sequencing

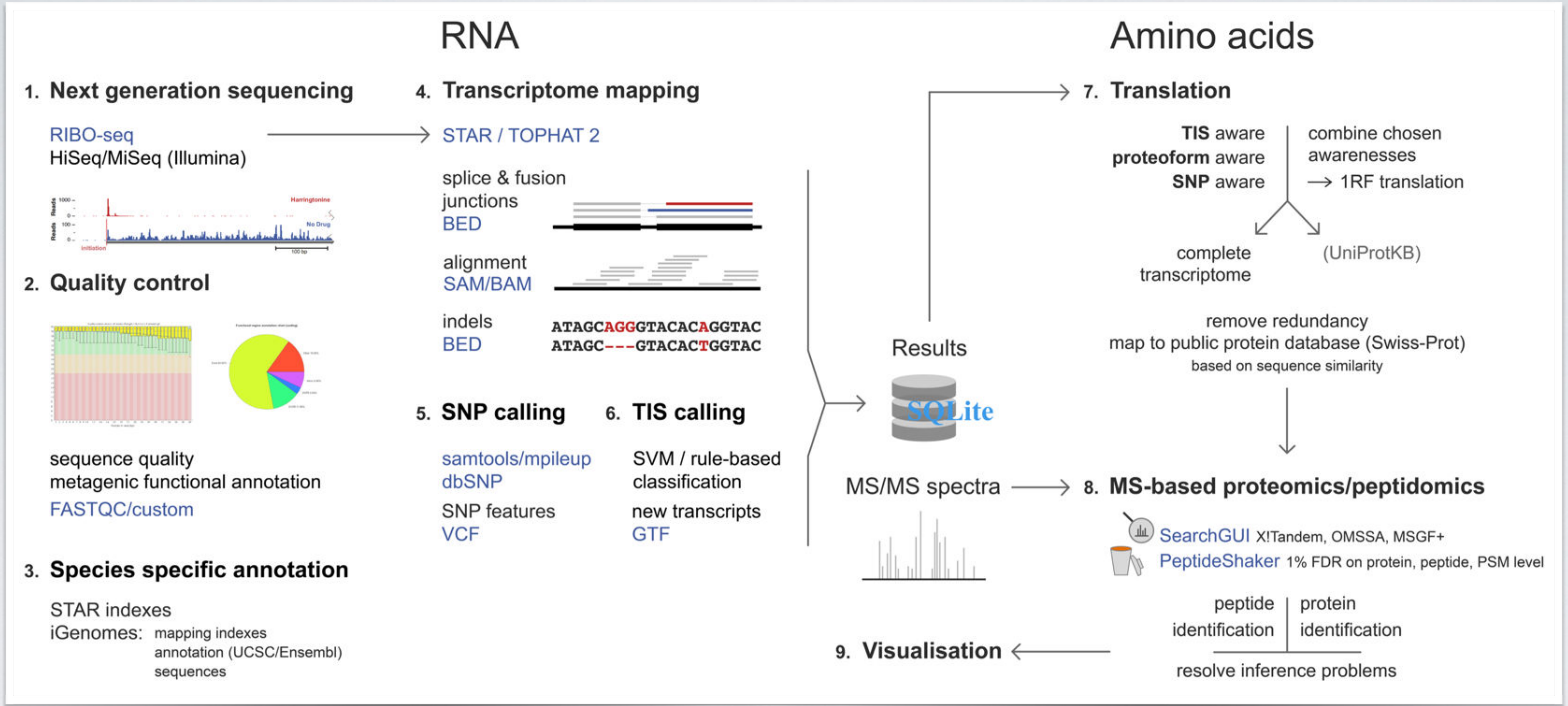


- Harringtonine	Ingolia N. et al., Cell, 2011
- Lactimidomycin (LTM)	Lee S. et al., PNAS, 2012
- Puromycin	Fritch C. et al., Gen. Research, 2012

causes ribosome accumulation at translation initiation site (TIS)



PROTEOFORMER PIPELINE



RNA

1. Next generation sequencing

RIBO-seq
HiSeq/MiSeq (Illumina)

2. Quality control

sequence quality
metagenic functional annotation
FASTQC/custom

3. Species specific annotation

STAR indexes
iGenomes: mapping indexes
annotation (UCSC/Ensembl)
sequences

4. Transcriptome mapping

STAR / TOPHAT 2

splice & fusion junctions
BED

alignment
SAM/BAM

indels
BED

5. SNP calling

samtools/mpileup
dbSNP

SNP features
VCF

6. TIS calling

SVM / rule-based classification

new transcripts
GTF

Results
SQLite

MS/MS spectra

Amino acids

7. Translation

TIS aware
proteoform aware
SNP aware

combine chosen awarenesses
→ 1RF translation

complete transcriptome
(UniProtKB)

remove redundancy
map to public protein database (Swiss-Prot)
based on sequence similarity

8. MS-based proteomics/peptidomics

SearchGUI X!Tandem, OMSSA, MSGF+
PeptideShaker 1% FDR on protein, peptide, PSM level

peptide identification | protein identification

resolve inference problems

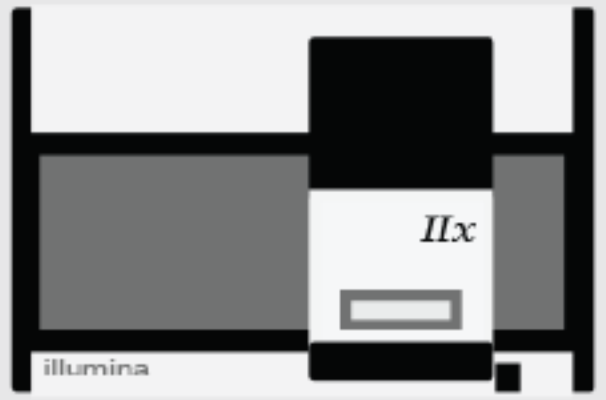
9. Visualisation

http://www.biobix.be/PROTEOFORMER

Crappé J. et al., NAR, 2015; Koch A. et al., Proteomics, 2014; Menschaert G. et al., MCP, 2013

RiboSorfs Pipeline

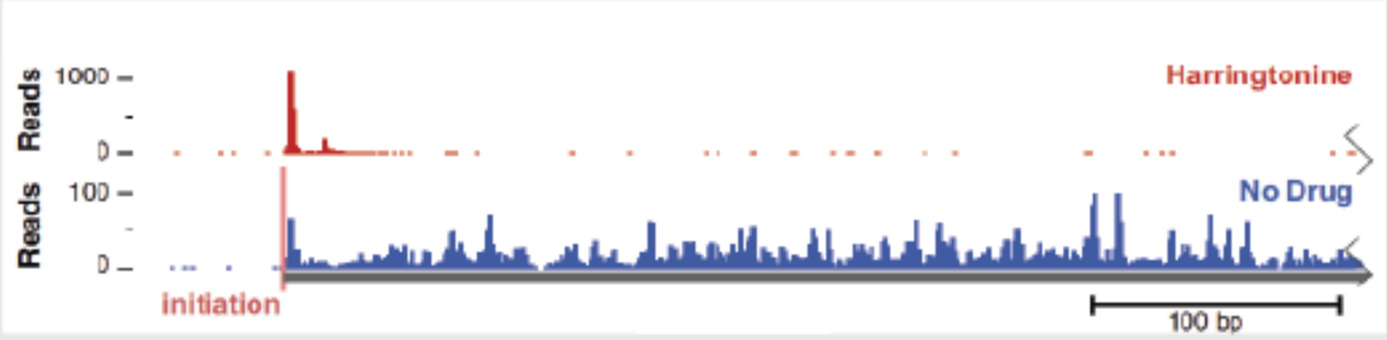
Next Generation Sequencing



RIBO-seq
Illumina HiSeq

1 Sequence processing

STAR / TopHat



RESULTS

SQLite

2 TIS Calling

Rule based classification
- min count
- Rltm - Rchx

Custom Perl

3 sORF Assembly

TIS aware

Genic & intergenic regions

Custom Perl

Annotation

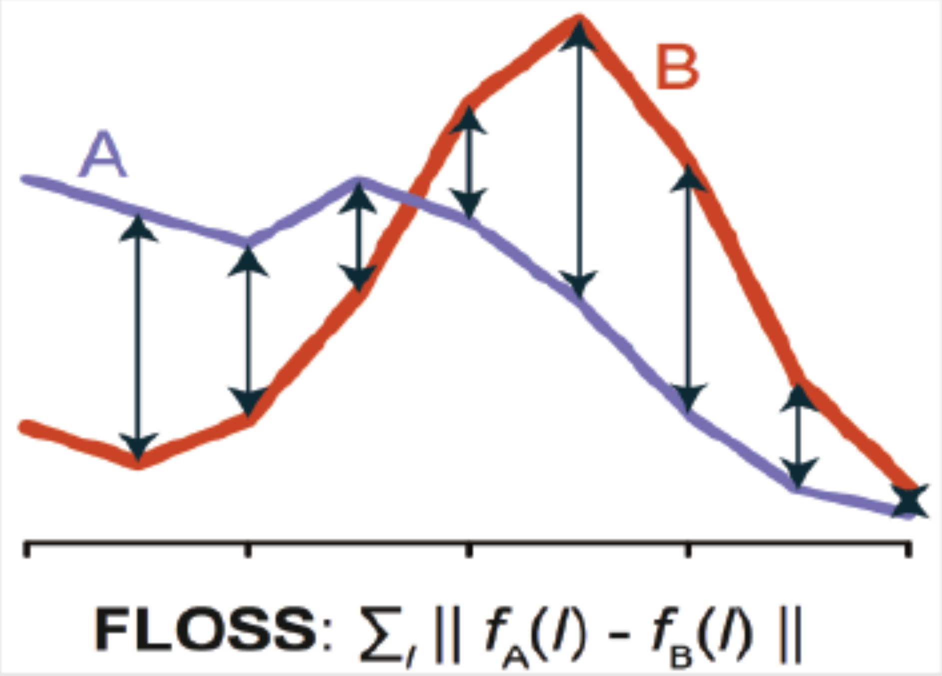
Igenomes / Ensembl

4 Conservation

```
mm9        ATGGCGGCTCCCAGGGCGGCCGCCACAGCCTGA
rn4        ATGGCGGCTCCCAGGGCGGCCGGCGCAGCCTGA
oryCun1    ATGGCGGCTCCCAGGGCGGCAGGCGCAGCCTGA
hg18       ATGGCGGCTCCCACGGCGGCAGGCGCAGCCTGA
sorAra1    ATGGCGGCTCCC------------------TGA
bosTau3    ATGGCGGCTCCCAGGGCGGCAGGCGCAGCCTGA
echTel1    ATGGCGGCTCCCACAGCGGCAGGCGCAGCTTGA
monDom4    ATGGCGGCTCCCAGCGCTGTGGCAGCGGCCTGA
```

PhyloCSF

5 Floss Score

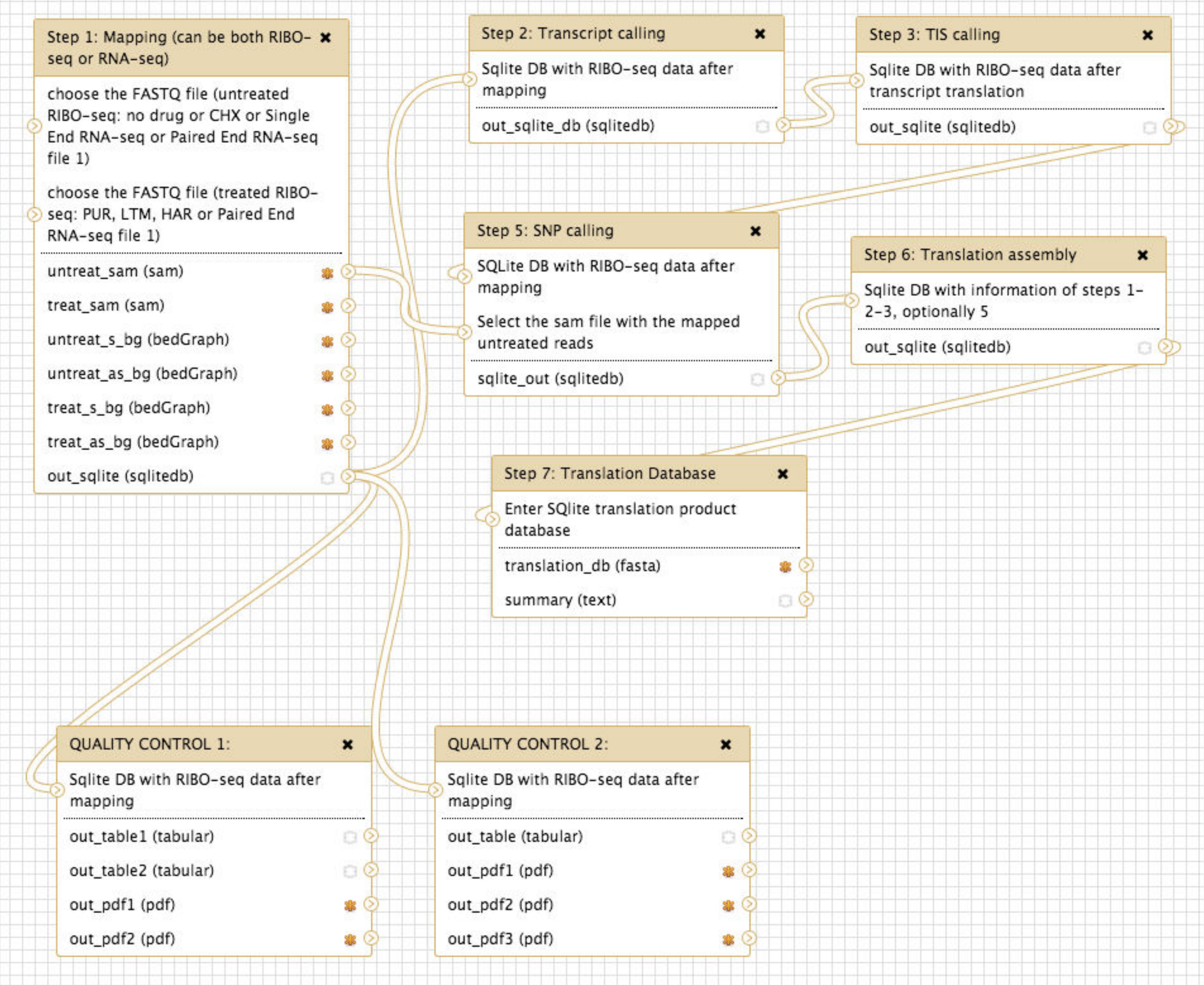


FLOSS: $\sum_l \| f_A(l) - f_B(l) \|$

Custom Perl/R

PIPELINES IN GALAXY

PROTEOFORMER



Step 1: Mapping (can be both RIBO-seq or RNA-seq)
- choose the FASTQ file (untreated RIBO-seq: no drug or CHX or Single End RNA-seq or Paired End RNA-seq file 1)
- choose the FASTQ file (treated RIBO-seq: PUR, LTM, HAR or Paired End RNA-seq file 1)
- untreat_sam (sam)
- treat_sam (sam)
- untreat_s_bg (bedGraph)
- untreat_as_bg (bedGraph)
- treat_s_bg (bedGraph)
- treat_as_bg (bedGraph)
- out_sqlite (sqlitedb)

Step 2: Transcript calling
- Sqlite DB with RIBO-seq data after mapping
- out_sqlite_db (sqlitedb)

Step 3: TIS calling
- Sqlite DB with RIBO-seq data after transcript translation
- out_sqlite (sqlitedb)

Step 5: SNP calling
- SQLite DB with RIBO-seq data after mapping
- Select the sam file with the mapped untreated reads
- sqlite_out (sqlitedb)

Step 6: Translation assembly
- Sqlite DB with information of steps 1-2-3, optionally 5
- out_sqlite (sqlitedb)

Step 7: Translation Database
- Enter SQlite translation product database
- translation_db (fasta)
- summary (text)

QUALITY CONTROL 1:
- Sqlite DB with RIBO-seq data after mapping
- out_table1 (tabular)
- out_table2 (tabular)
- out_pdf1 (pdf)
- out_pdf2 (pdf)

QUALITY CONTROL 2:
- Sqlite DB with RIBO-seq data after mapping
- out_table (tabular)
- out_pdf1 (pdf)
- out_pdf2 (pdf)
- out_pdf3 (pdf)

RIBOSORFS

Step 1: Mapping (RIBO-seq)
- choose the FASTQ file (untreated RIBO-seq: no drug or CHX)
- choose the FASTQ file (treated RIBO-seq: PUR, LTM, HAR)
- untreat_sam (sam)
- treat_sam (sam)
- untreat_s_bg (bedGraph)
- untreat_as_bg (bedGraph)
- treat_s_bg (bedGraph)
- treat_as_bg (bedGraph)
- out_sqlite (sqlitedb)

Step 2: TIS calling
- Sqlite DB with mapped RIBO-seq data
- out_sqlite (sqlitedb)

Step 5: sORF translation assembly
- Sqlite DB with information of steps 1-2, optionally 4
- out_sqlite (sqlitedb)

Step 6a: Create PhyloCSF input
- Sqlite DB with information of steps 1-2, optionally 4
- out_sqlite (sqlitedb)

Step 6b: sORF coding/non-coding analysis
- Sqlite DB with information of steps 1-2-5-6a, optionally 4
- out_sqlite (sqlitedb)

Step 7: FLOSS calculation
- Sqlite db with information from at least mapping, TIS calling and assembly.
- out_sqlite (sqlitedb)

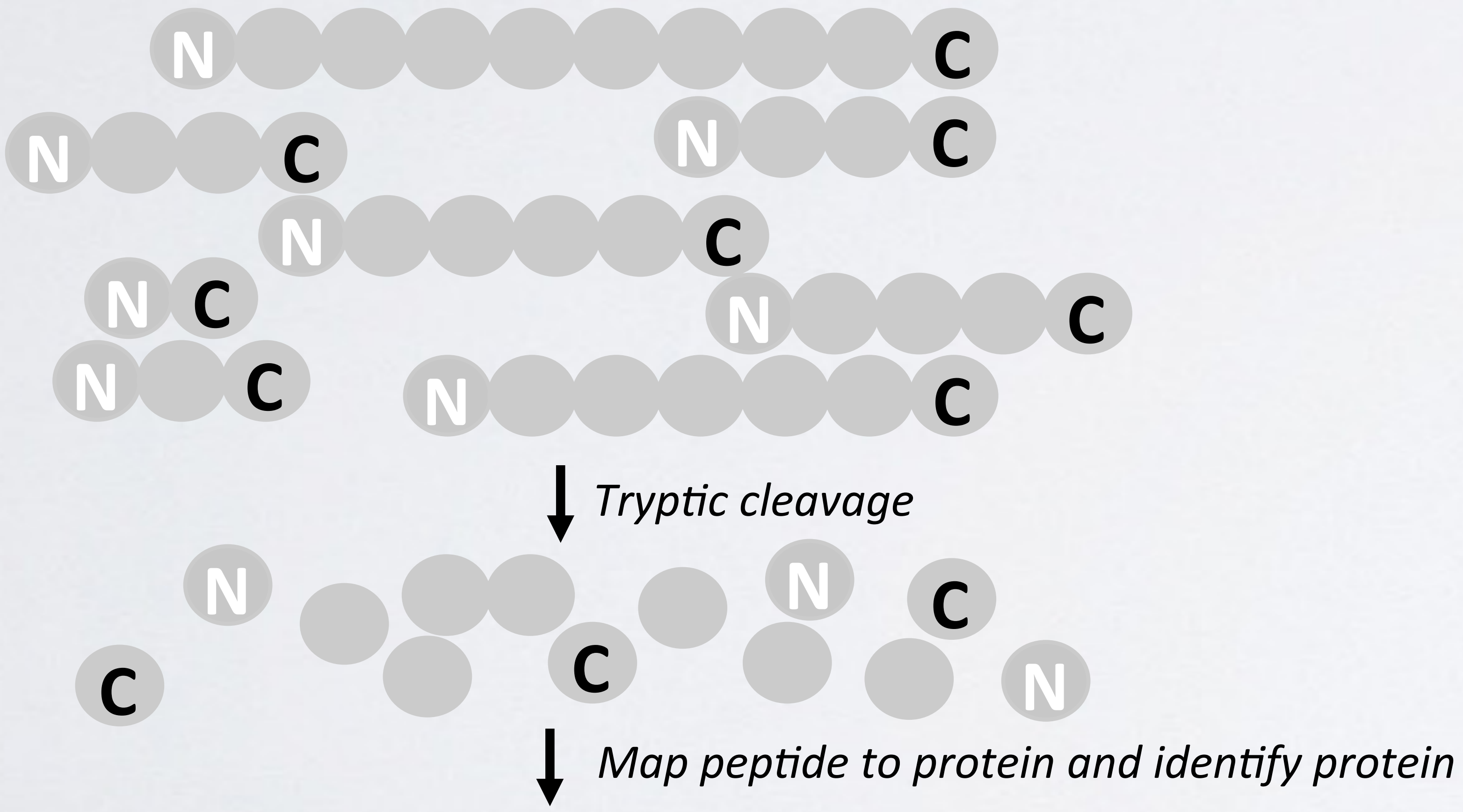
Step 8: Translation Assembly
- Sqlite db with information from mapping, TIS calling, assembly and optionally PhyloCSF and Floss
- out_sqlite (sqlitedb)
- fasta_out (fasta)

http://www.biobix.be/PROTEOFORMER

PROTEOFORMER: EXPERIMENTAL SETUP

- The PROTEOFORMER method was evaluated and optimized using available mouse embryonic stem cell (**mESC**) RIBO-seq data. *Ingolia N. et al., Cell, 2011*
- Matching shotgun and N-terminal COFRADIC proteomics experiments data were generated to validate our setup.
- RIBO-seq and matching MS data on **HCT116** cells were obtained to repeat our approach.

Shotgun proteomics



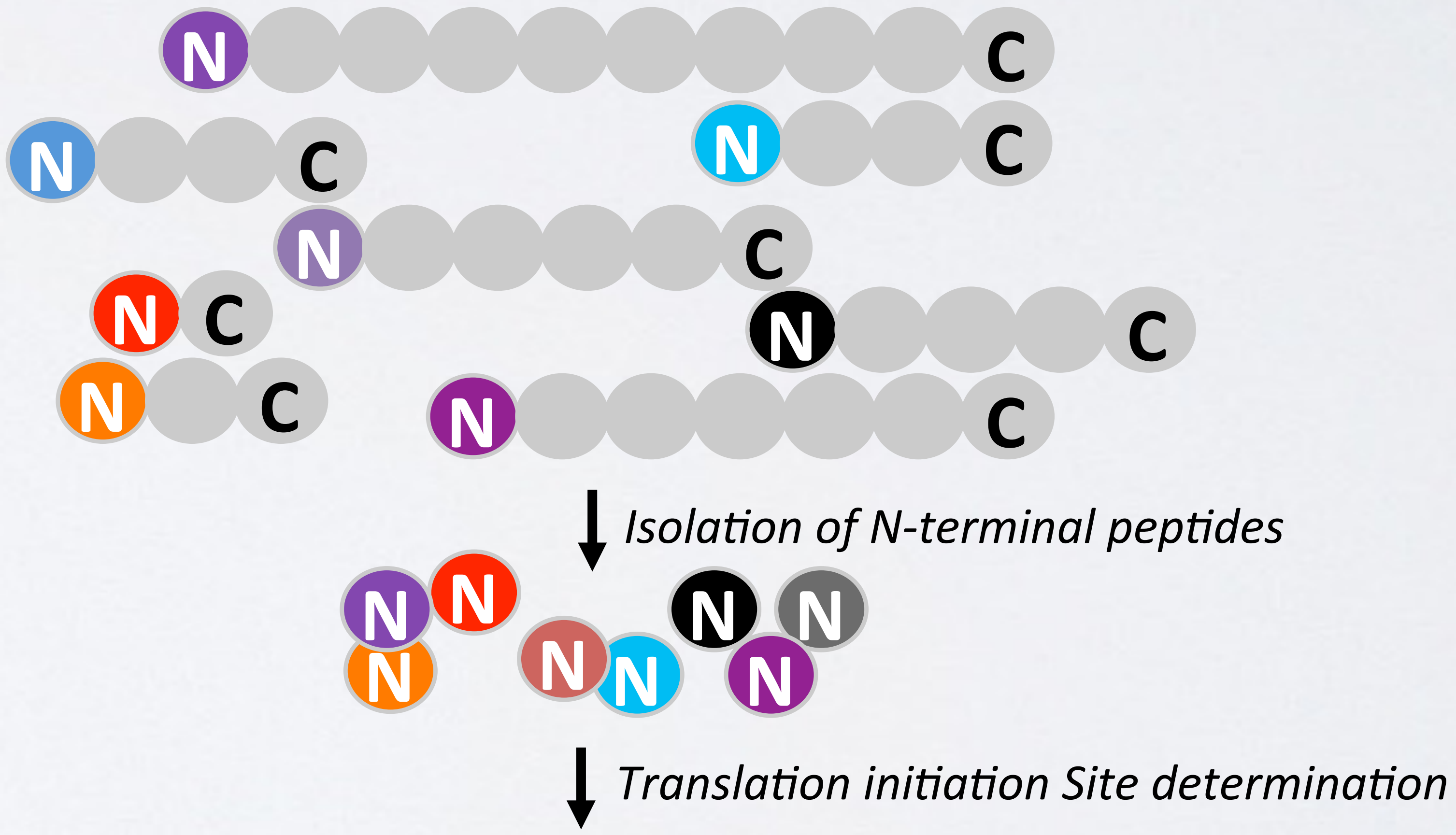
Tryptic cleavage

Map peptide to protein and identify protein

LTQ Orbitrap Velos/XL

N–terminal COFRADIC positional proteomics

Staes A. et al., Nature Protocols, 2011

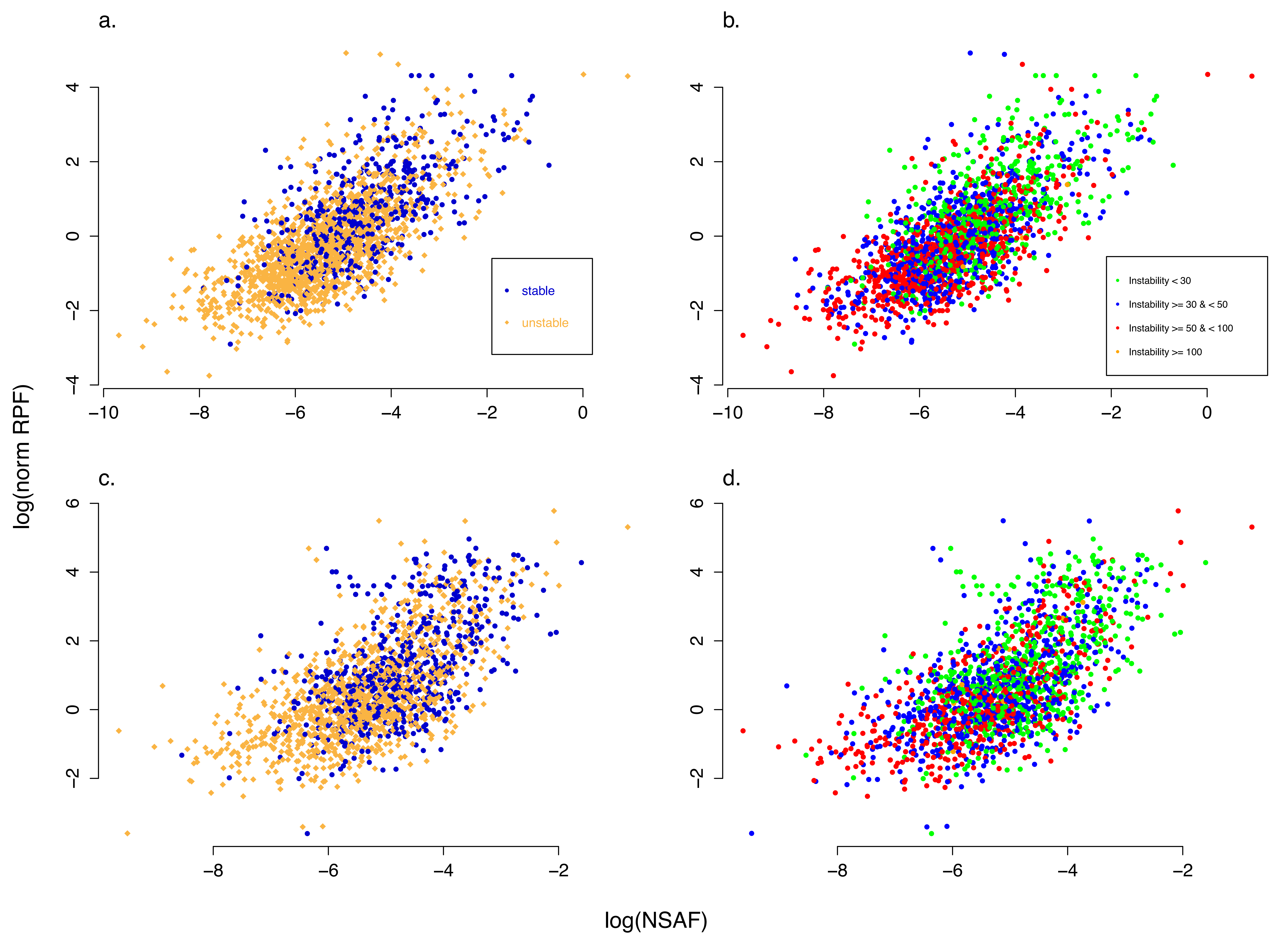


Isolation of N-terminal peptides

Translation initiation Site determination

PROTEOFORMER: QUANTITATIVE CORRELATION

Mouse



$r^2 = 0.714$

Human

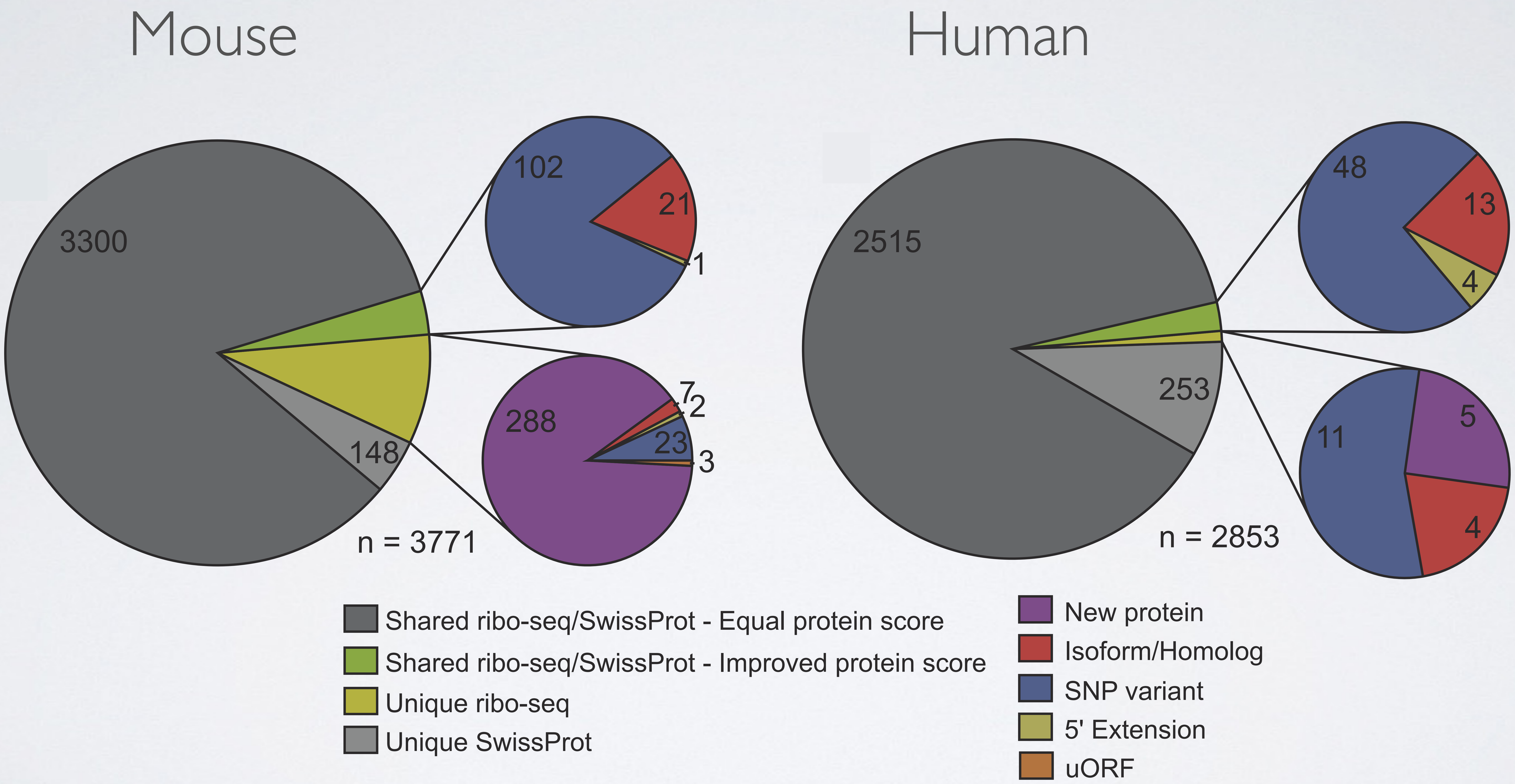
$r^2 = 0.643$

- aTIS
- RFP count >= 200
- spectral count >= 2

PROTEOFORMER: ALLOWS DEEP PROTEOME COVERAGE

1. overall improved identification rate
2. improved protein score

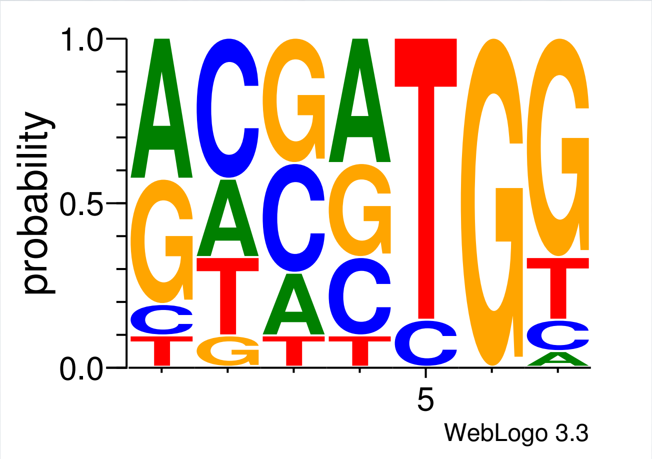
Shotgun proteomics



PROTEOFORMER: ALLOWS DEEP PROTEOME COVERAGE

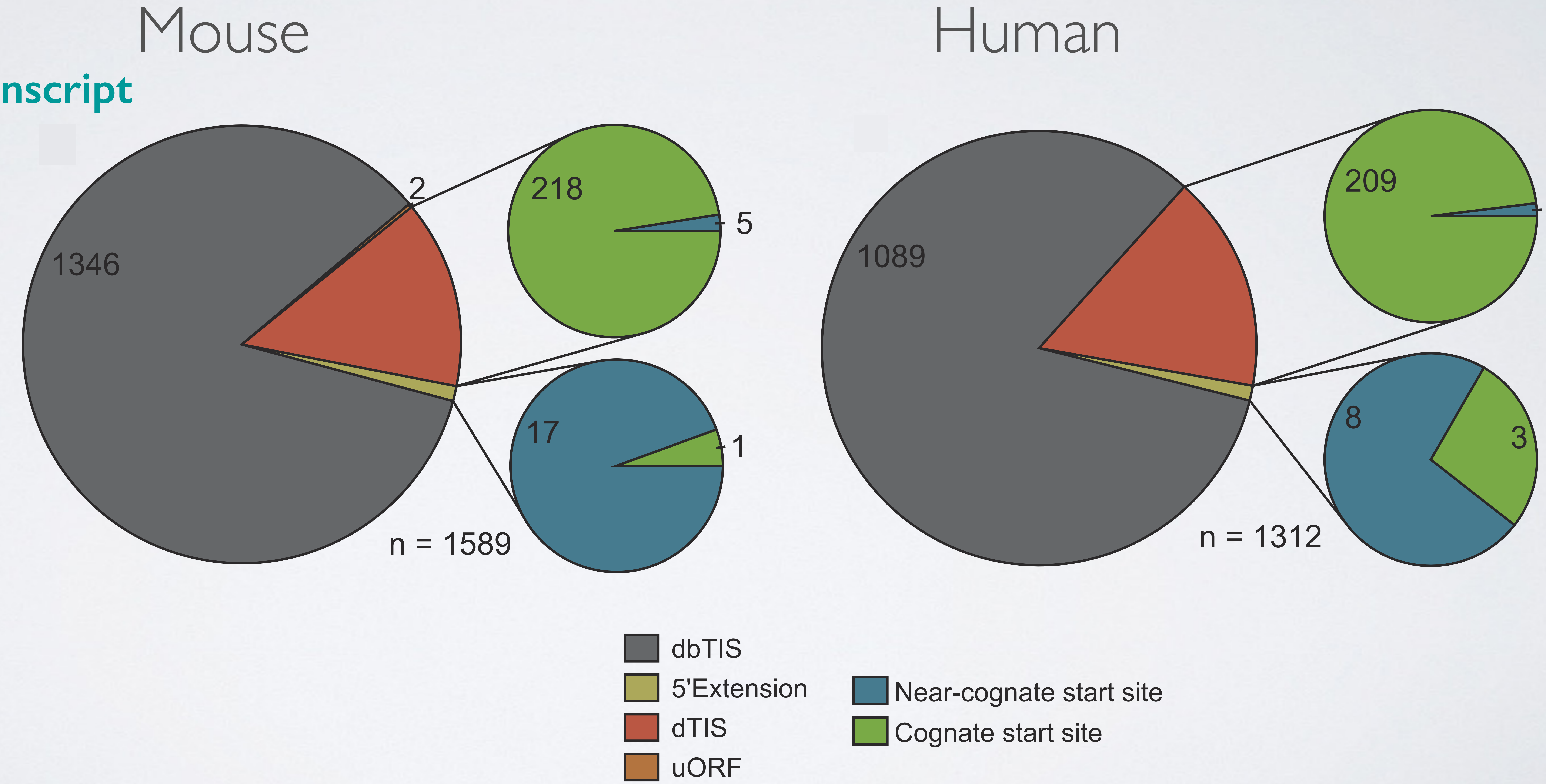
1. new translation products: 5’ext, 5’trunc, uORF

2. @ near cognate start sites

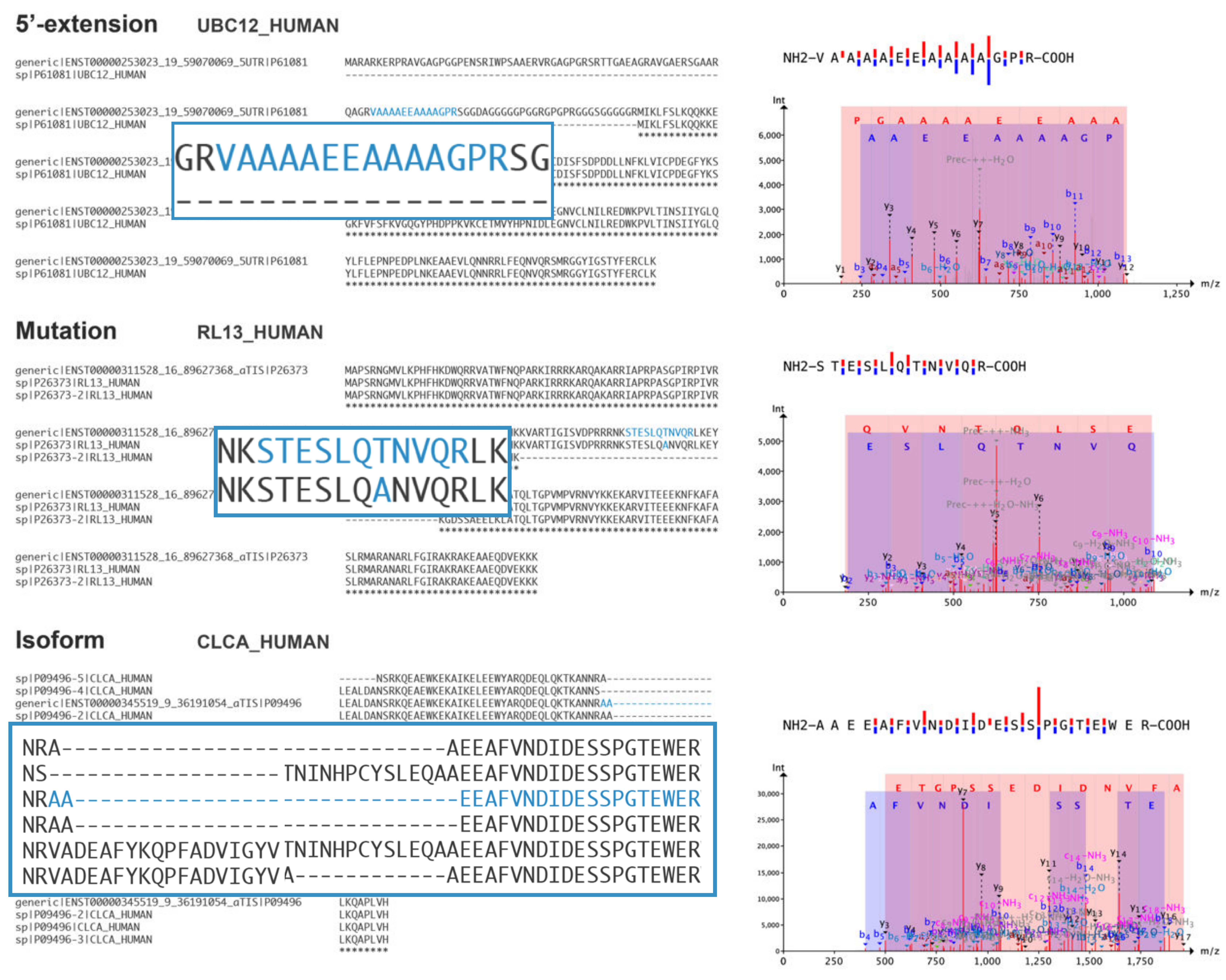


N-terminal COFRADIC positional proteomics

3. multiple TIS per transcript



PROTEOFORMER: NOVEL EVENTS



5'-extension UBC12_HUMAN

```
generic|ENST00000253023_19_59070069_5UTR|P61081    MARARKERPRAVGAGPGGPENSRIWPSAAERVRGAGPGRSRTTGAEAGRAVGAERSGAAR
sp|P61081|UBC12_HUMAN                              ------------------------------------------------------------

generic|ENST00000253023_19_59070069_5UTR|P61081    QAGRVAAAAEEAAAAGPRSGGDAGGGGGPGGRGPGPRGGGSGGGGRMIKLFSLKQQKKE
sp|P61081|UBC12_HUMAN                                                                     -------------MIKLFSLKQQKKE
                                                                                                       *************

generic|ENST00000253023_1                                     DISFSDPDDLLNFKLVICPDEGFYKS
sp|P61081|UBC12_HUMAN                                         DISFSDPDDLLNFKLVICPDEGFYKS
                                                              **************************

generic|ENST00000253023_1                                     EGNVCLNILREDWKPVLTINSIIYGLQ
sp|P61081|UBC12_HUMAN                              GKFVFSFKVGQGYPHDPPKVKCETMVYHPNIDLEGNVCLNILREDWKPVLTINSIIYGLQ
                                                   ************************************************************

generic|ENST00000253023_19_59070069_5UTR|P61081    YLFLEPNPEDPLNKEAAEVLQNNRRLFEQNVQRSMRGGYIGSTYFERCLK
sp|P61081|UBC12_HUMAN                              YLFLEPNPEDPLNKEAAEVLQNNRRLFEQNVQRSMRGGYIGSTYFERCLK
                                                   **************************************************
```

GRVAAAAEEAAAAGPRSG

Mutation RL13_HUMAN

```
generic|ENST00000311528_16_89627368_aTIS|P26373    MAPSRNGMVLKPHFHKDWQRRVATWFNQPARKIRRRKARQAKARRIAPRPASGPIRPIVR
sp|P26373|RL13_HUMAN                               MAPSRNGMVLKPHFHKDWQRRVATWFNQPARKIRRRKARQAKARRIAPRPASGPIRPIVR
sp|P26373-2|RL13_HUMAN                             MAPSRNGMVLKPHFHKDWQRRVATWFNQPARKIRRRKARQAKARRIAPRPASGPIRPIVR
                                                   ************************************************************

generic|ENST00000311528_16_89627                              KKVARTIGISVDPRRRNKSTESLQTNVQRLKEY
sp|P26373|RL13_HUMAN                                          KKVARTIGISVDPRRRNKSTESLQANVQRLKEY
sp|P26373-2|RL13_HUMAN                                        K--------------------------------
                                                              *

generic|ENST00000311528_16_89627                              TQLTGPVMPVRNVYKKEKARVITEEEKNFKAFA
sp|P26373|RL13_HUMAN                                          TQLTGPVMPVRNVYKKEKARVITEEEKNFKAFA
sp|P26373-2|RL13_HUMAN                             --------------KGDSSAEELKLATQLTGPVMPVRNVYKKEKARVITEEEKNFKAFA
                                                                 *********************************************

generic|ENST00000311528_16_89627368_aTIS|P26373    SLRMARANARLFGIRAKRAKEAAEQDVEKKK
sp|P26373|RL13_HUMAN                               SLRMARANARLFGIRAKRAKEAAEQDVEKKK
sp|P26373-2|RL13_HUMAN                             SLRMARANARLFGIRAKRAKEAAEQDVEKKK
                                                   *******************************
```

NKSTESLQTNVQRLK
NKSTESLQANVQRLK

Isoform CLCA_HUMAN

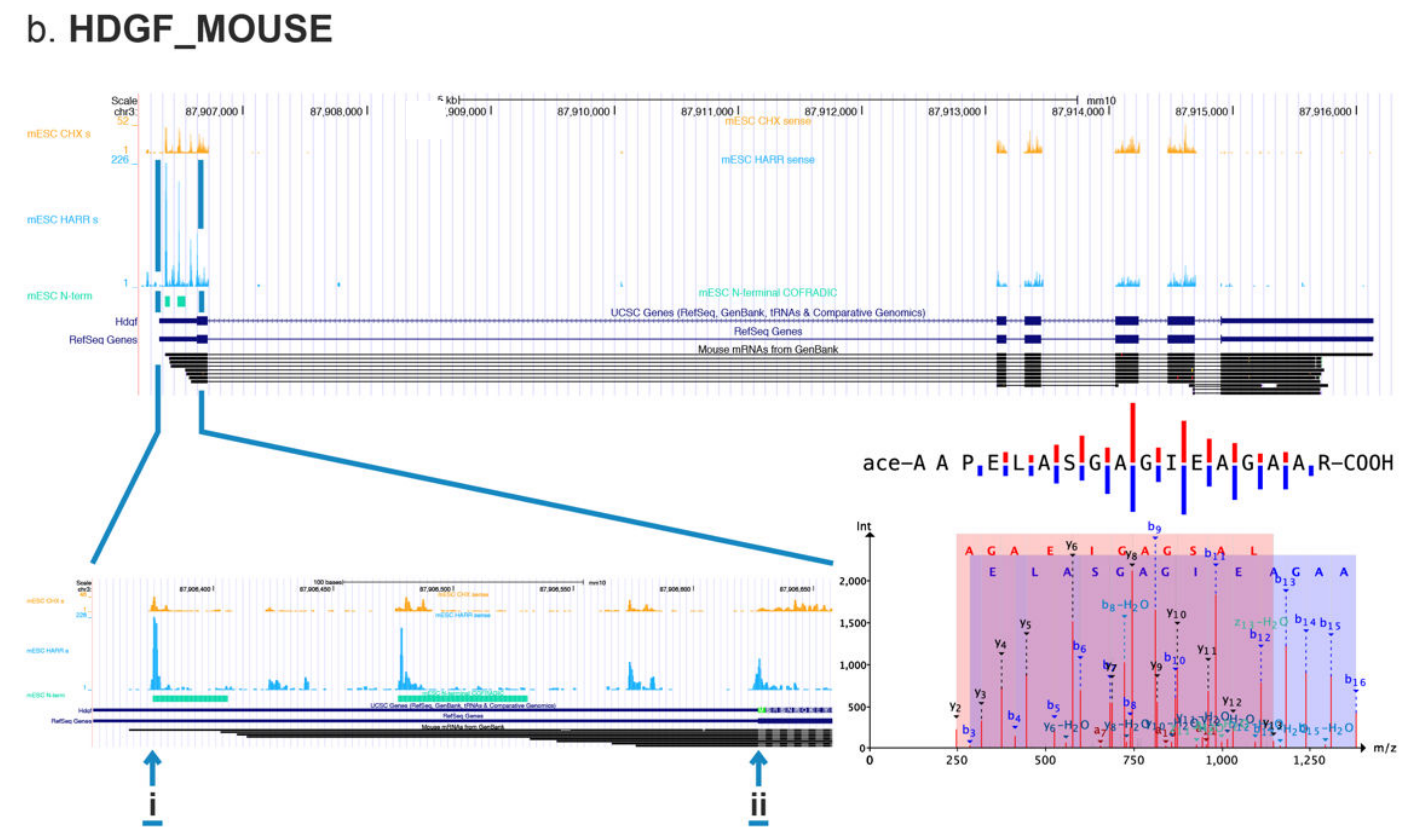
```
sp|P09496-5|CLCA_HUMAN                            ------NSRKQEAEWKEKAIKELEEWYARQDEQLQKTKANNRA----------------
sp|P09496-4|CLCA_HUMAN                            LEALDANSRKQEAEWKEKAIKELEEWYARQDEQLQKTKANNS-----------------
generic|ENST00000345519_9_36191054_aTIS|P09496    LEALDANSRKQEAEWKEKAIKELEEWYARQDEQLQKTKANNRAA---------------
sp|P09496-2|CLCA_HUMAN                            LEALDANSRKQEAEWKEKAIKELEEWYARQDEQLQKTKANNRAA---------------
```

```
NRA---------------------------------AEEAFVNDIDESSPGTEWER
NS-------------------TNINHPCYSLEQAAEEAFVNDIDESSPGTEWER
NRAA--------------------------------EEAFVNDIDESSPGTEWER
NRAA--------------------------------EEAFVNDIDESSPGTEWER
NRVADEAFYKQPFADVIGYVTNINHPCYSLEQAAEEAFVNDIDESSPGTEWER
NRVADEAFYKQPFADVIGYVA-------------AEEAFVNDIDESSPGTEWER
```

```
generic|ENST00000345519_9_36191054_aTIS|P09496    LKQAPLVH
sp|P09496-2|CLCA_HUMAN                            LKQAPLVH
sp|P09496|CLCA_HUMAN                              LKQAPLVH
sp|P09496-3|CLCA_HUMAN                            LKQAPLVH
                                                  ********
```

PROTEOFORMER: NOVEL EVENTS

TIS @ near cognate start sites



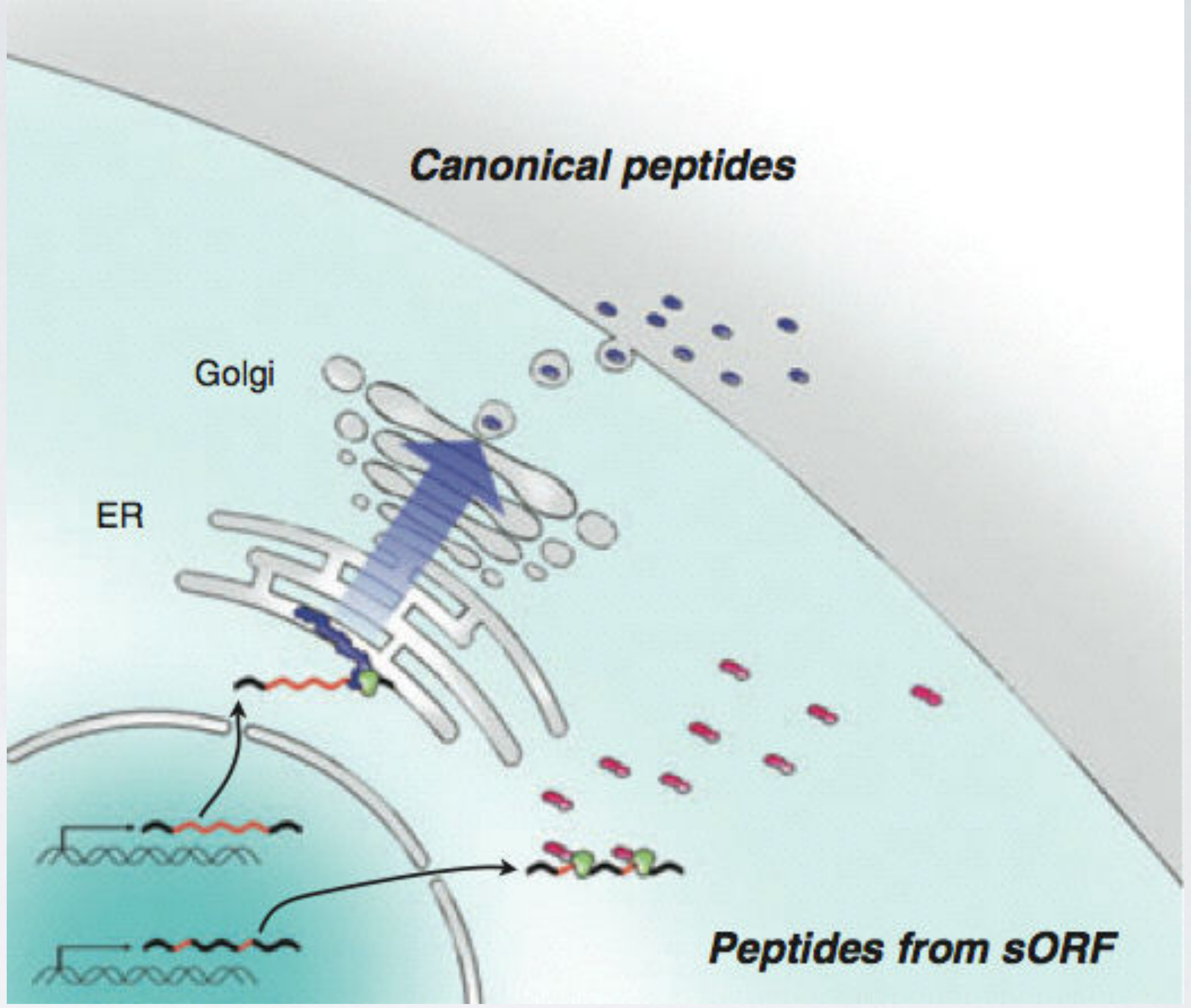
b. **HDGF_MOUSE**

- near-cognate initiation site (GTG)
- N-terminus was acetylated (Ace-)
- initiator methionine removed (MAP)

CODING sORFs

Canonical bio-active peptides:
- cleaved from precursor
- signal peptide @ N-terminus
- secretory pathway

Micropeptides:
- translated directly from sORF
- lacking signal sequence
- released in cytoplasm

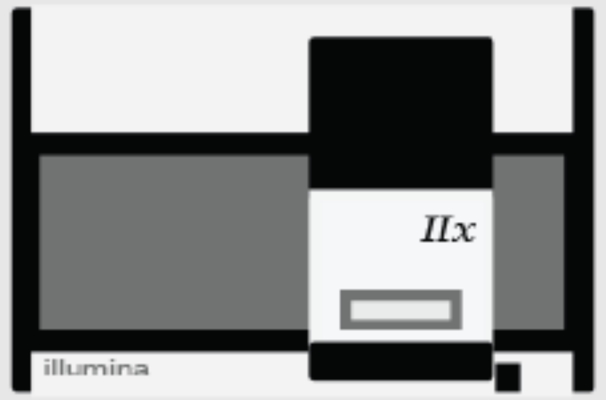


Polaris: 3 peptides: 8, 9, 36 AA
Rotundifolia4: 1 peptide: 53 AA
Enod40: 2 peptides: 12 and 24 AA
Tarsal-less/pri: 4 peptides: 11 and 32 AA
Sarcolipin/Sarcolamban/Myoregulin: 29/29/46 AA

Andrews S.J., Rothnagel J.A., 2014, Nat Gen. Rev; Crappé J. et. al., 2014, Eupa Open Proteomics

RIBOSORFS PIPELINE

Next Generation Sequencing

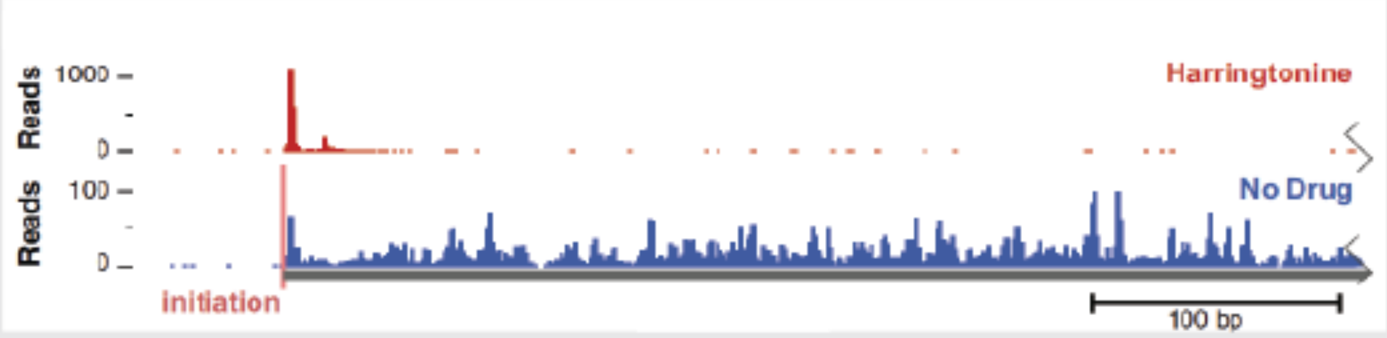


RIBO-seq

Illumina HiSeq

1 Sequence processing

STAR / TopHat



RESULTS

SQLite

2 TIS Calling

Rule based classification
- min count
- Rltm - Rchx

Custom Perl

3 sORF Assembly

TIS aware

Genic & intergenic regions

Custom Perl

Annotation

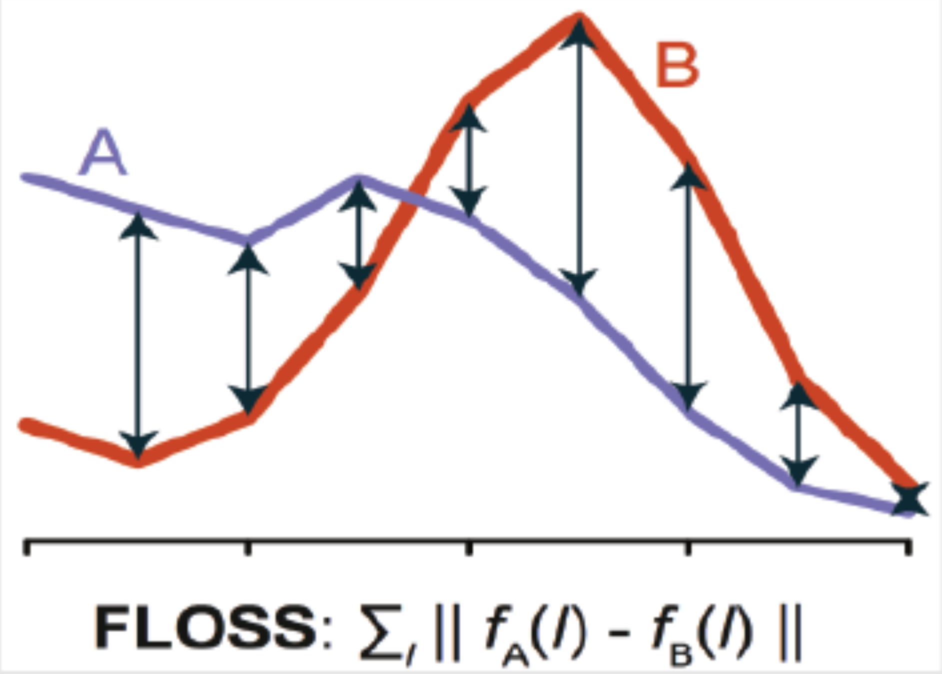
Igenomes / Ensembl

4 Conservation

```
mm9        ATGGCGGCTCCCAGGGCGGCCGCCACAGCCTGA
rn4        ATGGCGGCTCCCAGGGCGGCCGGCGCAGCCTGA
oryCun1    ATGGCGGCTCCCAGGGCGGCAGGCGCAGCCTGA
hg18       ATGGCGGCTCCCACGGCGGCAGGCGCAGCCTGA
sorAra1    ATGGCGGCTCCC------------------TGA
bosTau3    ATGGCGGCTCCCAGGGCGGCAGGCGCAGCCTGA
echTel1    ATGGCGGCTCCCACAGCGGCAGGCGCAGCTTGA
monDom4    ATGGCGGCTCCCAGCGCTGTGGCAGCGGCCTGA
```

PhyloCSF

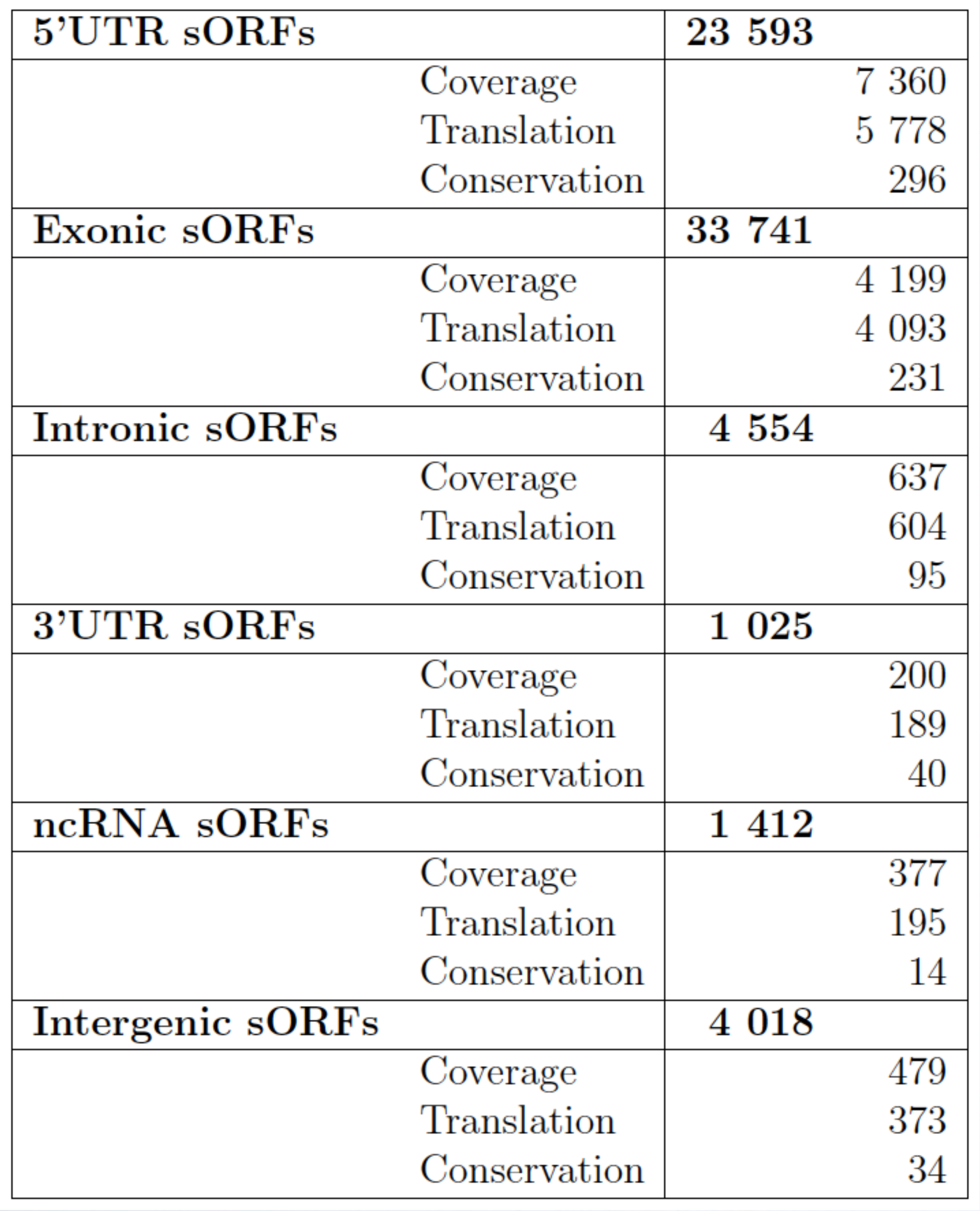
5 Floss Score



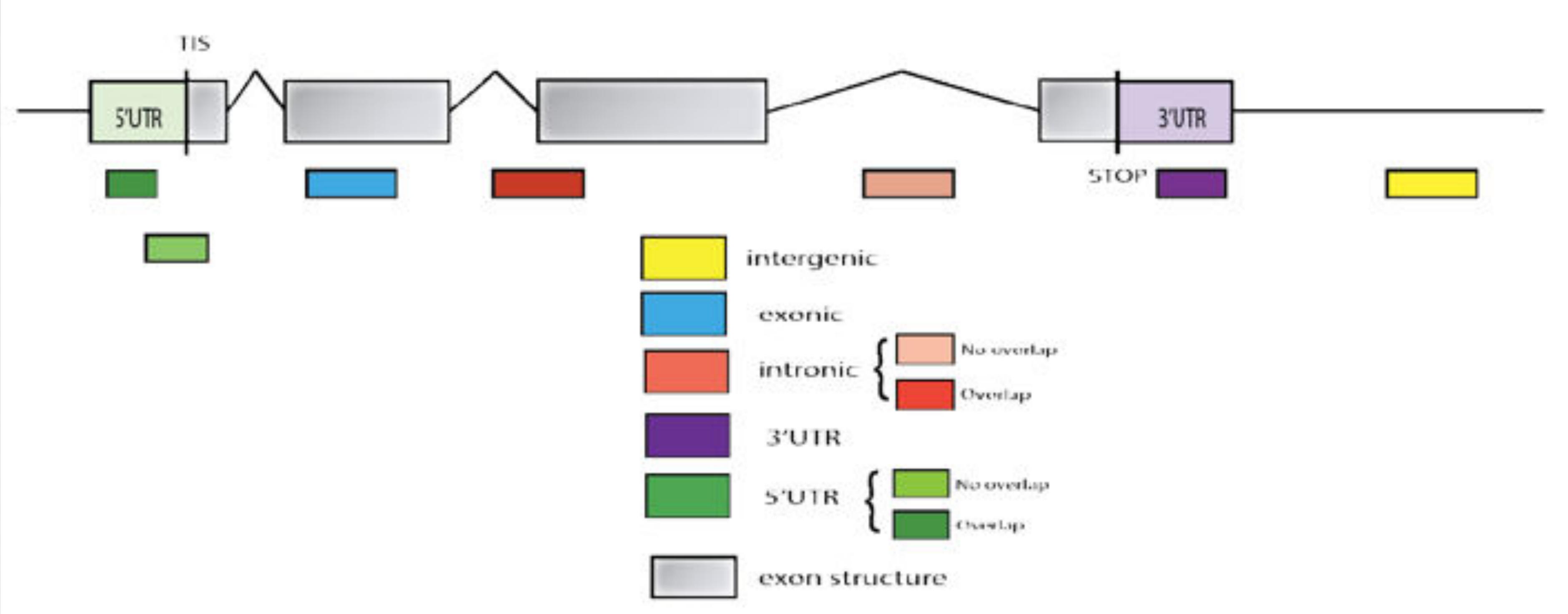
FLOSS: $\sum_l || f_A(l) - f_B(l) ||$

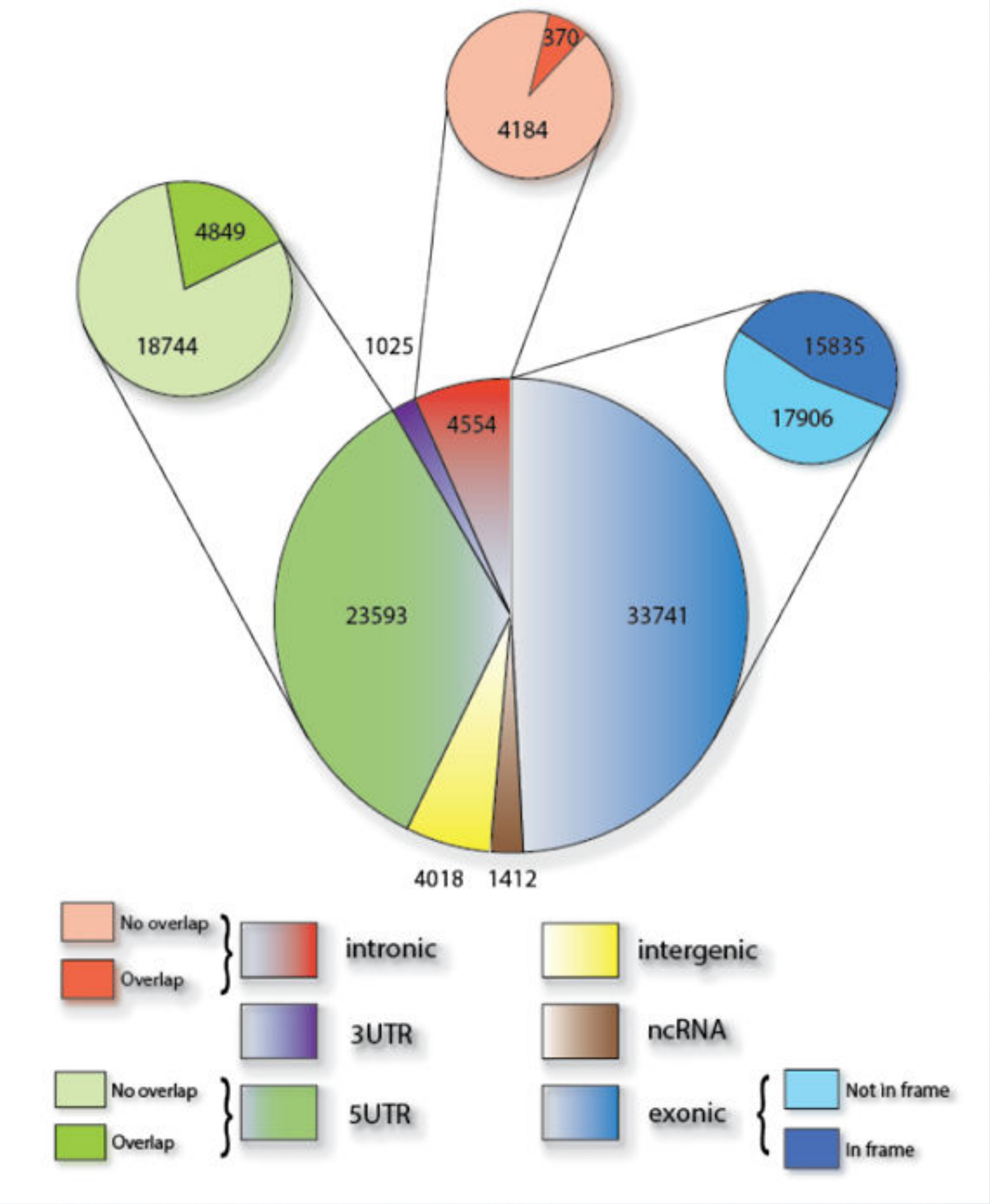
Custom Perl/R

RIBOsORF: Results



5'UTR sORFs		**23 593**
	Coverage	7 360
	Translation	5 778
	Conservation	296
Exonic sORFs		**33 741**
	Coverage	4 199
	Translation	4 093
	Conservation	231
Intronic sORFs		**4 554**
	Coverage	637
	Translation	604
	Conservation	95
3'UTR sORFs		**1 025**
	Coverage	200
	Translation	189
	Conservation	40
ncRNA sORFs		**1 412**
	Coverage	377
	Translation	195
	Conservation	14
Intergenic sORFs		**4 018**
	Coverage	479
	Translation	373
	Conservation	34





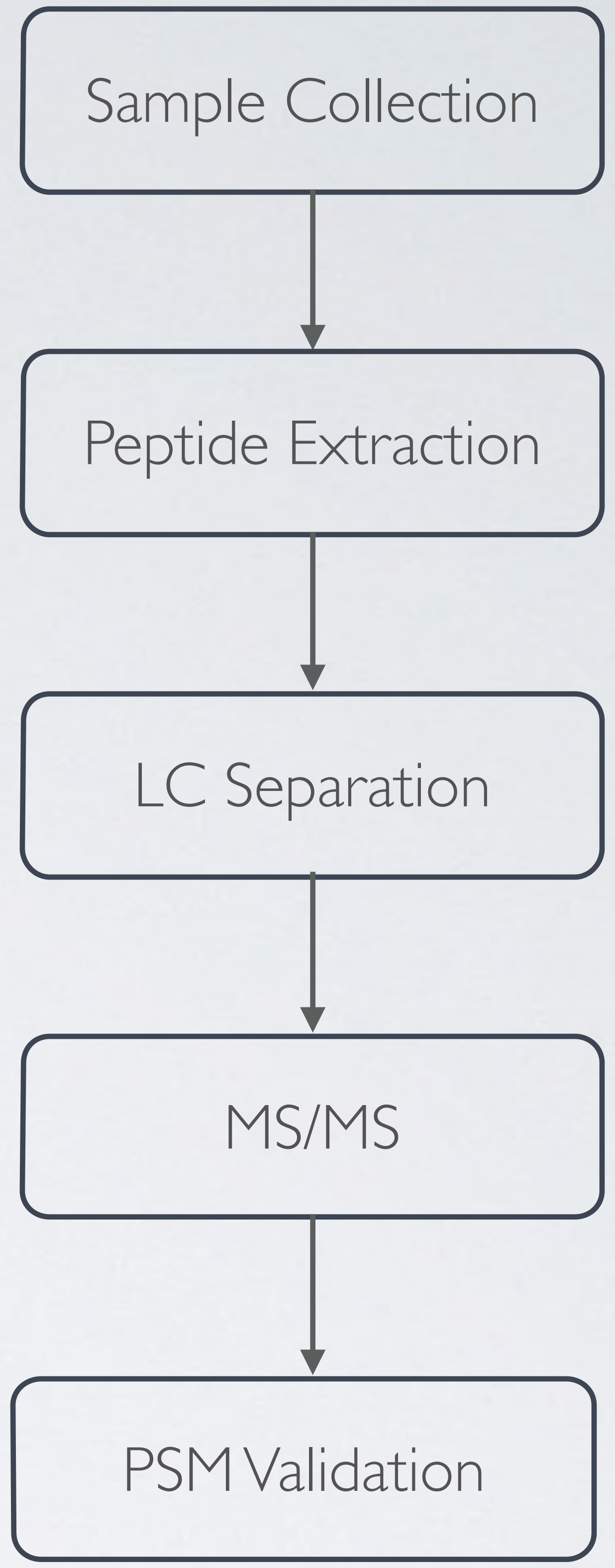
RIBOsORF: MS VALIDATION

➡ specific **enrichment, extraction** and **separation**

- Exosome enrichment
- Synaptosome separation
- Specific cell lines

- 90/9/1 (methanol/H20/acid) extraction solvent
- heat-inactivation
- size filtration (3kDa/ 10 kDa)

- 2D LC separation (RP-LC at high/low pH)

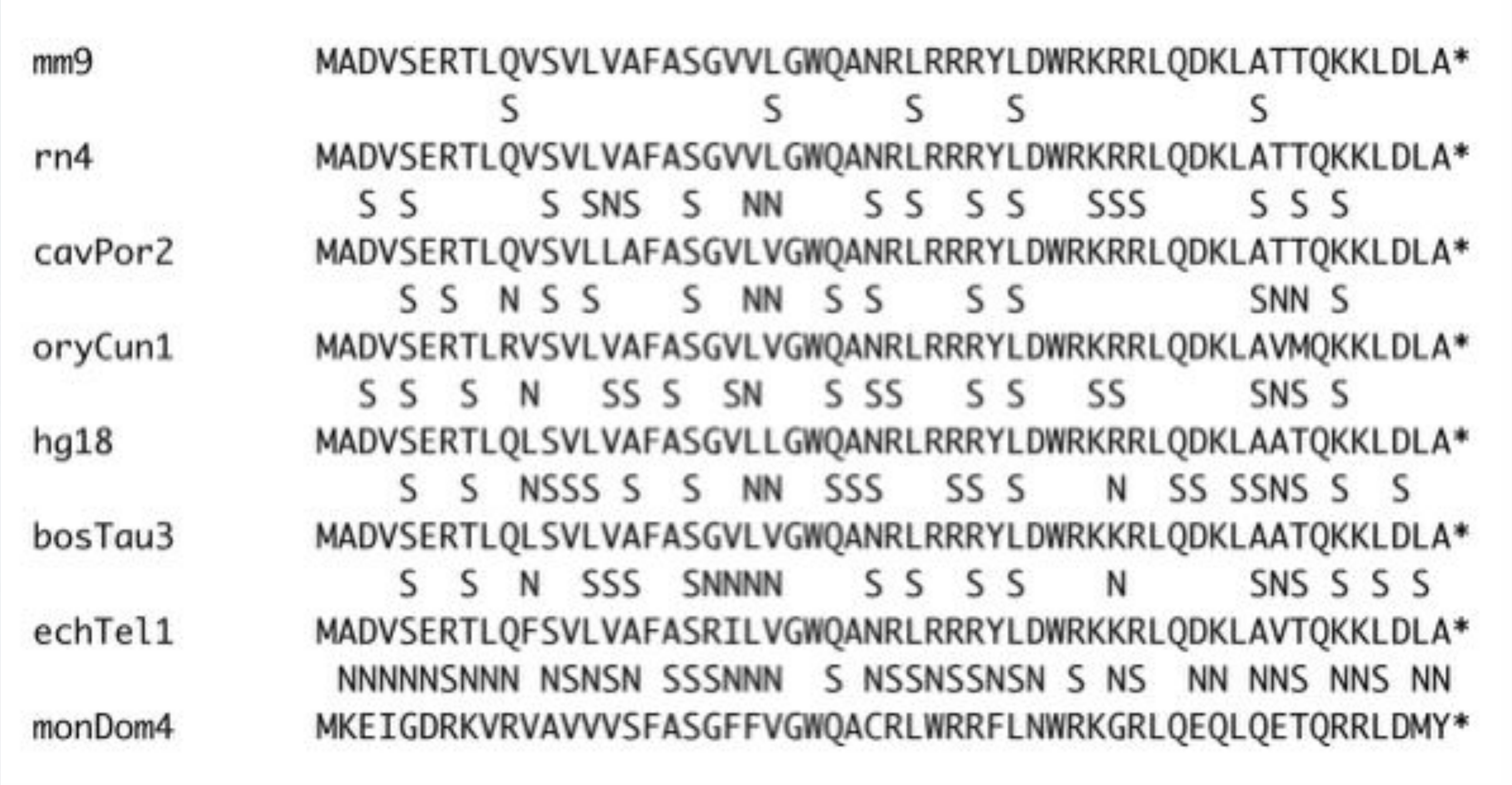

Sample Collection
Peptide Extraction
LC Separation
MS/MS
PSM Validation

➡ search **in-house and public** (**PRIDE Reshake**) MS data

RIBOsORF: Results

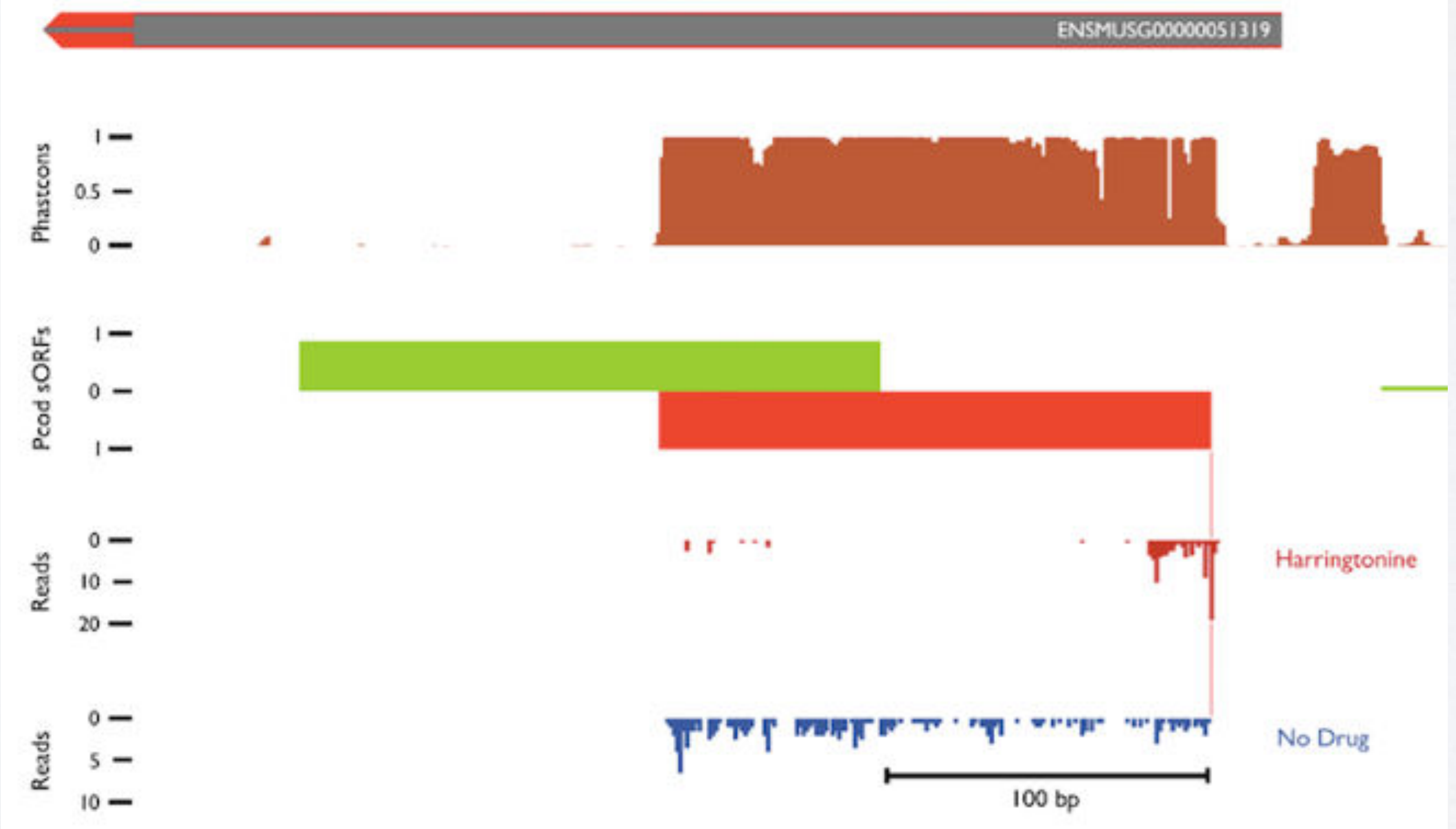
lnc-NPHP1-1:5 sORF

good PhyloCSF



```
mm9         MADVSERTLQVSVLVAFASGVVLGWQANRLRRRYLDWRKRRLQDKLATTQKKLDLA*
                     S            S      S    S           S
rn4         MADVSERTLQVSVLVAFASGVVLGWQANRLRRRYLDWRKRRLQDKLATTQKKLDLA*
              S S      S SNS  S  NN    S S  S S   SSS     S S S
cavPor2     MADVSERTLQVSVLLAFASGVLVGWQANRLRRRYLDWRKRRLQDKLATTQKKLDLA*
                S S  N S S    S  NN  S S    S S           SNN S
oryCun1     MADVSERTLRVSVLVAFASGVLVGWQANRLRRRYLDWRKRRLQDKLAVMQKKLDLA*
              S S  S  N   SS S  SN   S SS   S S   SS      SNS S
hg18        MADVSERTLQLSVLVAFASGVLLGWQANRLRRRYLDWRKRRLQDKLAATQKKLDLA*
                S  S  NSSS S  S  NN  SSS   SS S    N  SS SSNS S  S
bosTau3     MADVSERTLQLSVLVAFASGVLVGWQANRLRRRYLDWRKKRLQDKLAATQKKLDLA*
                S  S  N  SSS  SNNNN    S S  S S    N      SNS S S S
echTel1     MADVSERTLQFSVLVAFASRILVGWQANRLRRRYLDWRKKRLQDKLAVTQKKLDLA*
             NNNNNSNNN NSNSN SSSNNN  S NSSNSSNSN S NS  NN NNS NNS NN
monDom4     MKEIGDRKVRVAVVVSFASGFFVGWQACRLWRRFLNWRKGRLQEQLQETQRRLDMY*
```

good RIBOseq coverage



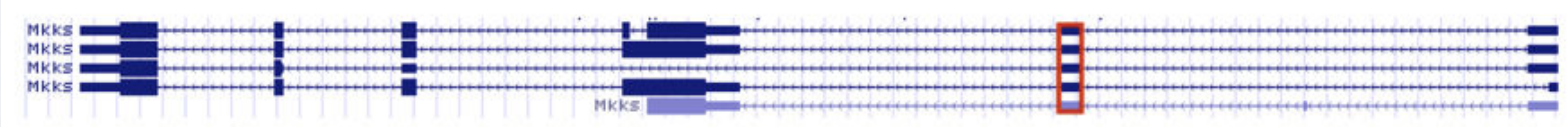
lnc-SPATA21-1:2 sORF

good PhyloCSF

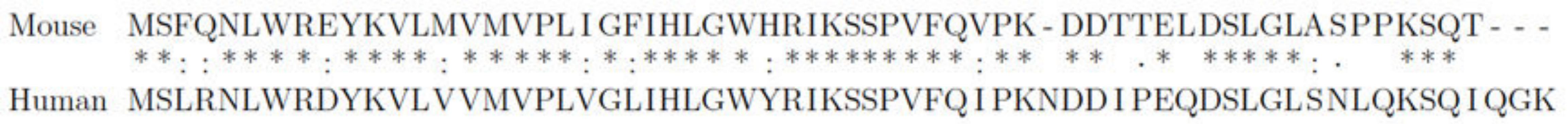
good RIBOseq coverage

MS validation: PRIDE ReShake

mKKS uORF

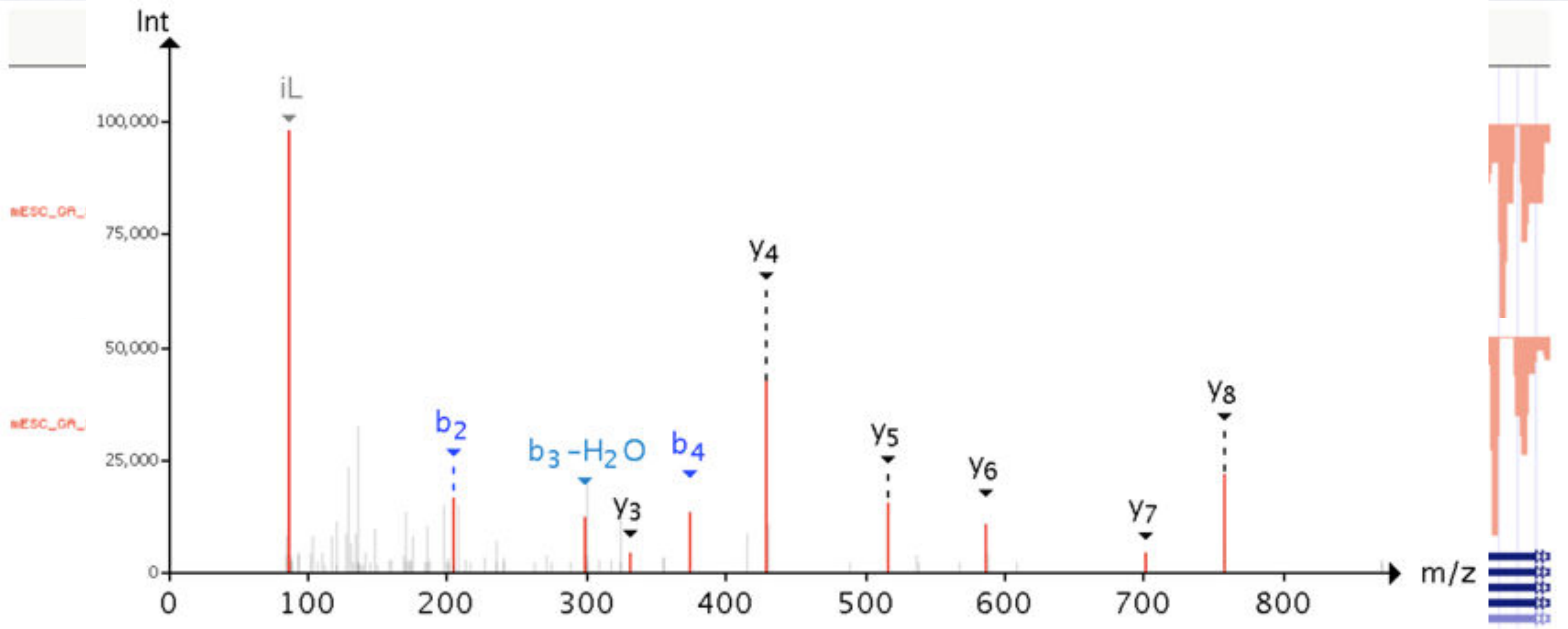


good PhyloCSF



```
Mouse  MSFQNLWREYKVLMVMVPLIGFIHLGWHRIKSSPVFQVPK-DDTTELDSLGLASPPKSQT---
       **::****:****:*****:*:**** *:*********:**  **  .*  *****:.   ***
Human  MSLRNLWRDYKVLVVMVPLVGLIHLGWYRIKSSPVFQIPKNDDIPEQDSLGLSNLQKSQIQGK
```

good RIBOseq coverage and MS validation



Attention while validating the sORF identifications!

- **uORFs** or **5’extension**? (NELFB)
- **below threshold** scoring PSM vs. **reference** DB
- Mouse SMIM20 (69 AA): **trEMBL** to **Swiss-Prot** (re-annotate)
- sORF or **variant reference peptide**?

Search multiple DB’s (multi-stage)

Swiss-Prot
trEMBL
RefSeq
Ensembl
CRAP-DB

Check for possible contaminations

Isobaric masses (I,L)
Near-isobaric masses (K, Q)
2 AA equal 1 AA (G-A = Q)
PTMs (E = Ace-S)

Automate these validation steps: Work in Progress…

Nesvizhskii A., 2014, Nature Methods

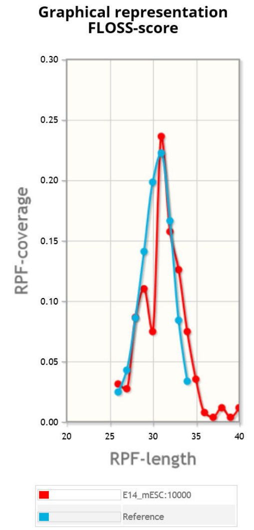
SORFS.ORG

Poster

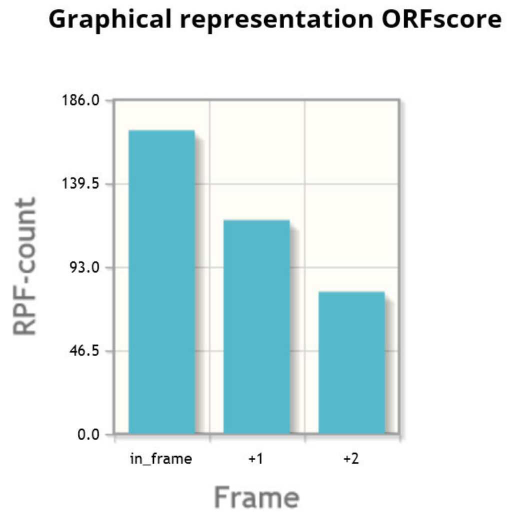
Combining ribosome profiling and proteomics to discover micropeptides, translation products from small open reading frames.

Poster 467
Session: Informatics: Systems Biology
10:30am - 2:30pm
Thu, Jun 04

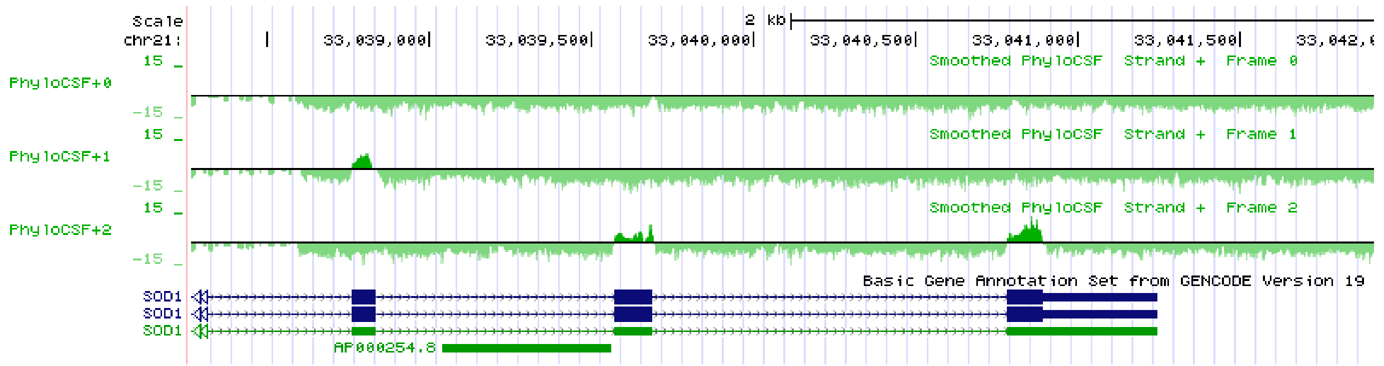
A.



B.



C.

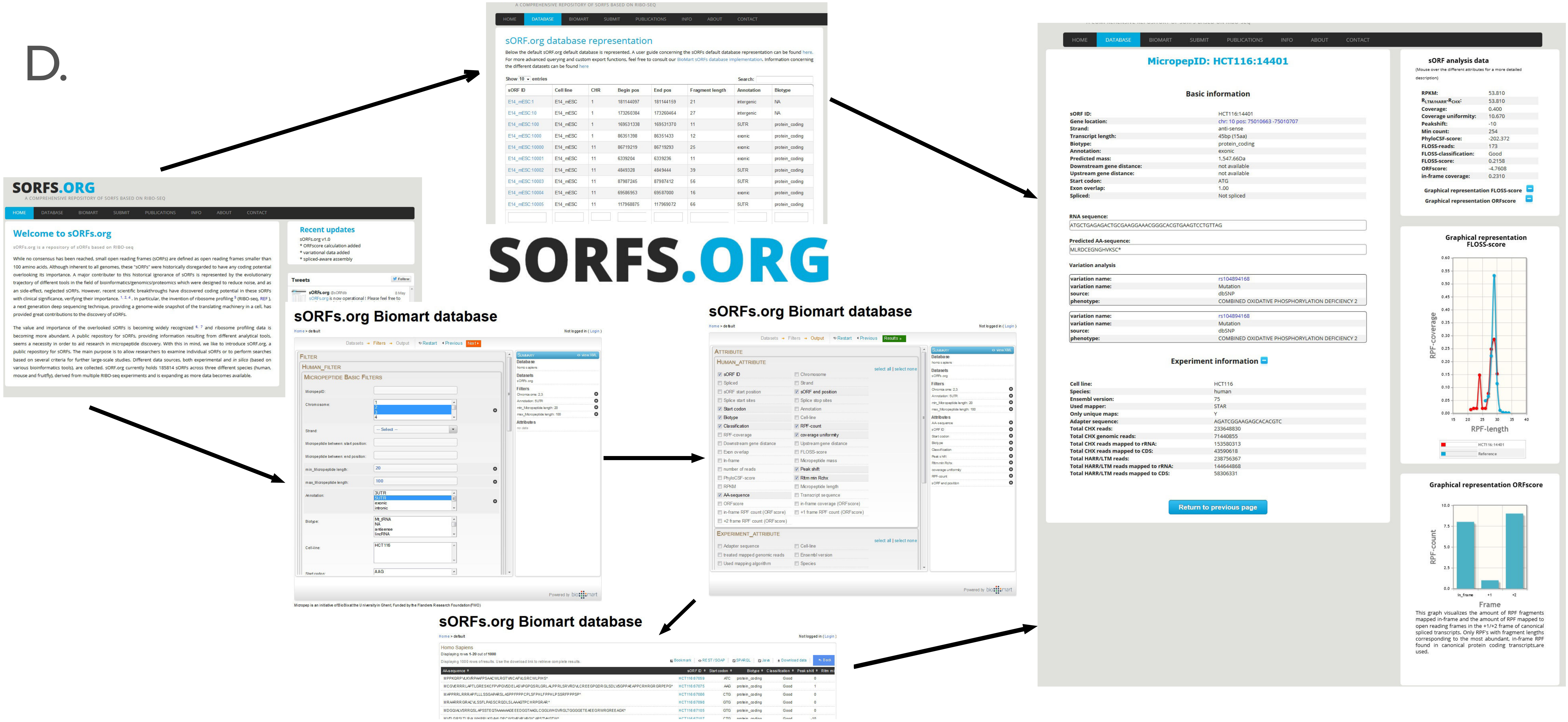


[C] FLOSS[4]: The FLOSS algorithm distinguises true coding from non-coding sequences based on the ***RPF-length distribution***.

[D] ORFscore[5]: The ORFscore calculates the preference of RPFs to accumulate in the ***first frame*** of coding sequences.

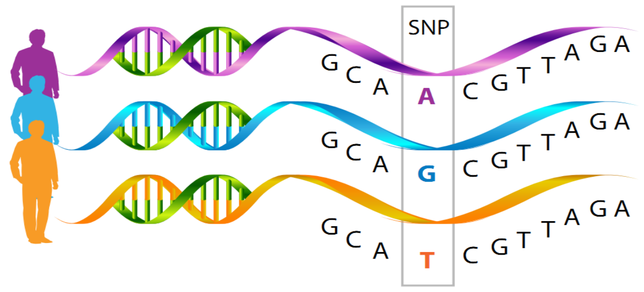
[E] PhyloCSF[6] :Cross species ***conservation*** is a general adopted technique in order to acquire evidence for genomic important regions.

D.



[F] sORFs.org: A ***public repository*** for sORFs based on RIBO-seq. Providing a platform for scientist to inspect, query and retrieve information regarding sORFs.

E.



F.

[G] Variation analysis: Information embedded in phenotype related ***mutations, insertions and deletions*** provide an additional source for evidence in favor of the coding potential of sORFs.

[H] Automated PRIDE resprocessing[7][8][9]: The PRIDE[10] database is explored to find ***Mass spectrometry*** evidence for the translation of sORFs into functional micropeptides.

CONCLUSIONS

✓ **Multi-omis** approaches help the identification of novel proteoforms

✓ **PROTEOFORMER** pipeline: www.biobix.be/proteoformer
✓ **RIBOsORF** pipeline and sORFs DB: www.sORFs.org

Galaxy

ACKNOWLEDGEMENTS

- Biobix, Lab for Bioinformatics and Computational Genomics
  Faculty of Bioscience Engineering
  UGent

  Jeroen CRAPPÉ
  Alexander KOCH
  Geert TROOSKENS
  Elvis NDAH
  Sandra STEYAERT
  Wim VAN CRIEKINGE
  Volodimir OLEXIOUK
  Steven VERBRUGGEN

- Dept. Medical Protein Research, VIB-UGent

  Petra VAN DAMME
  Daria GAWRON
  Kris GEVAERT

- CFP, Center for Proteomics, Antwerp University / VITO

  Geert BAGGERMAN

- Computational Omics and Systems Biology Group, VIB-UGent

  Lennart MARTENS
  Harald BARSNES
  Marc VAUDEL

- NXTGNT

  Sarah DE KEULENAER
  Ellen DE MEESTER

Gerben.Menschaert@Ugent.be