



Identifying Novel Genes and Gene Refinements in *Thermoanaerobacter tengcongensis*

Presented by Kun Zhang

Institute of Computing Technology, CAS, Beijing, China
<http://pfind.ict.ac.cn>

2014-06-20



Thermoanaerobacter tengcongensis

- *Thermoanaerobacter tengcongensis* (TTE)
 - Discovered in Tengchong, Yunnan, China
 - Complete genome sequencing in 2002¹, 2.7Mb
- Genome annotation²
 - Computational: *ab initio*, comparative
 - Experimental: RNA-seq, EST, etc. (cDNA-seq related)
 - TTE: 2588 coding genes (through *ab initio*, comparative)

1. Bao Q, *et al.* A complete sequence of the *T. tengcongensis* genome. *Genome Res* **12**, 689-700 (2002).
2. 1. Brent MR. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet* **9**, 62-73 (2008).

Background

- Deficiencies of the
 - High error rate in the genes predicted
 - Discriminate coding genes from non-coding ones
 - Explain the translational event, such as alternative translational initiate site, translational UTR (untranslational region)

Background

- Another experimental evidence: proteomics data (LC-MS/MS based)
- For genome annotation: Proteogenomics
 - Direct evidence of ORF translation
 - Find novel genes and refine gene models that the traditional approaches could not reach

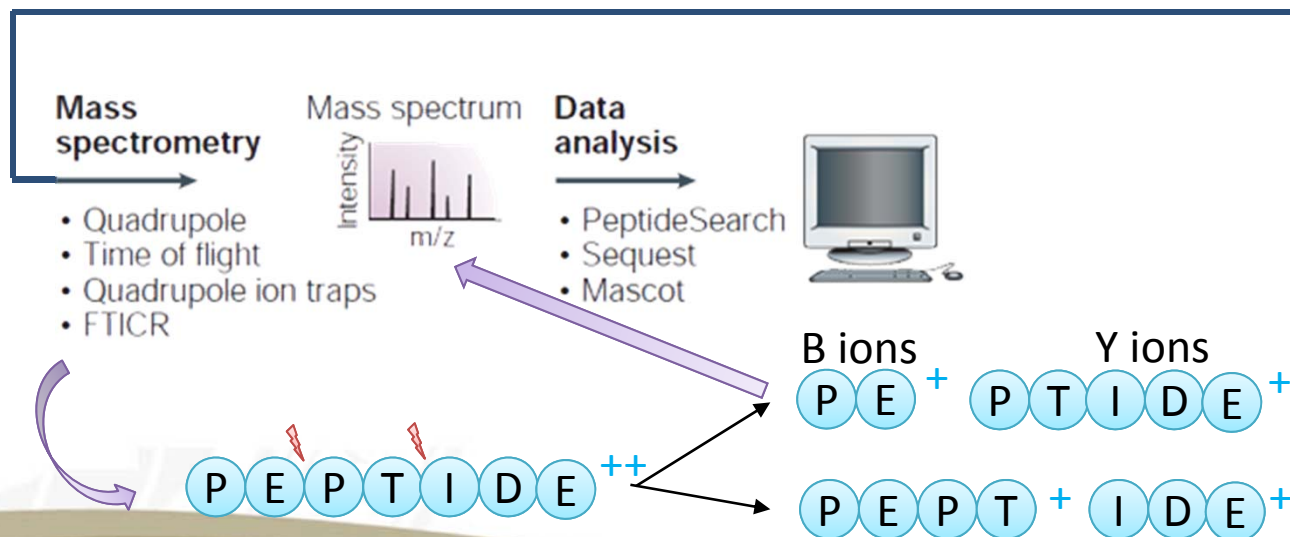
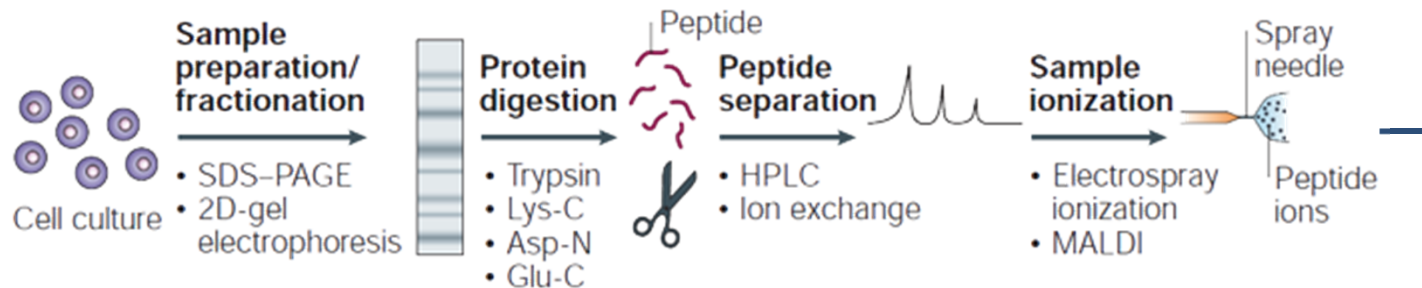
Utilize proteomics data to find novel genes and gene refinements in TTE.

1. Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R. Generating and navigating proteome maps using mass spectrometry. *Nat Rev Mol Cell Biol* 11, 789-801 (2010).
2. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 15, 193-204 (2014).



LC-MS/MS based proteomics

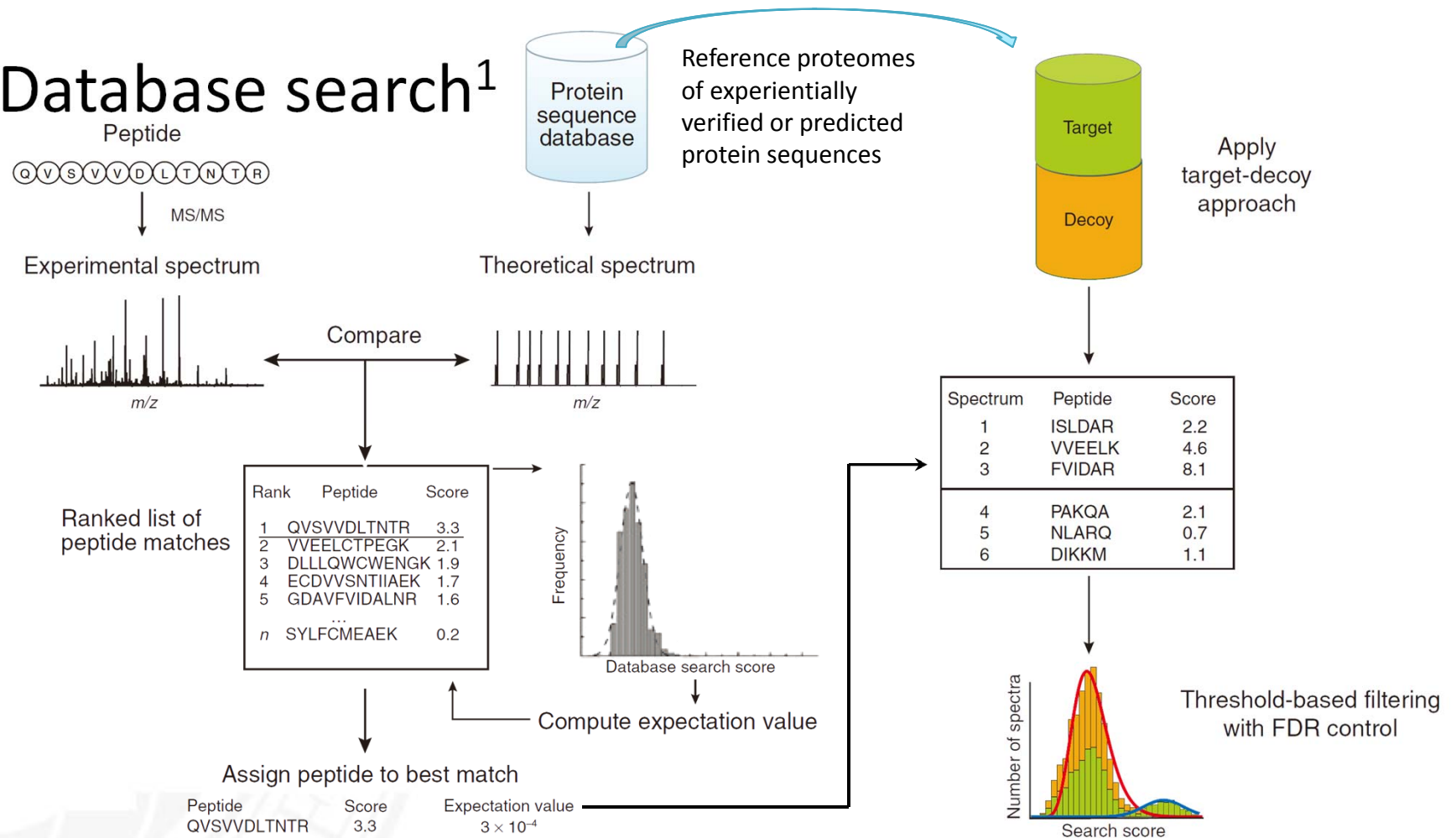
- LC-MS/MS based proteomics¹



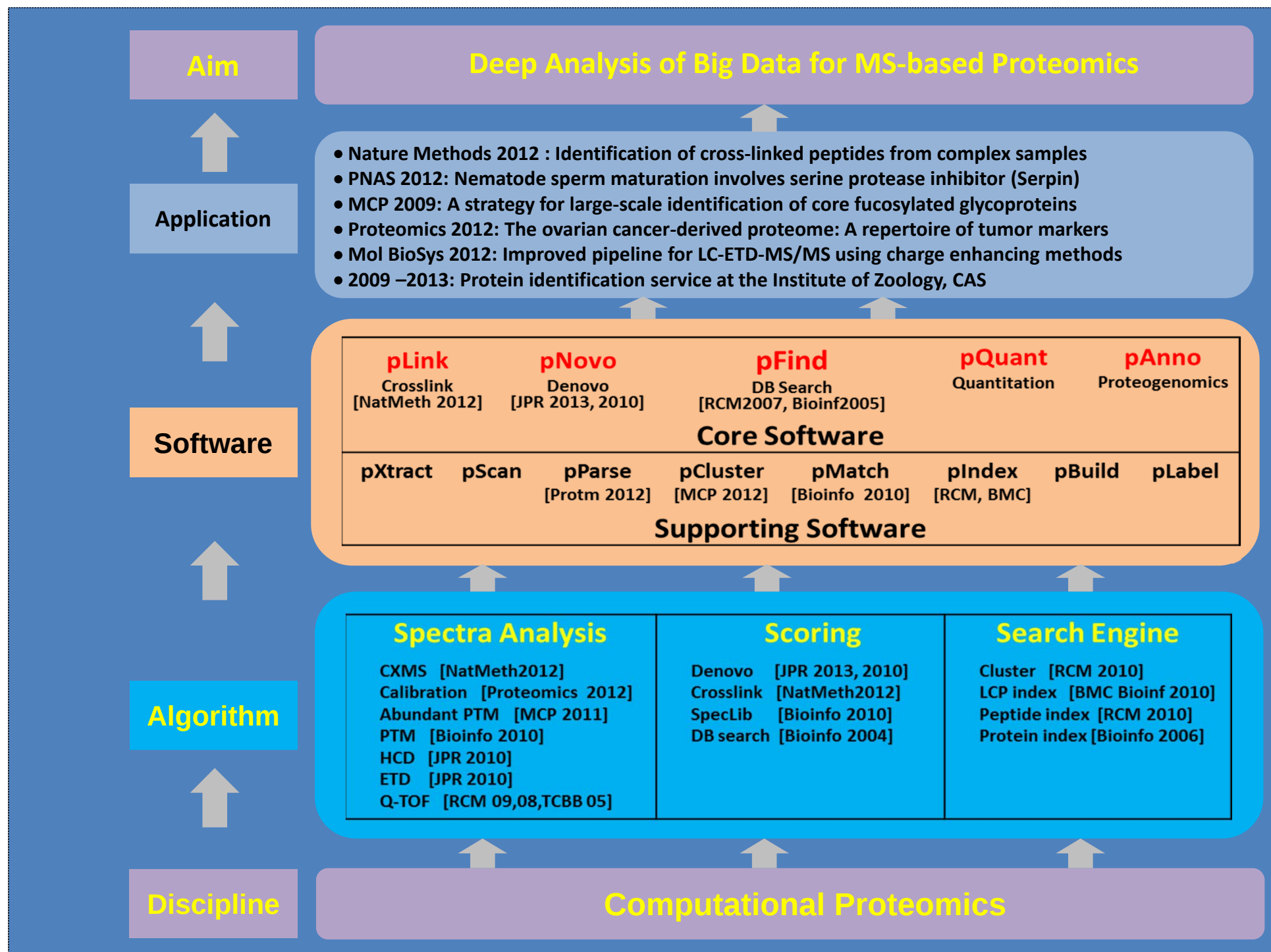
1. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5, 699-711 (2004).

Searching the protein database

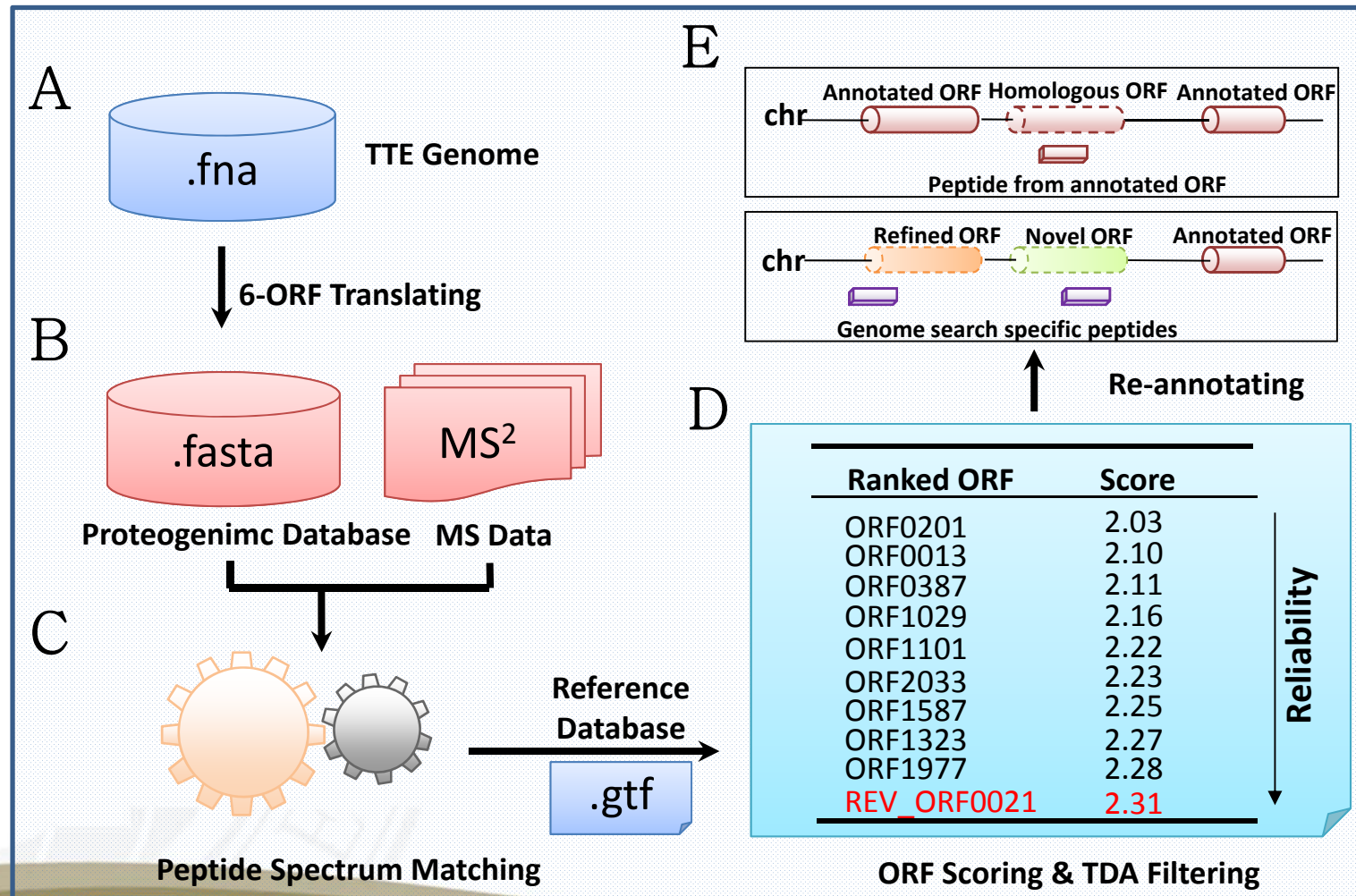
• Database search¹



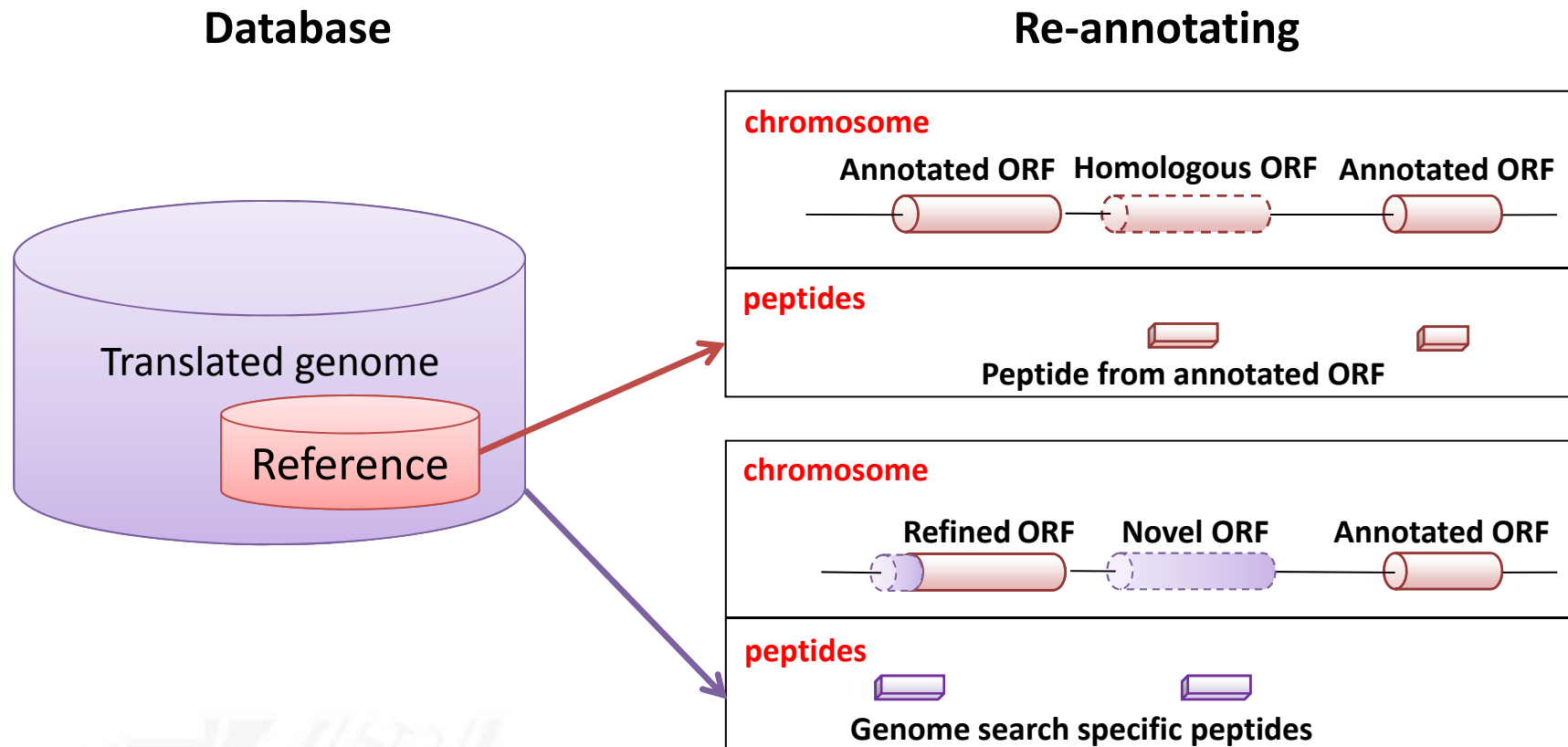
1. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 4, 787-797 (2007).



Proteogenomic analysis



Novel genes and gene refinements



Data and database

- Database: translated TTE genome
- Data
 - Mass spectrometry: Q-Exactive (HCD-FT)
 - Tandem mass spectra: ~5 million
- Search parameter
 - Engine: pFind v3.0
 - Default params for HCD-FT mode
- Quality control: ORF level FDR <1%



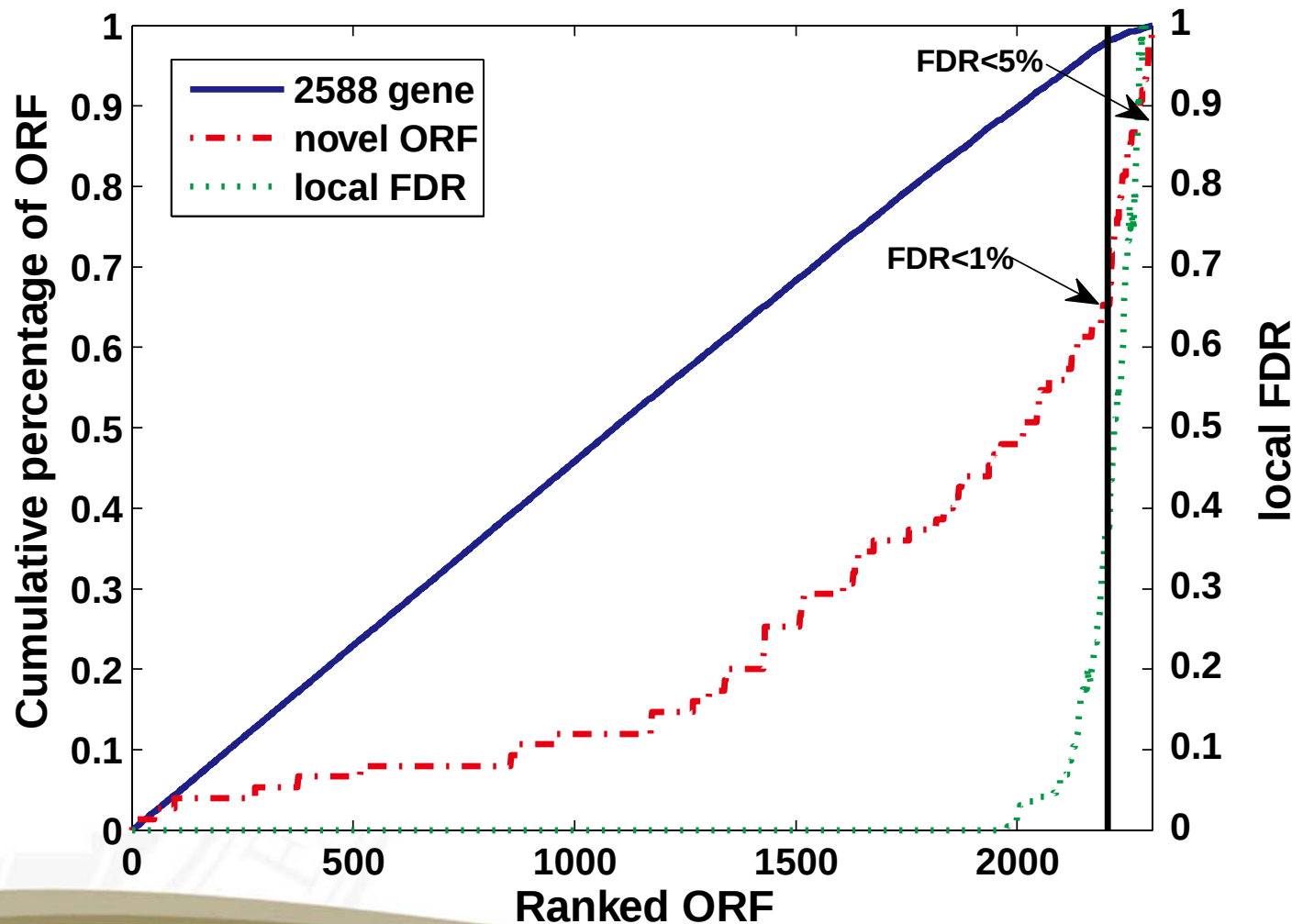
All ORFs identified

Table. All ORFs identified in this study.

Category	Number	Ratio
Annotated ORF	2002	91%
Refined ORF	112	5%
Novel ORF	49	2%
Homologous ORF	42	2%

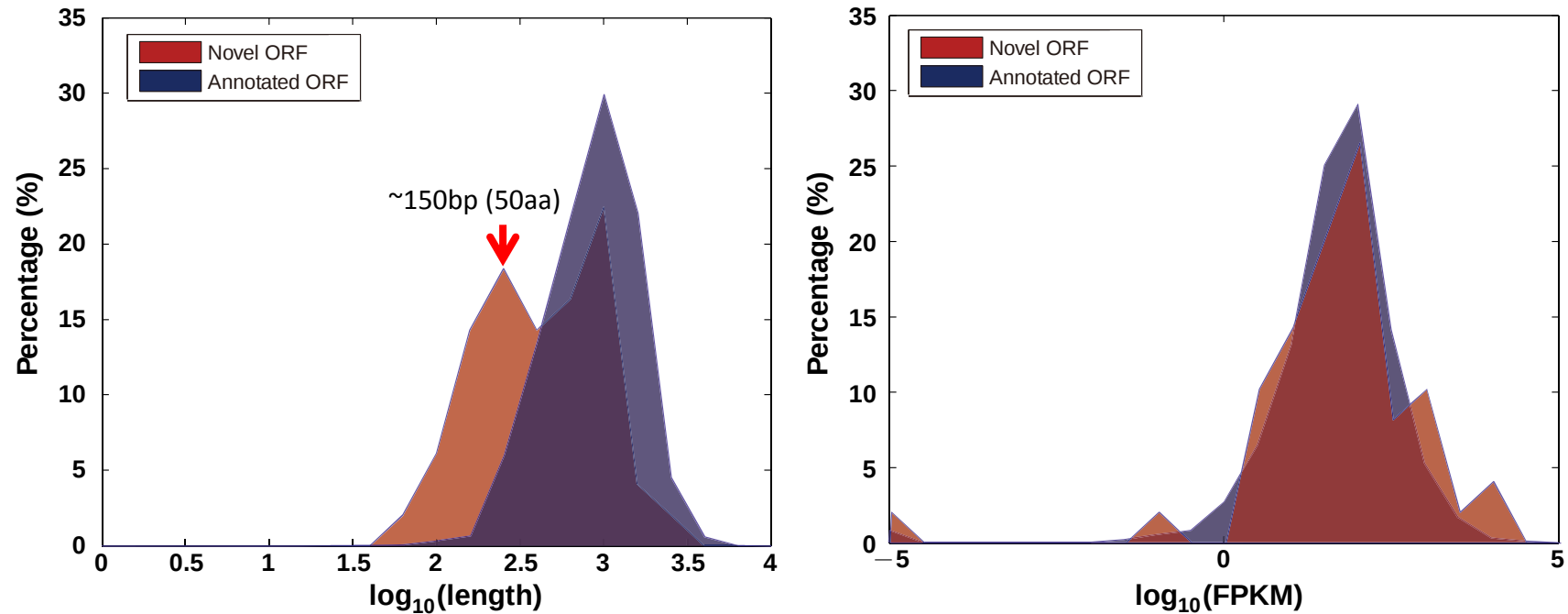
- Validation of novel ORFs
 - Computational: the local FDR(posterior error probability) measures the reliability of each ORF was calculated
 - Transcriptional evidence: the strand specific RNA-seq data was also used as a supporting evidence

Distribution of novel and reference ORFs



The cumulative percentage of novel and annotated ORFs, together with their local FDRs.

Novel ORFs in length



The length (bp) and RNA expression of the novel ORFs

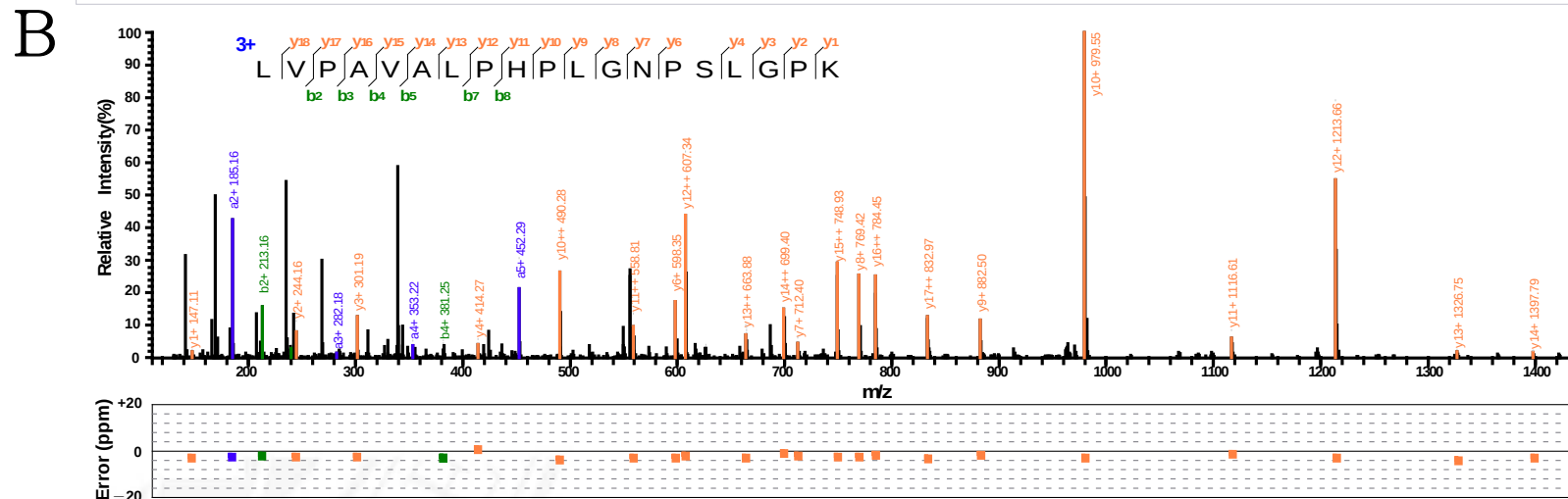
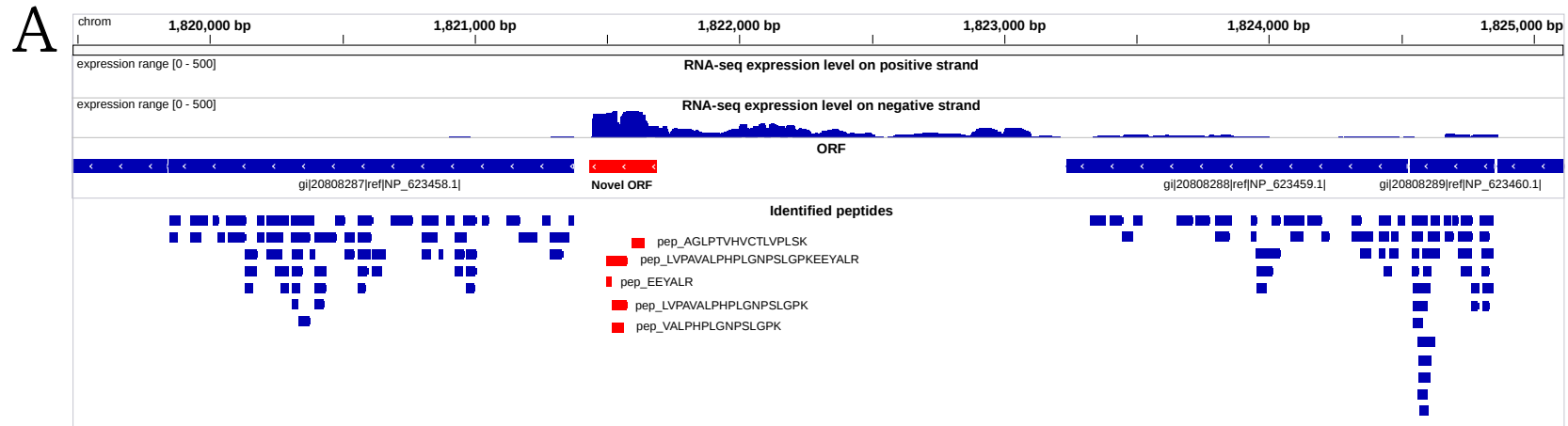
Genomic location

- Novel ORFs were shorter in length (sORF)
- Classification based on the distance (1000kb) to annotated ORFs
 - UTR translation or operon structure (46)
 - Intragenic regions (1)
 - Overlapping with annotated ORFs (2)

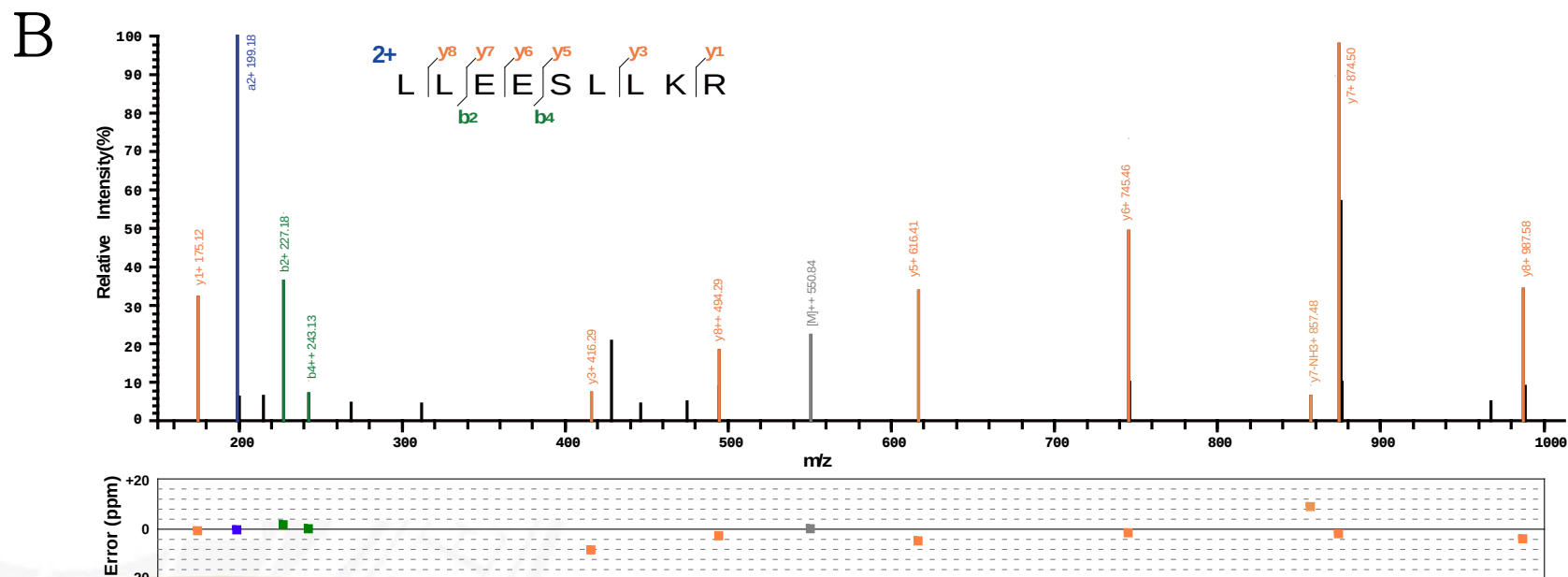
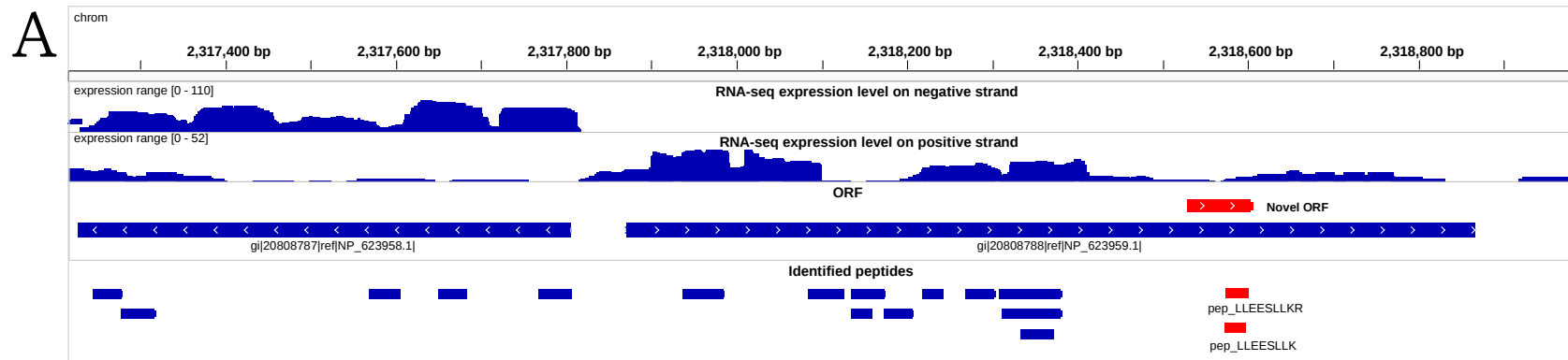
- Examples of identified novel ORFs
 - UTR translation
 - Overlapping ORF
 - Non-canonical start codon



UTR translation

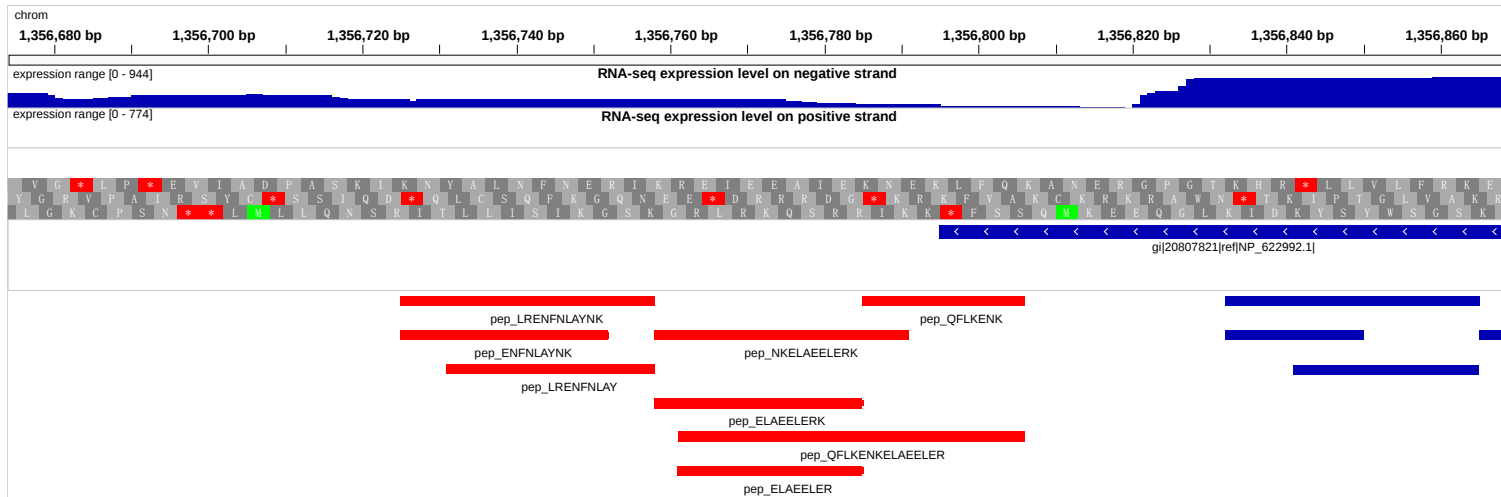


Overlapping ORF

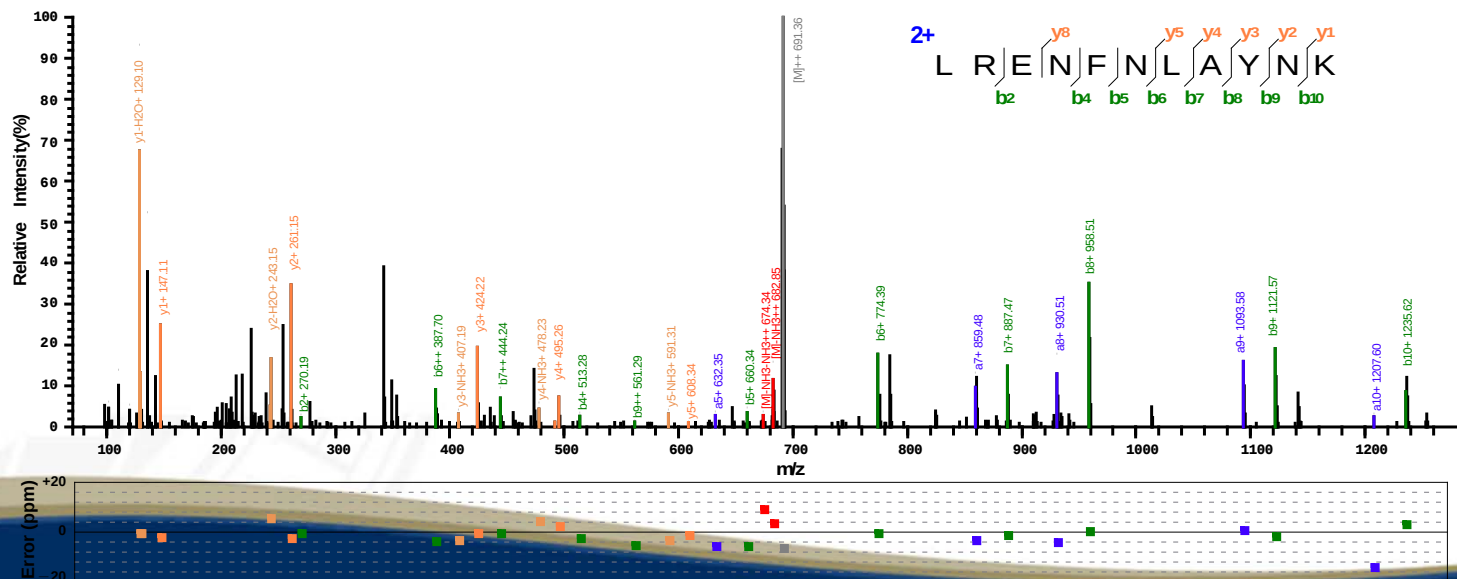


Non-canonical start codon

A



B



Category	Sensitivity	Accuracy
Transcriptomic	★★★	★★
Computational	★★	★☆
Proteomic	★★☆	★★★★

Conclusion

- 2% ORFs are missing, short ORFs are tend to be eliminated in TTE.
- Most of the novel ORFs identified in this study come from UTR region, operon, non-canonical start codons.
- Proteogenomics is a good source for genome refinement and novel gene discovery, which may beyond the reach of genomic and transcriptomic approaches.

Acknowledgement



**Professor
Si-Min He**



**Professor
Si-Qi Liu**



**Associate Professor
Yi Zhao**



**Professor
Lu Xie**



**Professor
Xiao-Hong Qian**



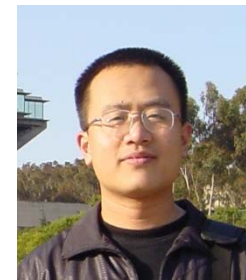
**Associate Professor
Rui-Xiang Sun**



**Associate Professor
Quan-Hui Wang**



**Ph.D. candidate
De-Chao Bu**



**Assistant Researcher
Hao Chi**



**M.D.
Long Wu**



And, many thanks to our pFinders



Q & A

