

Proteogenomics: Finding Fusion Genes from Human Cancer Mass Spectrometry Data

XIE Lu (谢鹭), 2013-06-18

The 11th Japan-Korea-China Bioinformatics
Training Course & Symposium

June 17-18 2013, Soochow University, Suzhou, China



上海生物信息技术研究中心

SHANGHAI CENTER FOR BIOINFORMATION TECHNOLOGY

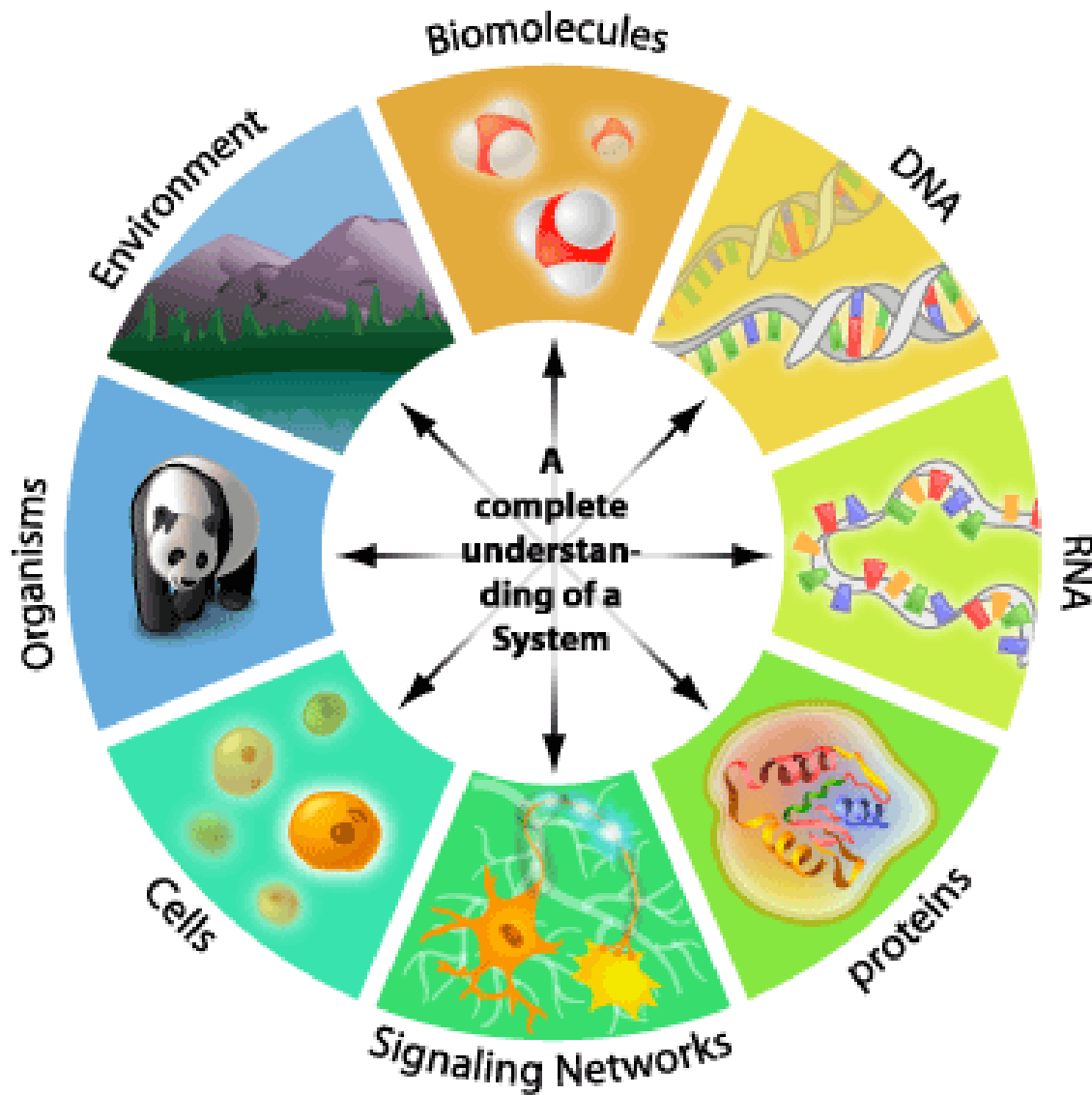




Proteogenomics tutorial

Based on references:

- Santosh Renuse, Raghothama Chaerkady and Akhilesh Pandey. Proteogenomics. ***Proteomics***. 2011, 11, 620–630.
- Natalie Castellana, Vineet Bafna. Proteogenomics to discover the full coding content of genomes: A computational perspective. ***Journal of Proteomics***. 2010, 73, 2124-2135.
- ZHANG Kun, WANG Le-Heng, CHI Hao, BU De-Chao, YUAN Zuo-Fei, LIU Chao, FAN Sheng-Bo, CHEN Hai-Feng, ZENG Wen-Feng, LUO Hai-Tao, SUN Rui-Xiang, HE Si-Min, **XIE Lu**, ZHAO Yi. Proteogenomics: Improving Genomes Annotation by Proteomics. ***Progress in Biochemistry and Biophysics***. 2013, 40(4): 297~308



Genetic Information Flow

DNA Sequences



RNA Transcript



Encoded Proteins



Signaling and Functional Networks



Introduction

- High-throughput and cost-effective genome sequencing approaches have led to completion of more than a thousand genome sequences with the sequencing of thousands of additional species underway.
- Genome annotations of a majority of sequenced genomes are mostly based on predictions, do not guarantee that the corresponding proteins do exist.



- ‘Proteogenomics’ refers to the correlation of the proteomic data with the genomic and transcriptomic data with the goal of enhancing the understanding of the genome.



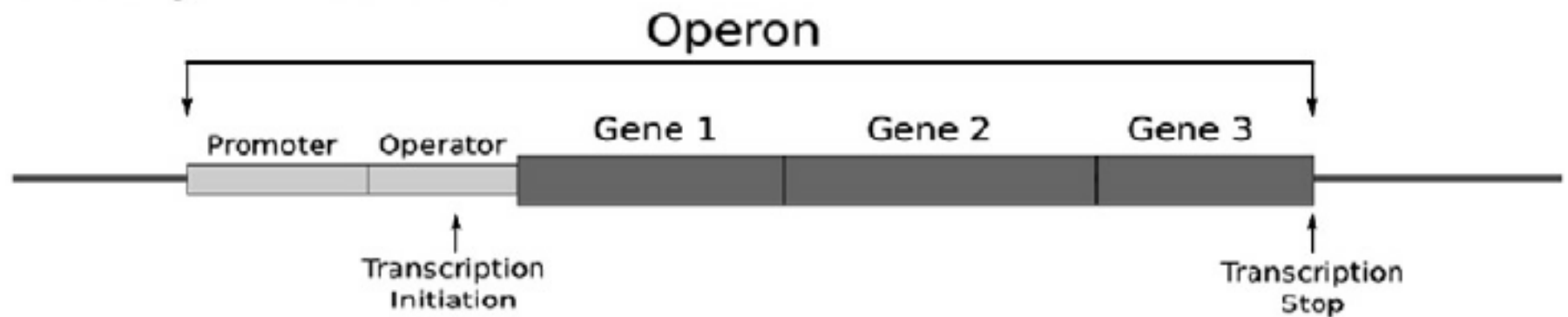
Why proteogenomics?

- Although many gene prediction programs are available, accurate gene identification in a given organism decreases drastically from the nucleotide level to exons to whole gene structures.
- Gene finding in prokaryotes is relatively easier owing to their compact genomes with simple gene structure.
- Gene finding in eukaryotes is difficult because of introns and complex regulatory regions.
- Most often short genes are missed, alternative splice isoforms are also difficult to predict.



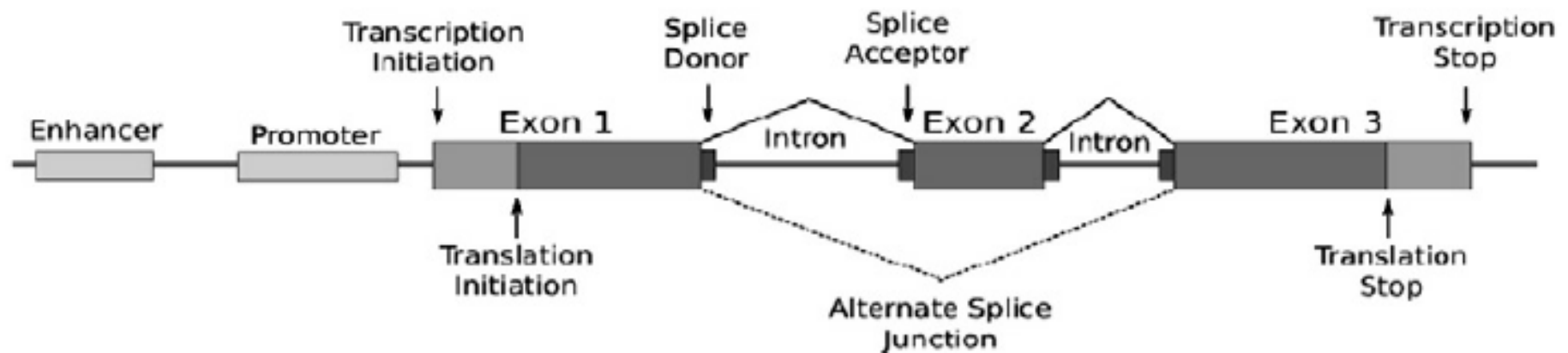
Prokaryotic Genes

A

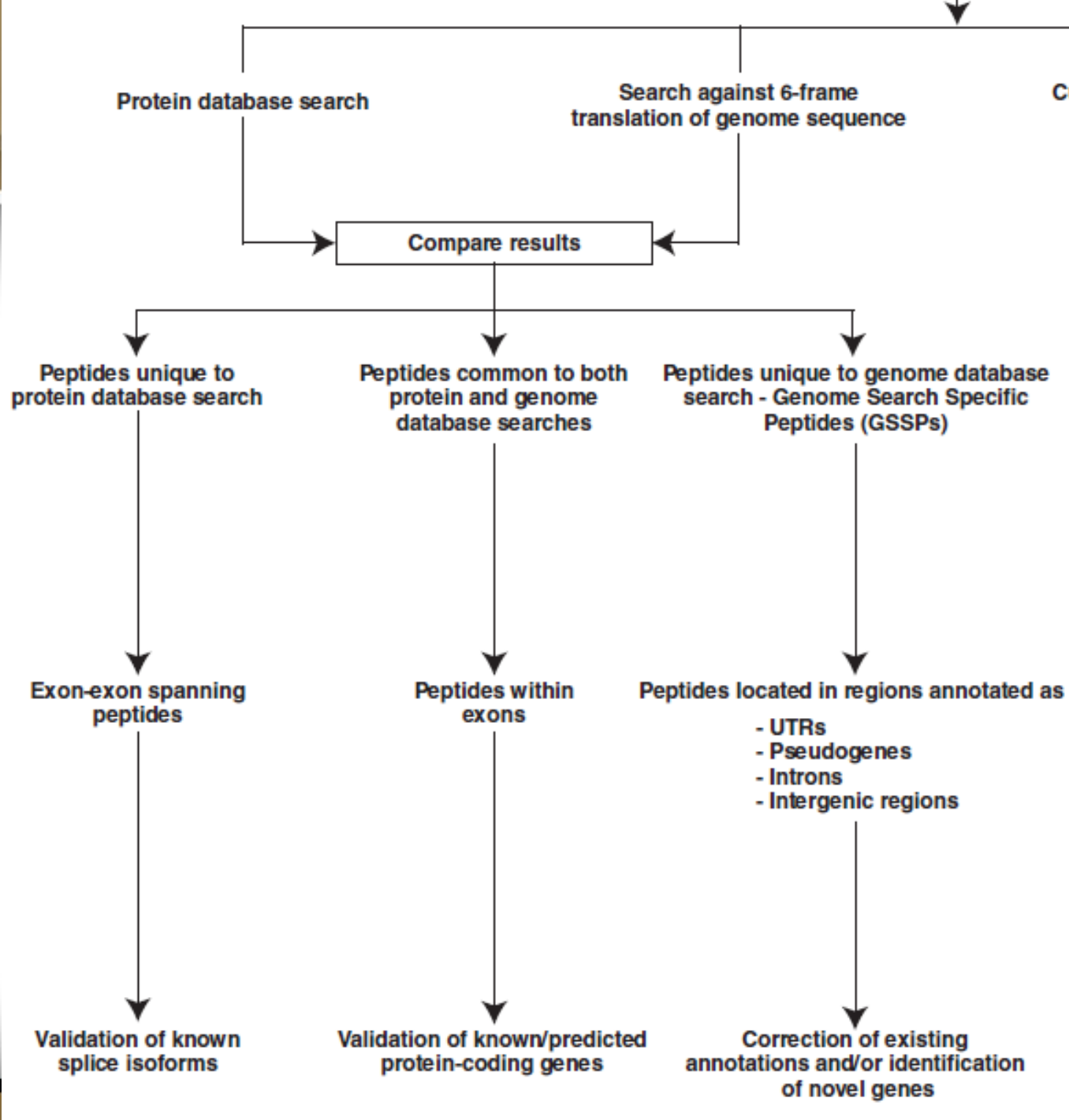


Eukaryotic Gene

B



Work pipeline I



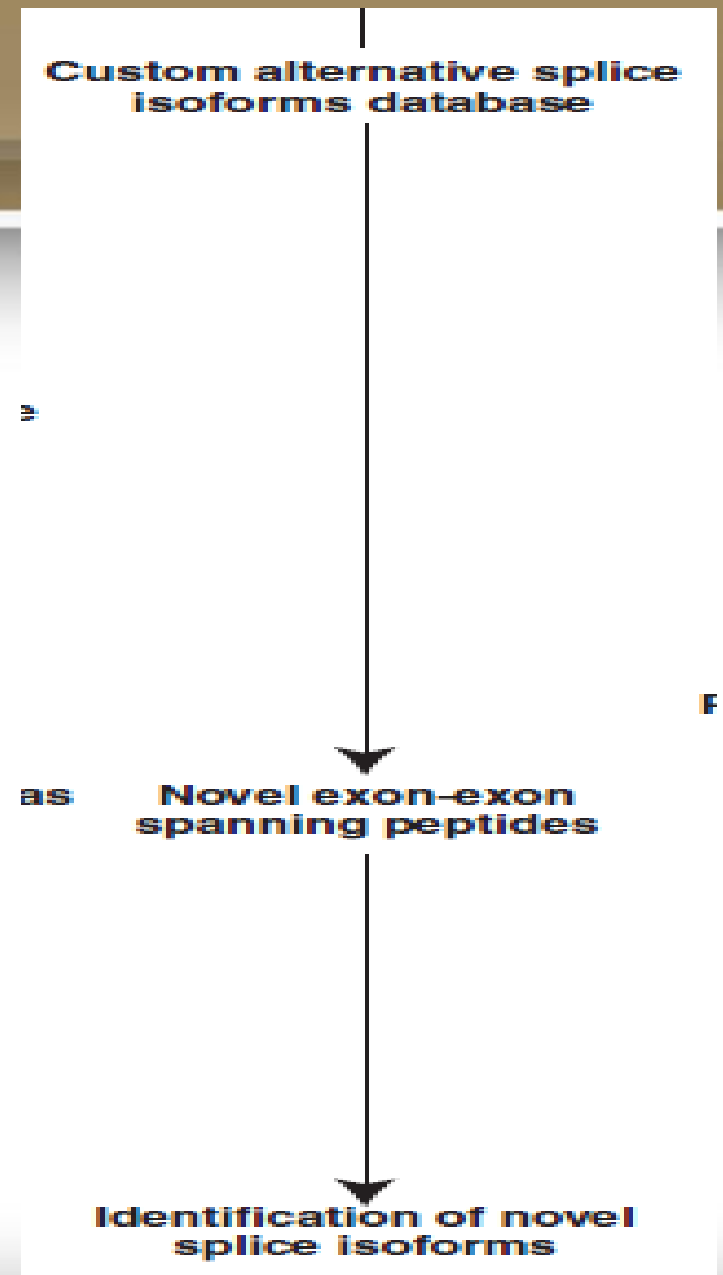
High resolution and
high accuracy
mass spectrometry data



Work pipeline II

High resolution and
high accuracy
mass spectrometry data

Mass spectrometry data are also searched against
custom alternative splice isoform database to
identify novel splice isoforms





Within One Gene---Intragenic

AT2G32235

Translated UTR



AT1G63500

Out of Frame



AT1G67350

Exon Boundary

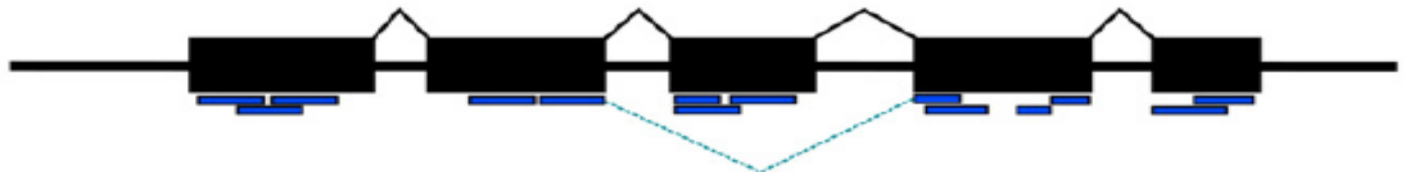


AT1G33790

Novel Exon

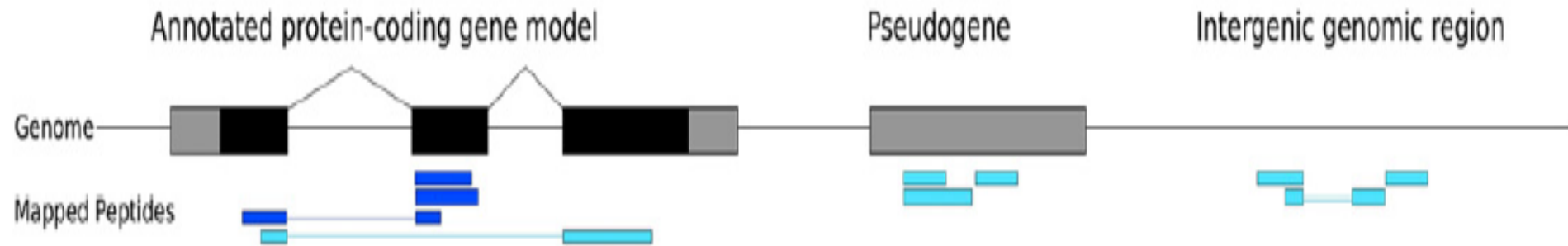


Novel Splice Junction





Between Genes---Intergenic

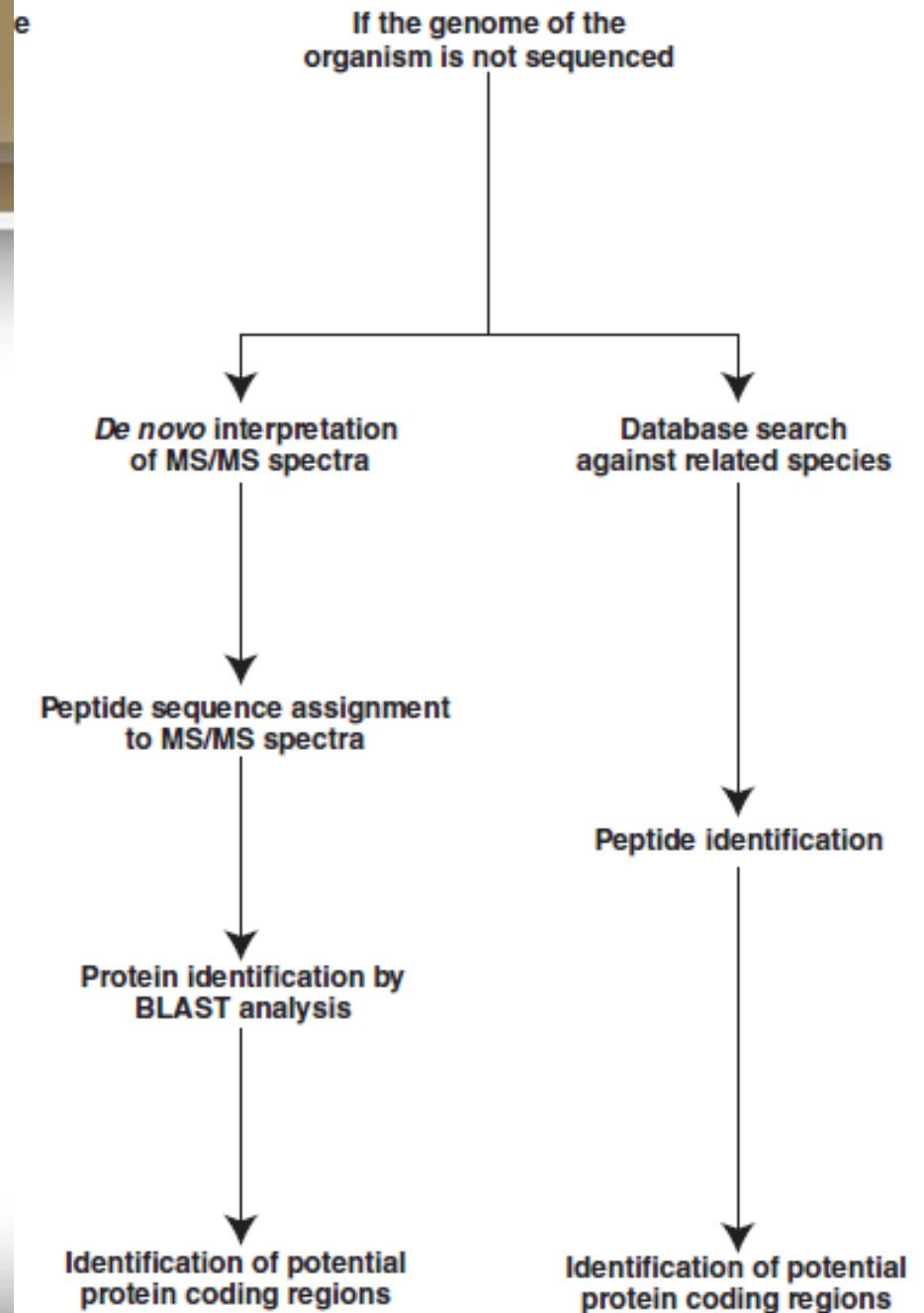




Work pipeline III

**High resolution and
high accuracy
mass spectrometry data**

**Unsequenced genomes can be
analyzed using homology-based or de
novo sequencing based approaches to
identify protein coding regions**

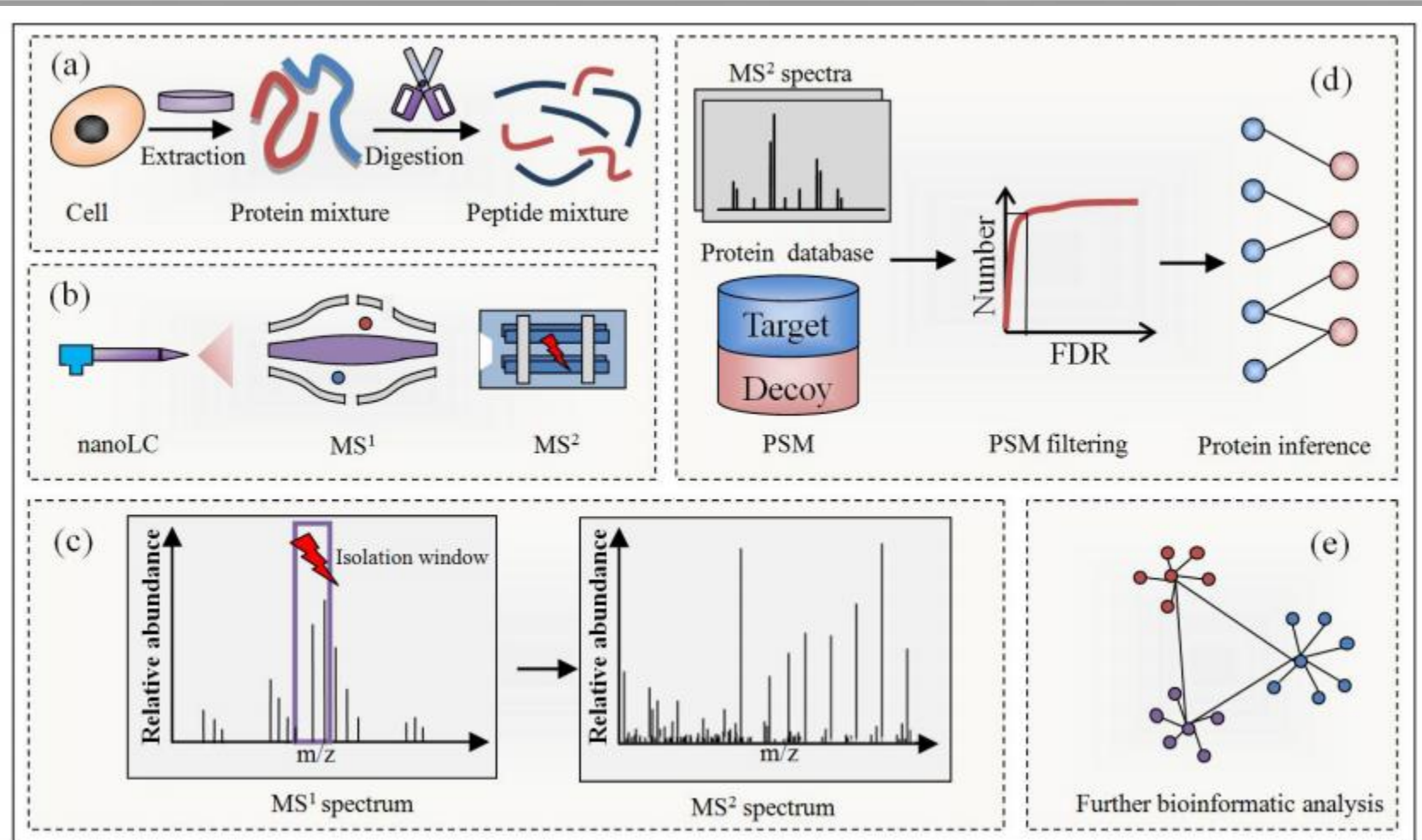




Platforms for proteogenomics

- The advancements in mass spectrometry instrumentation allow acquisition of tandem mass spectral data with **fast scanning capabilities, high resolution and high mass accuracy**, as the basis for carrying out comprehensive proteogenomic analysis.
- **Deeper sequence coverage** can be obtained using multiple sample fractionation methods and mass spectrometry analysis techniques.

MS/MS based Proteomics



Proteogenomics: Improving Genomes Annotation by Proteomics,
Kun Zhang, et al., Prog. Biochem. Biophys, 2012



MS/MS

Instrumentation



Fragmentation



Protein digestion

**Sample
Prepa-
ration**

Protein fractionation
---Depth

Protein extraction
---Broadth

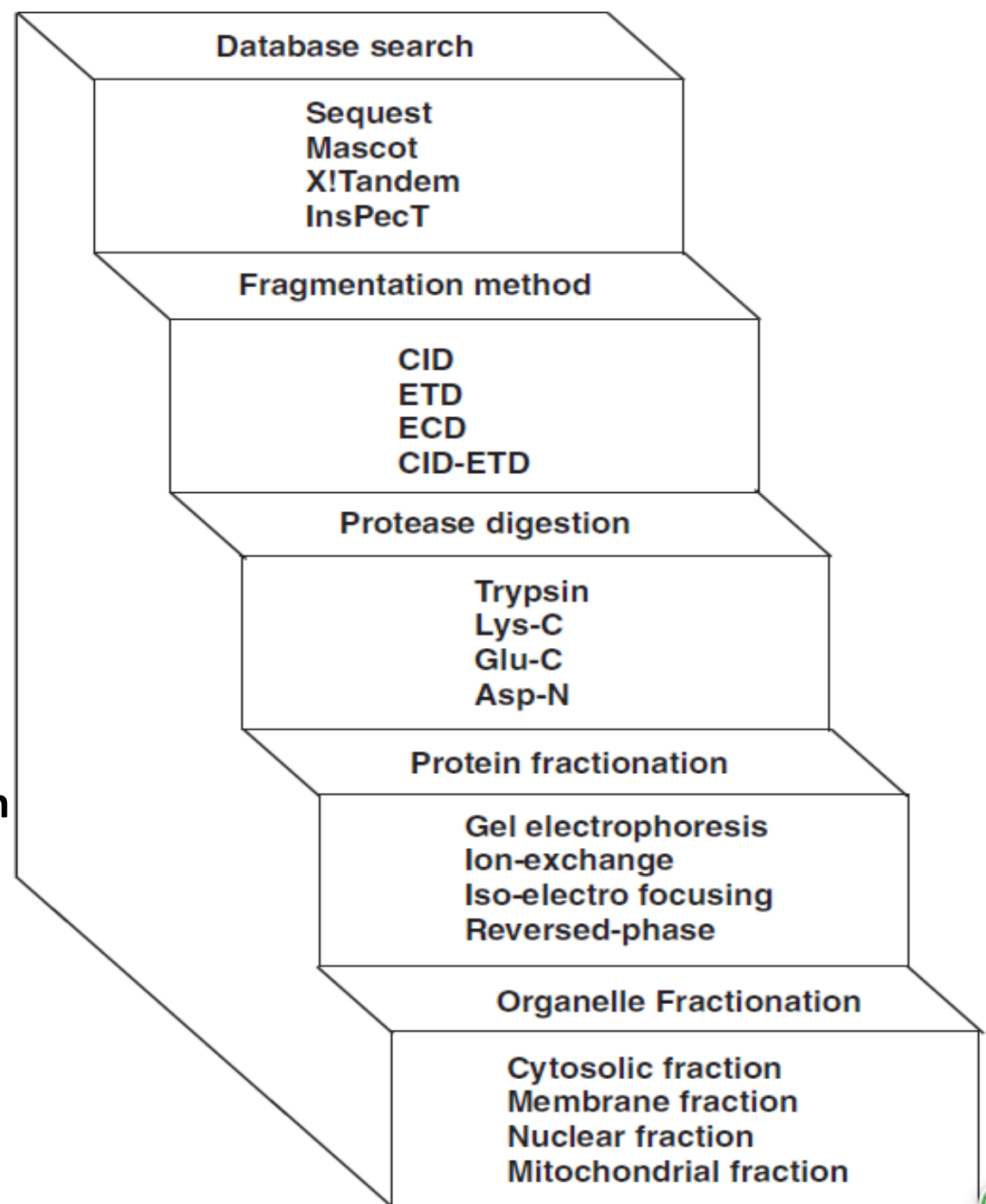




Table 2. Essentials for proteogenomics

Availability of genome sequence data

It is important to have the genome sequence data (ideally assembled) for the organism being studied. In the absence of available genome sequence information, the availability of closely related genomes is helpful

High-resolution and high accuracy mass spectrometry data

Because the search space is increased when searching genome databases, it is important to have high-resolution and high accuracy data to reduce the occurrence of false hits

Genome database search and annotation tools

Most database search algorithms do not permit direct searches against six-frame translated genome databases. Even when this is possible, direct mapping of peptides onto gene structures is not trivial



Applications of proteogenomics

- **Genome annotation**

Mass spectrometry-derived peptide data can be used to annotate genomes for the confirmation and/or correction of existing gene annotations.



Table 3. Genome annotation through proteogenomics

- **Confirmation of existing gene models**

- Peptides mapping to predicted genes confirm the existence of annotated gene models
- N-terminally acetylated peptides can be used to confirm the N-termini of proteins
- Exon–exon spanning peptides confirm splice isoforms

- **Correction of existing gene models**

- Peptides mapping to annotated intronic regions signify the presence of a novel exon or extension of an existing exon
- Peptides mapping to exon–intron boundaries extend the existing exon
- Peptides mapping to different reading frames indicates identification of additional reading frames
- Peptides overlapping a novel exon–exon junction represent novel splice isoforms

- **Identification of novel genes**

- Mapping of peptides in the non-genic region indicates a novel gene
- Peptides mapping to an annotated pseudogene indicates a wrongly annotated pseudogene

- **Others**

- Fusion protein identification from peptide mapping to two different protein-coding genes
 - Peptides with sequences that diverge from known genomic sequences reflect sequence polymorphisms in the genome
-



AT2G32235

Translated UTR



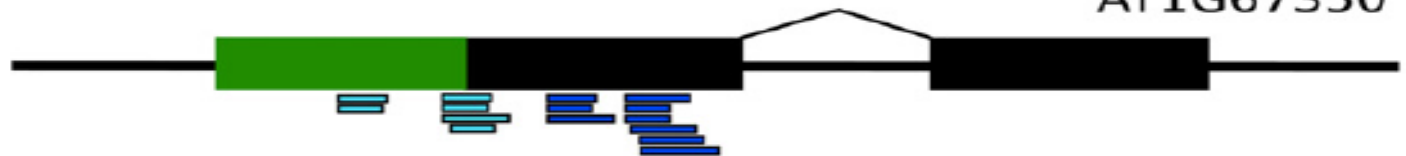
AT1G63500

Out of Frame



AT1G67350

Exon Boundary

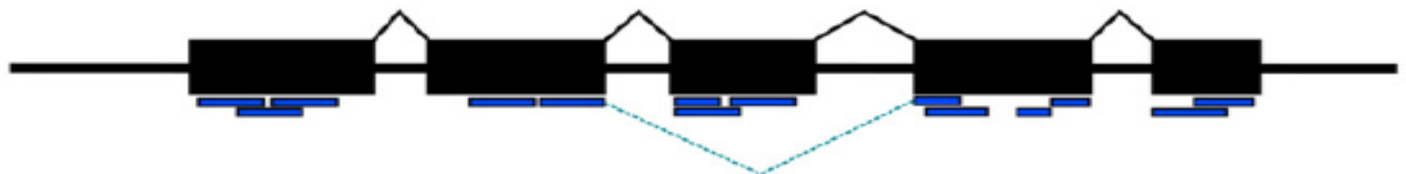


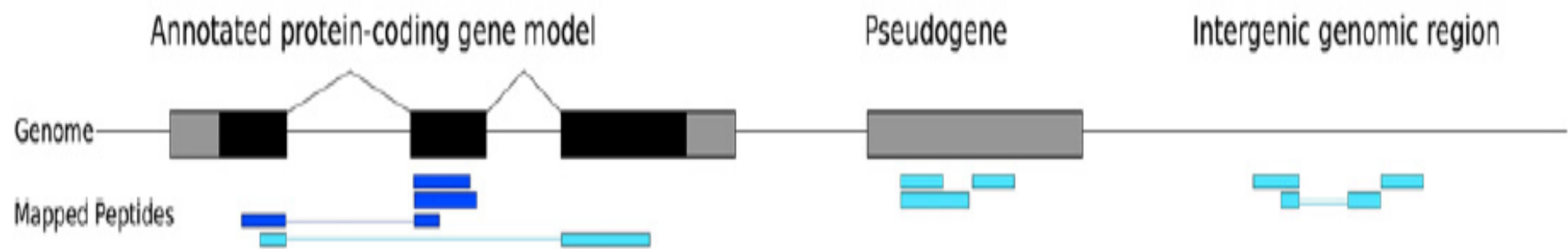
AT1G33790

Novel Exon



Novel Splice
Junction







Annotation of unsequenced genomes

- Gene conservation between closely related species permits one to study unsequenced organisms to some extent using mass spectrometry derived data. MS/MS data of an unsequenced organism are searched against proteins as well as six-frame translated genome database of a closely related species.



Studying individual genome variation

- Organisms of the same species can have almost 99.9% similarity at the level of genome sequence. Single-nucleotide polymorphisms (SNPs), insertions and deletions and copy number variations (CNVs) contribute to variations in the genomes.
- Proteogenomics can be used to identify these variations in the form of coding SNPs (cSNPs), can result in the identification of certain genome variations leading to the identification of individual-specific protein variants.



Studying disease mechanisms

- Altered protein function can result from the specific genomic alterations in a person's genome. In some of these situations, the altered protein function can result in a disease.
- For instance, some cancers are known to be caused by fusion proteins. For example, BCR-ABL gene fusion results in chronic myelogenous leukemia, and TMPRSS2-ETS gene fusion results in prostate cancer.



- Although mostly genomic approaches are used for the identification of fusion genes, proteomic approaches like mass spectrometry can also be used for the identification and characterization of fusion proteins, which may be useful in studying the altered or diseased states.
- Identification of fusion peptides using proteogenomics can be useful to unravel disease mechanisms.



Biomarker discovery

- Biomarkers can be used for disease diagnosis, making therapeutic decisions, or monitoring of disease. A proteogenomic approach is especially suitable for the identification of non-predicted or wrongly annotated protein-coding genes which would otherwise be missed.
- Many different aberrant splice isoforms are known to be associated with cancers. For instance, Her2/neu alternative splice isoforms have been shown to be associated with pancreatic and breast cancer.



Table 4. Proteogenomics for human diseases

- **Studying disease mechanisms**

Fusion genes, mutated or truncated proteins and unique splice isoforms are known to be involved in various diseases. Proteogenomics can identify such events

- **Identifying potential biomarkers**

Proteogenomics data can be used for the identification of novel fusion proteins or alternate splice isoforms, which can serve as potential biomarkers for diagnosis and/or monitoring of diseases

- **Finding genome variation**

Genomic variations such as SNPs play an important role in certain diseases. Such variations can be identified in coding regions by proteogenomics and confirmed by DNA sequencing

- **Clinical diagnostics**

Metaproteomics efforts based on proteogenomic strategies could be applied for the diagnosis of pathogenic microbes in clinical samples where other diagnosis modalities are not available



Challenges

- Although proteogenomics can be a vital tool for genome annotation, it has not been effectively combined with genome sequencing efforts.
- It is necessary to overcome some challenges.



- **Sampling:** In multicellular organisms, one has to analyze multiple organs and tissues to obtain good proteome coverage, which requires significantly more work than genome sequencing
- **Genome sequence unavailability:** However, because of the ease of sequencing genomes, this should not be a major problem in the future because the genomes of all species are likely to be available or can be obtained prior to proteomic analysis.



- **Data analysis bottleneck:** Peptide identification of known or predicted proteins is straightforward. Proteogenomic identification of novel peptides includes six-frame translated database searching and such large genome database sizes results in increased search space. This can increase the error rate significantly contributing to higher false positives. That is why it is important to carry out such searches with the highest quality of data.



Software

- It is necessary to have a software application specifically designed for proteogenomic analysis – a software which can be used for protein database search as well as six-frame translation of genome sequence search of MS/MS data. As of now, there is no single software application in use where one can perform an entire proteogenomic analysis.



Future prospects

- Though the human genome sequence became available 10 years ago, there is still ambiguity in the exact number of protein-coding genes. Multi-pronged approaches such as transcriptomics and proteomics in addition to genomics are being carried out only in limited instances.

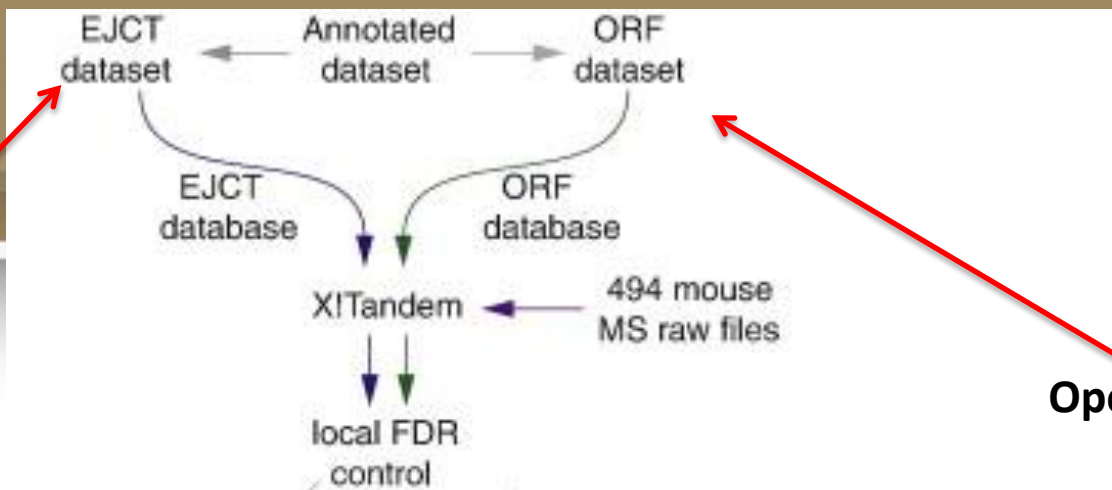
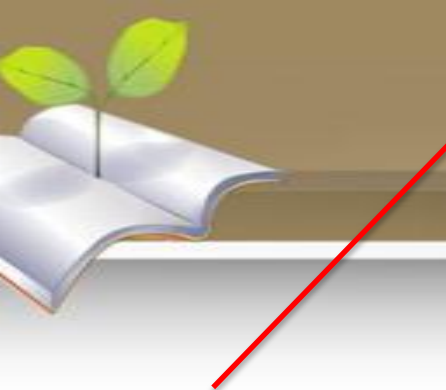


- It may be suggested that every genome sequencing project should include proteogenomic analysis from the initiation of the project to provide a more accurate catalog of protein-coding genes.
- Apart from this genome annotation, proteogenomics data will be increasingly used for unraveling disease mechanisms, biomarker discovery and studying genome-wide variation.



Proteogenomics: Our related works

- Xiaobin Xing, Qingrun Li, Han Sun, Xing Fu, Fei Zhan, Xiu Huang, Jing Li, Chunlei Chen, Yu Shyr, Rong Zeng, Yixue Li, **Lu Xie**. The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. **Genomics**. 2011, 98(5): 343-351.
- ZHANG Kun, WANG Le-Heng, CHI Hao, BU De-Chao, YUAN Zuo-Fei, LIU Chao, FAN Sheng-Bo, CHEN Hai-Feng, ZENG Wen-Feng, LUO Hai-Tao, SUN Rui-Xiang, HE Si-Min, **XIE Lu**, ZHAO Yi. Proteogenomics: Improving Genomes Annotation by Proteomics. **Progress in Biochemistry and Biophysics**. 2013, 40(4): 297~308
- Han Sun, Xiaobin Xing, Jing Li, Yunqin Chen, Ying He, Wei Li, Guangwu Wei, Xiao Chang, Jia Jia, Jing Li, Yixue Li, **Lu Xie**. Identification of Gene Fusions from Human Lung Cancer Mass Spectrometry Data. **BMC Genomics**. 2013, accepted.



Open Reading Frame

Alternative Splicing

The discovery of novel protein-coding features in mouse genome based on mass spectrometry data.

Xiaobin Xing, et al.,
Genomics, 2011.

88 novel coding region discovered, and 12 novel genes obtained by AUGUSTUS prediction

52 novel splicing events (involving 47 genes) inferred

Gene model refine event	No.
Intron imbeded	19
3'/5' UTR overlapped	6
Exon included	2
Processed transcript	3
Processed pseudogene	2

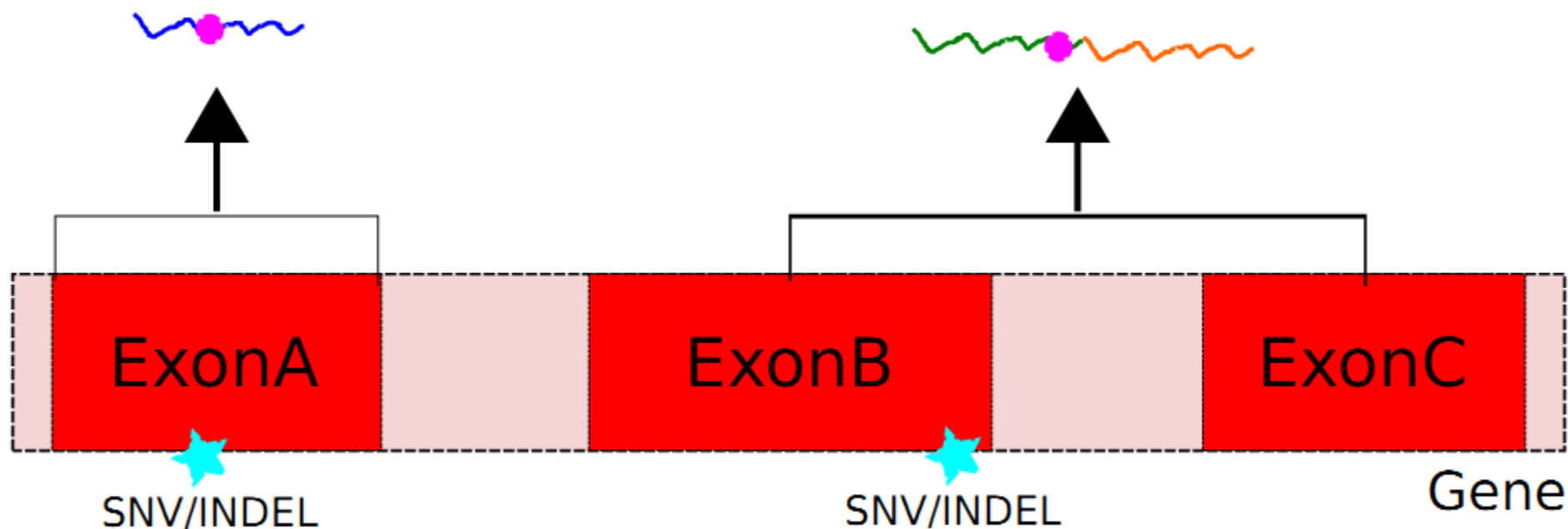


Proteogenomics: Our related works

- Cancer Variant Discovery
 - SNV/INDEL
 - Alternative Splicing
 - Gene Fusion



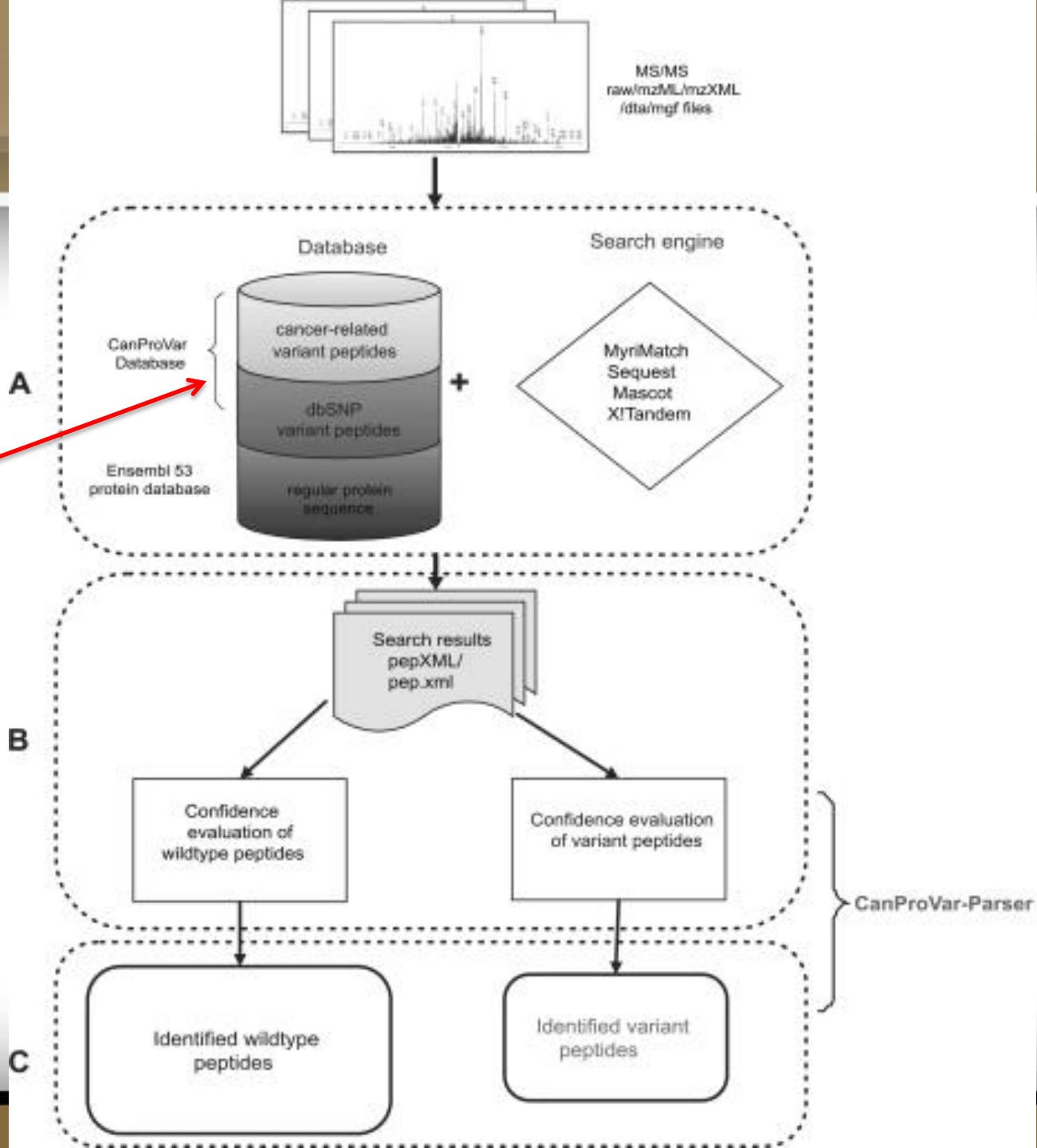
Cancer Variant Discovery



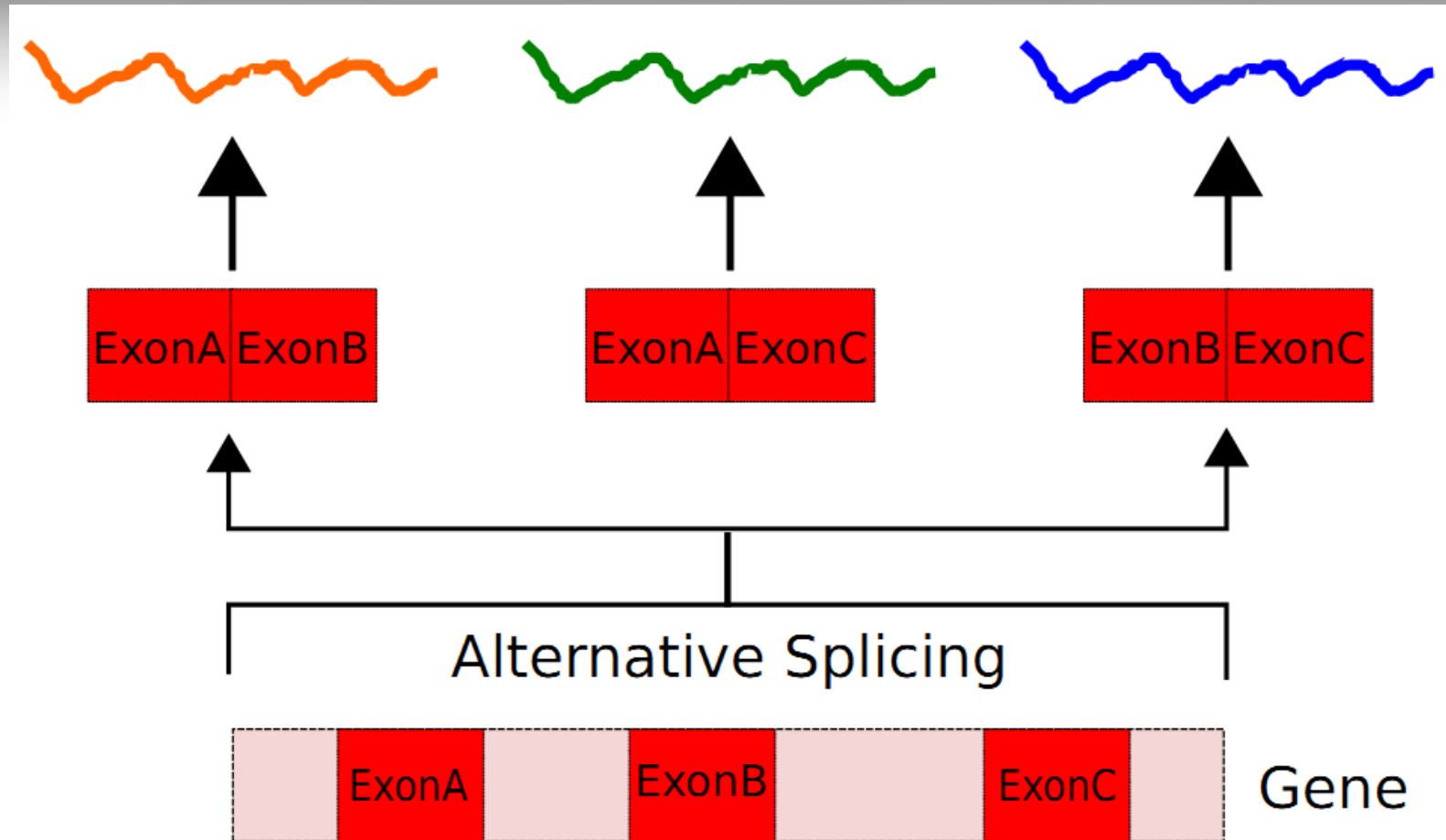
A bioinformatics workflow for variant peptide detection in shotgun proteomics,
Jing Li, et al., Mol Cell Proteomics, 2011



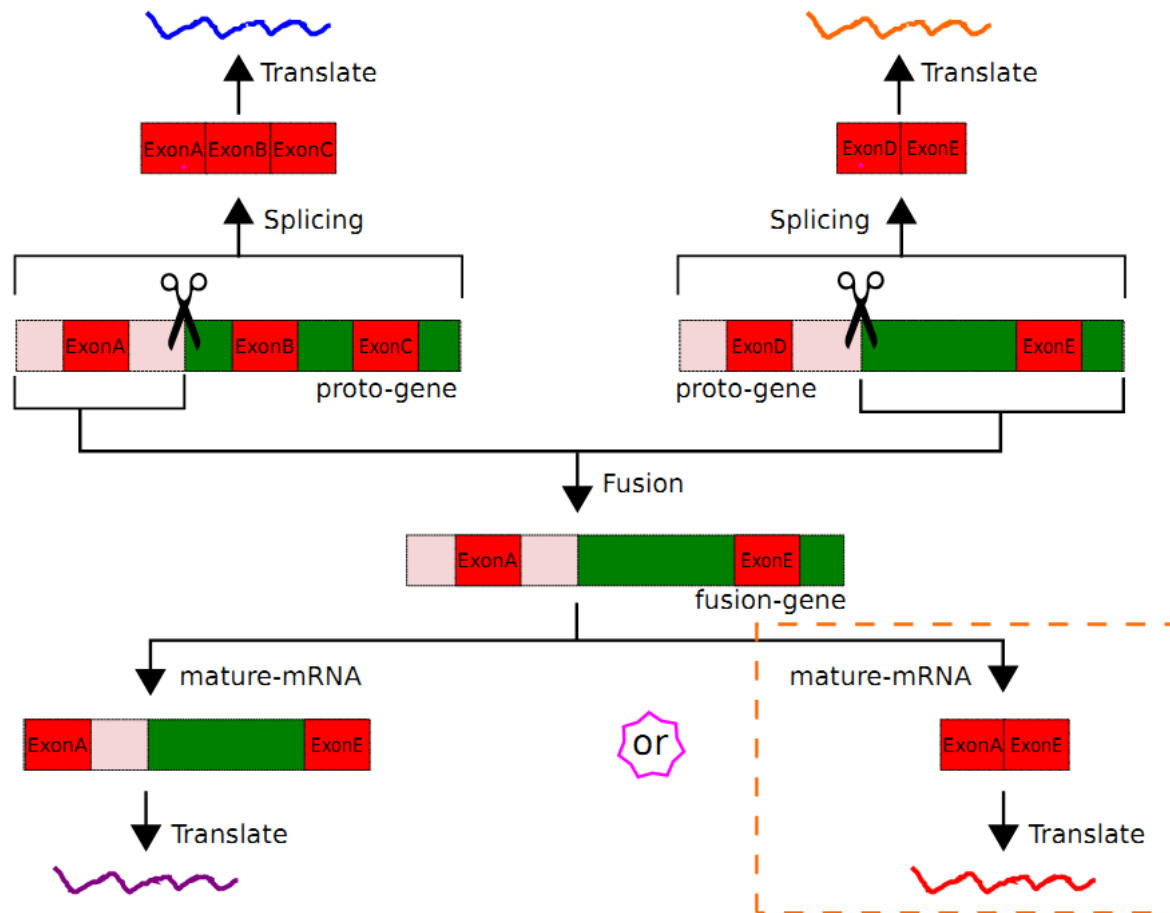
SNV



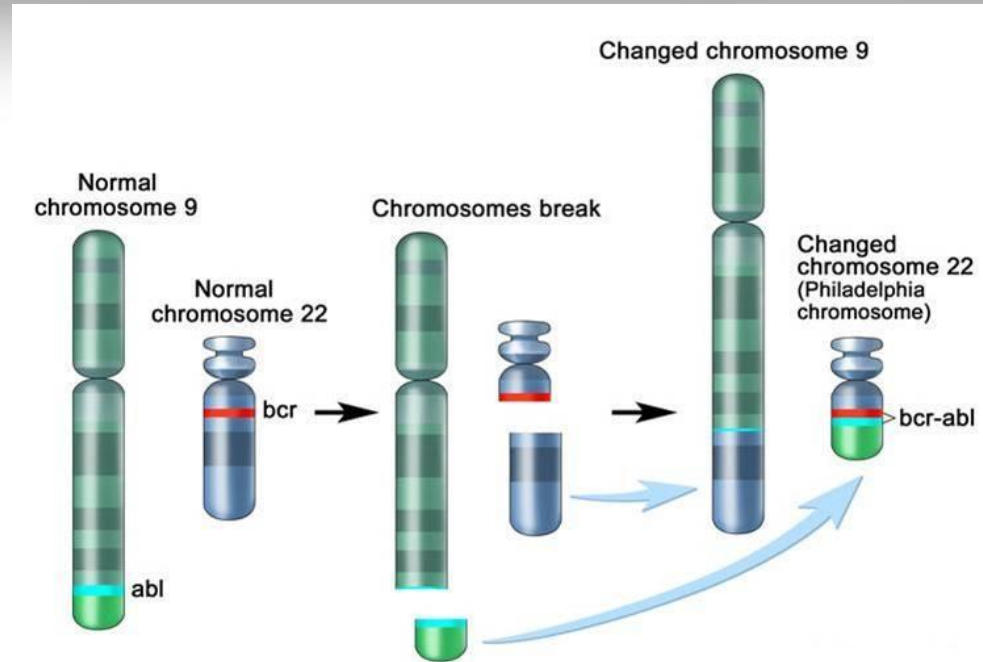
Cancer Variant Discovery



Cancer Variant Discovery



Gene Fusion Example



- 95% of people with CML have Philadelphia chromosome
- ABL and BCR are normal genes on chromosomes 9 and 22.
- ABL gene encodes a tyrosine kinase whose activity is tightly regulated
- BCR-ABL also encodes a tyrosine kinase, adding a phosphate group to tyrosine, whose activity is deregulated
- Understanding this process led to the development of the drug imatinib mesylate (Gleevec)

TQRHYLGHTDCVKCLAIHPDKIRIATGQIAGVDKDGRLQPHVRVWDSVTLSTLQIIIGLG
 TFERGVGCLDFSKADSGVHLCVIDDSNEHMLTVWDWQKKAKGAEIKTTNEVVLAVEFHPT
 DANTIITCGKSHIFFWTWSGNSLTRKQGIFGKYKPKFVQCLAFGLNGDVLTDGSSGGVML
IWSKTTVEPTPGKGPKVYRRKHQELQAMQMELOQSPEYKLSKLRTSTIMTDYNPNYCFAGK
 TSSISDLKEVPRKNITLIRGLGHGAFGEVYEGQVSGMPNDPSPLQVAVKTLPEVCSEQDE
 LDFLMEALIISKFNHQNIVRCIGVSLQSLPRFILLELMAGGDLKSFLRETRPRPSQPSSL
 AMLDLLHVARDIACGCQYLEENHFIHRDIAARNCLLTCPGPGRVAKIGDFGMARDIYRAS

```
>genefusions:ALK:EML4:ENSE00001735043:ENSE00000962700:53:136|2|17|0|45|
LHQ*PEGGAEEKHHPHSHYKPKFVQCLAFGLNGDVLTDGSSGGVMLIWSKTTVEPTPGKGPK
>genefusions:EML4:ALK:ENSE00000962700:ENSE00001154407:136:187|0|46|-2|61|
KYKPKFVQCLAFGLNGDVLTDGSSGGVMLIWSKTTVEPTPGKGPKVYRRKHQELQAMQMELOQSPEYKLSKLRTSTIMTDYNPNYCFAGKT
>genefusions:EML4:ALK:ENSE00000962700:ENSE00001154407:136:187|1|45|0|62|
NMKSQNLCSV*HSWGMEMFLLETQVESCLYGAKLL*SPHLGKDLKCTAGSTRSCKPCRWSCRALSTS*ASSAPRPS*PTTTPPTALLARP
```

EML4:ALK, Soda et al, Nature, 2007

MEDSMDMDMS	PLRPQNYLFG	CE	ADKDYH	FK	VQNDENEH	QLSLRTVSLG	AGAK	DELHIV	EAEAMNYEGS	PIKVTLATLK
MEDSMDMDMS	PLRPQNYLFG	CE	ADKDYH	FK	VQNDENEH	QLSLRTVSLG	AGAK	DELHIV	EAEAMNYE	VTIATLK
					QNDENEH	QLSLRTVSLG	AGAK			VTIATLK
					NEH	QLSLRTVSLG	AGAKDEL			GS PIKVTLATLK

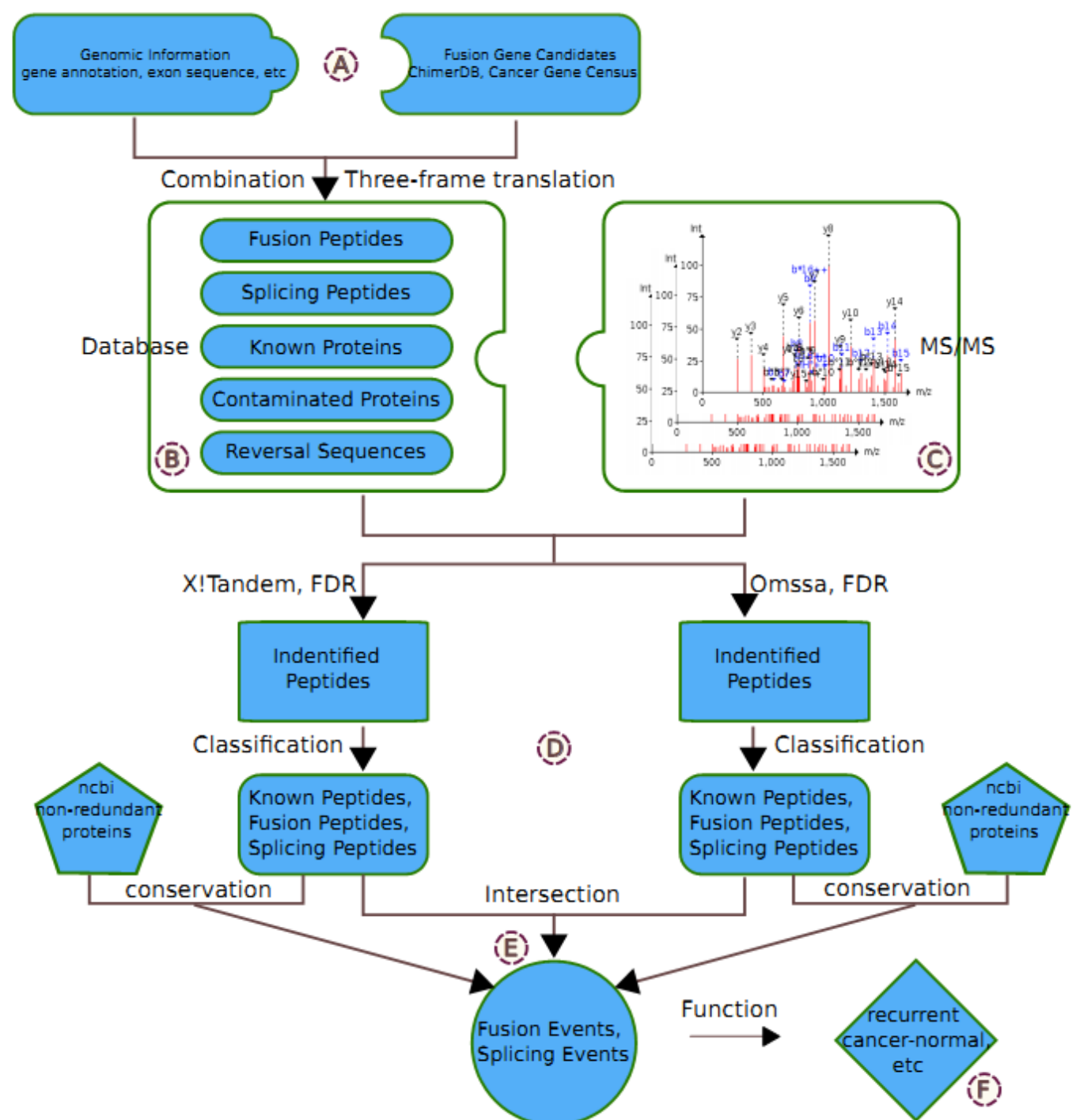
fusion site


MSVQPTVSLG	GFEITPPVVL	RLKCGSGPVH	ISGQHLVVYR	RKHQELQAMQ	MELQSPYKLSKLRTSTIMT	DYNPNYCFAG
MSVQPTVSLG	GFE	LKCGSGPVH	ISGQHLVVYR	HQELQAMQ	MELQSPYK	LRTSTIMT DYNPNYCFAG
		CGSGPVH	ISGQHLVVYR	RK	LQAMQ	MELQSPY
		PVH	ISGQHLVVYR	RKHQELQ		

KTSSISDLKE	VPRKNITLIR	GLGHGAFGEV	YEGQVSGMPN	DPSPLQVAVK	TLPEVCSEQD	ELDFLMEALI	ISKFNHQNIV
KTSSISDLKE	VPR				TLPEVCSEQD	ELDFLMEALI	ISKFNHQNIV
	VPRKNITLIR	GLGHGAFGE			VCSEQD	ELDFLME	NIV

```
>genefusions:ALK:NPM1:ENSE00001735043:ENSE00001084440:53:94|2|17|0|31|
LHQ*PEGGAEEKHHPHSHVSLGGFEITPPVVLRLKCGSGPVHISGQHLV
>genefusions:NPM1:ALK:ENSE00001084440:ENSE00001154407:94:187|0|32|-2|61|
VSLGGFEITPPVVLRLKCGSGPVHISGQHLVYRRKHQELQAMQMELOQSPEYKLSKLRTSTIMTDYNPNYCFAGKTSSISDLKEVPRKNITLI
>genefusions:NPM1:ALK:ENSE00001084440:ENSE00001154407:94:187|1|31|0|62|
FPLGALK*HHQWS*G*SVVQGCILVDST**CTAGSTRSCKPCRWSCRALSTS*ASSAPRPS*PTTTPPTALLARPPPSVT*RRCRGKTSPSF
```

NPM1:ALK, Elenitoba-Johnson et al, PNAS, 2006



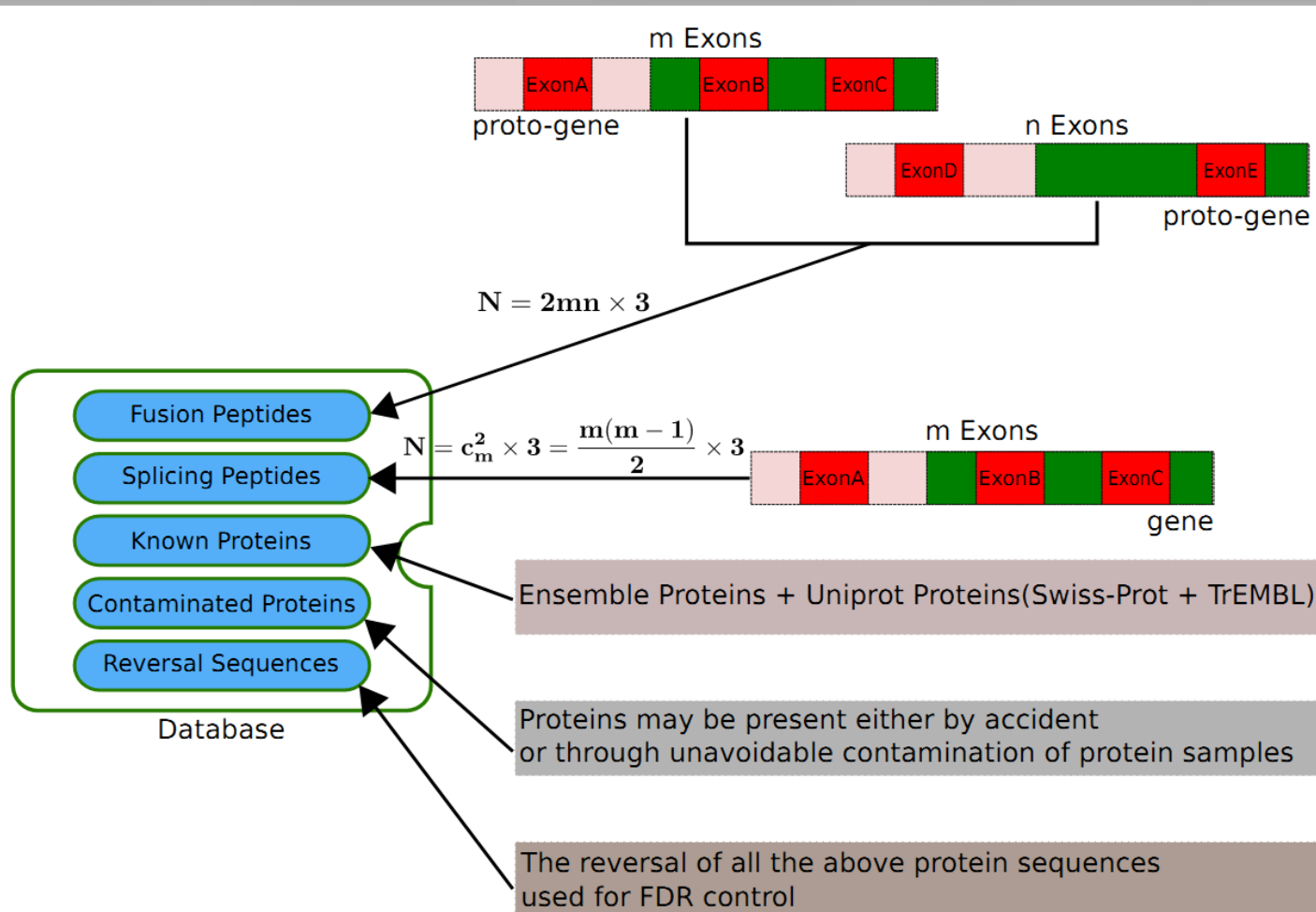
Han Sun, et al.
Identification of Gene
Fusions from Human
Lung Cancer Mass
Spectrometry Data.
BMC Genomics, 2013,
accepted.



A: 6259 potential gene pairs

- Cancer Gene Census
- ChimerDB 2.0
- Literature-based Annotation
 - Integrated literature knowledgebase
 - PubMed annotation
 - OMIM annotation
 - Sanger CGP
 - Mitelman Database
- Transcriptome Analysis
 - mRNA seq
 - EST seq
 - RNA-seq

B: CanProFu – the fusion peptide database





C: Lung Cancer MS/MS Data

- adenocarcinomas(**ADC**)
 - 1,199,542 spectrums
- squamous cell carcinomas(**SCC**)
 - 1,272,006 spectrums
- normal tissues from human lung (**Normal**)
 - 2,531,416 spectrums
- Each sample pools from 19 or 20 original tissues
- Each sample was analyzed in 4 IEF/RPLC technical replicates
- All the fractions were analyzed by LC-MS/MS on a Thermo-Fisher LTQ-Orbitrap

In-depth proteomic analysis of non-small cell lung cancer to discover molecular targets and candidate biomarkers,

Takefumi Kikuchi et al., Mol Cell Proteomics, 2012



D: Search Database using X!Tandem and Omssa

- 320 raw files
- ± 10 ppm parent monoisotopic mass error
- ± 0.4 da fragment monoisotopic mass error
- Sorted the matched spectrum by hyper score
- $$\text{FDR} = \frac{2R}{F+R} = 10^{-6}$$

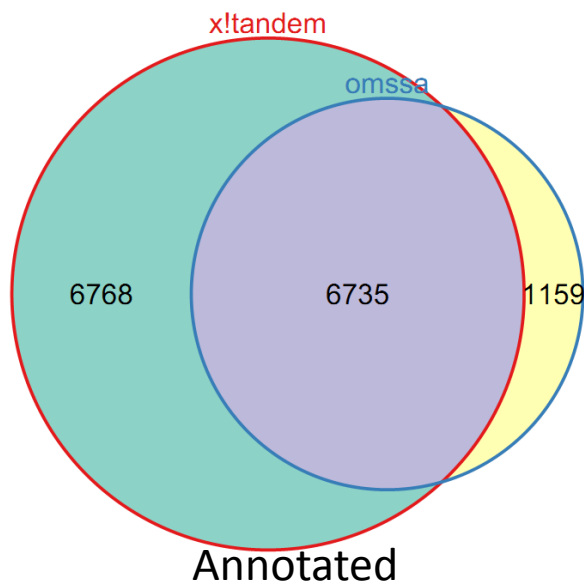
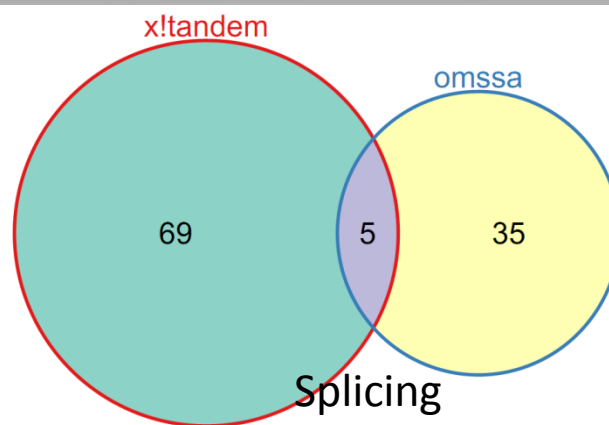
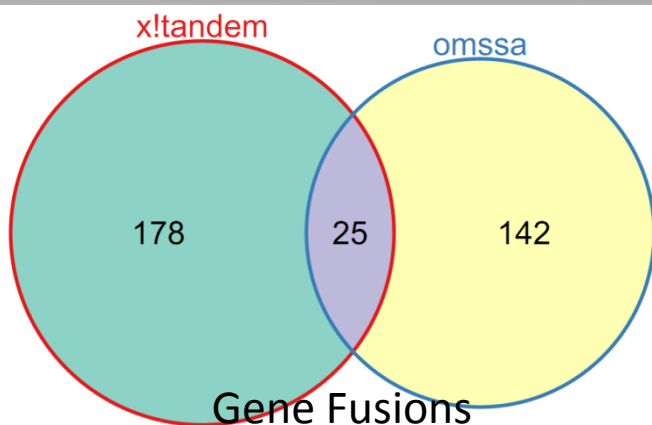


Determination of Fusion, Splicing peptides

- At least 3 aa each beside the fusion/splicing point
- At least 2 spectrums matched to the peptide
- The fusion and splicing peptides are not in known human proteins when blasted against NCBI non-redundant database

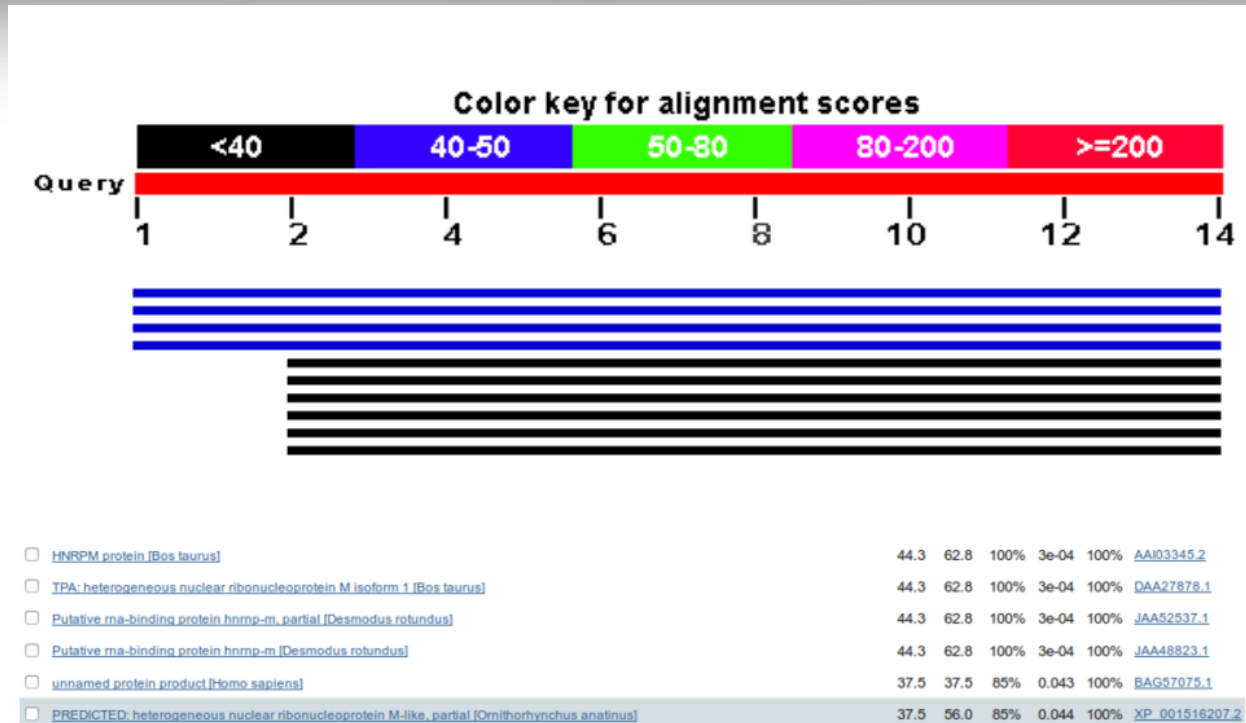


Classification of Fusion, Splicing or Annotated peptides





Save peptides through conservation



Peptide: **INGGGGSVPGIER**

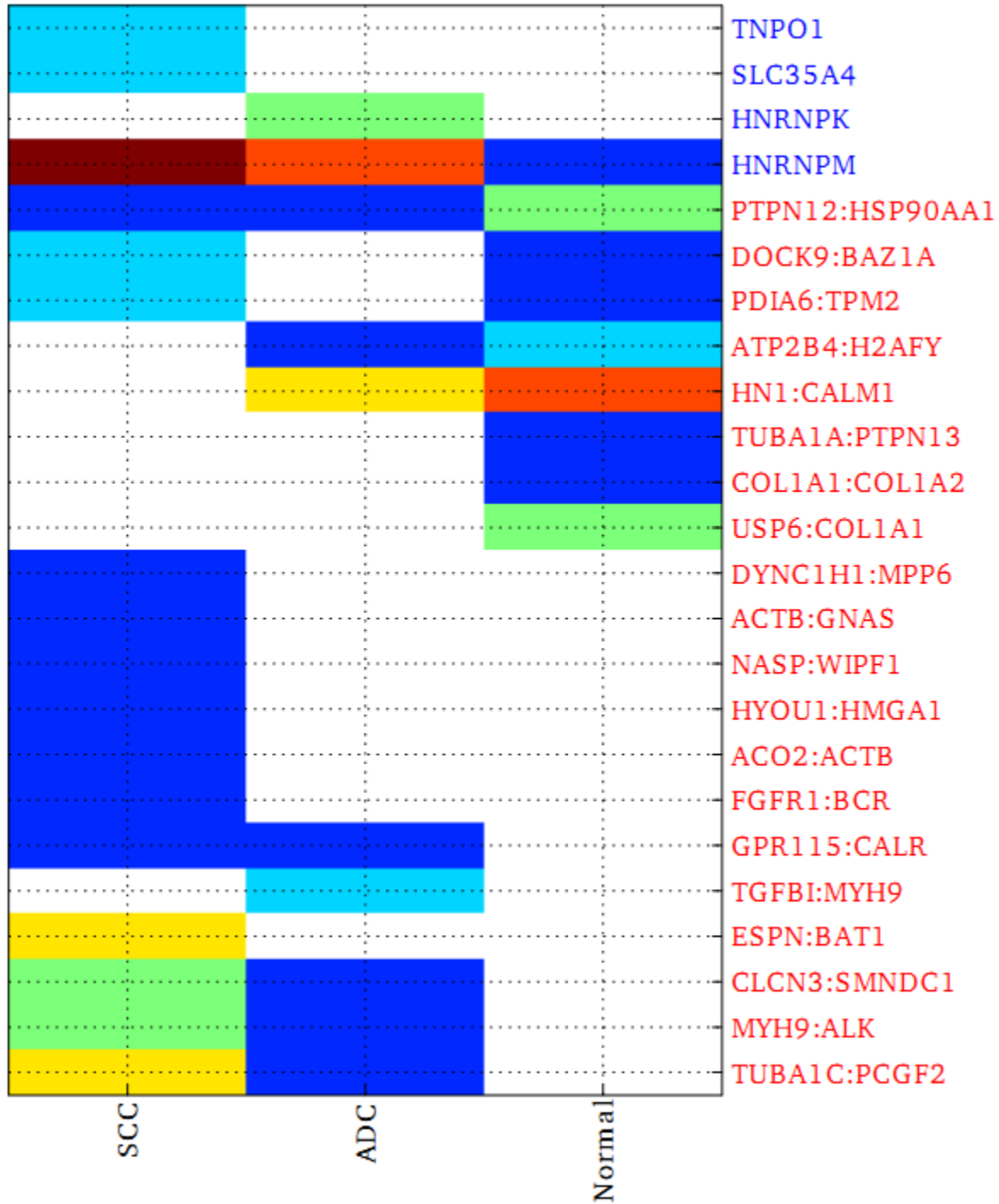
Conservation: Bos taurus(100%), Desmodus rotundus(100%), Homo sapiens(85%)

If a fusion or splicing peptide could be found in other species, even if they were only identified by one of the search engines, they were included as candidate identified peptides



27 Fusion Peptides and 7 Splicing Peptides

- 27 Fusion Peptides
 - 7 peptides were identified by both search engines, but the original spectra were completely different
 - 20 peptides
 - 12 peptides were fully digested by trypsin and with no mis-cleavage
 - 6 peptides were fully digested but with one mis-cleavage
 - 2 peptides were semi-digested
- 7 Splicing Peptides
 - 3 peptides were identified by both search engines, but the original spectra were completely different
 - 4 peptides
 - 3 peptides were fully digested by trypsin and with no mis-cleavage
 - 1 peptide was fully digested but with one mis-cleavage





GO process enrichment

- 41 unique genes (20 fusion peptides + 4 splicing peptides)

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0043933	macromolecular complex subunit organization	2.22E-8	2.57E-4	5.52 (17723,1204,40,15)
GO:0071822	protein complex subunit organization	7E-8	4.06E-4	6.17 (17723,933,40,13)
GO:0065003	macromolecular complex assembly	6E-7	2.32E-3	5.70 (17723,933,40,12)
GO:0022607	cellular component assembly	9.01E-7	2.61E-3	4.51 (17723,1374,40,14)
GO:0006461	protein complex assembly	2.38E-6	5.52E-3	6.37 (17723,696,40,10)
GO:0006457	protein folding	4.1E-6	7.92E-3	13.92 (17723,191,40,6)
GO:0016043	cellular component organization	1.69E-5	2.8E-2	2.54 (17723,3493,40,20)
GO:0071840	cellular component organization or biogenesis	1.93E-5	2.8E-2	2.51 (17723,3524,40,20)
GO:0006928	cellular component movement	2.16E-5	2.78E-2	4.96 (17723,894,40,10)



MYH9:ALK

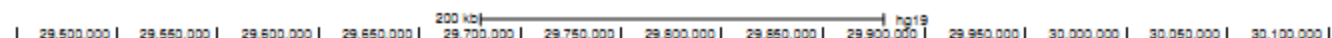
- Only found in ADC and SCC, not in Normal
- Non-muscle myosin heavy chain (**MYH9**): a new partner fused to **ALK** in anaplastic large cell lymphoma, Lamant L et al, Genes Chromosomes Cancer, 2003
- Identification of the transforming **EML4:ALK** fusion gene in non-small-cell lung cancer, Soda et al, Nature, 2007
 - The EML4-ALK fusion gene is responsible for approximately 3-5% of non-small-cell lung cancer(NSCLC)

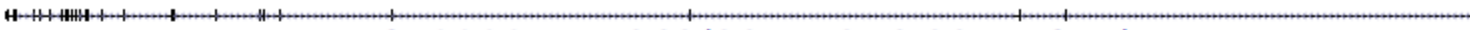



MYH9:ALK



- MYH9 normally locates on the complement strand of chr22, and encodes a conventional non-muscle myosin which was reported to be involved in several important functions, such as cytokinesis, cell motility and maintenance of cell shape
- Defects in this gene have been associated with non-syndromic sensorineural deafness autosomal dominant type 17, Epstein syndrome and so on
- ALK normally locates on the complement strand of chr2, and encodes a receptor tyrosine kinase
- Many translocations have been found with this ALK gene, including EML4:ALK, RANBP2 :ALK, PM1:ALK, TFG:ALK and so on

chr2 (p23.2-p23.1) 

 hg19

ALK 
ENSE00001745962(29551216-29551357)

MYH9 
ENSE00001600096(36716265-36716408)

 hg19
chr22 (q12.3) 

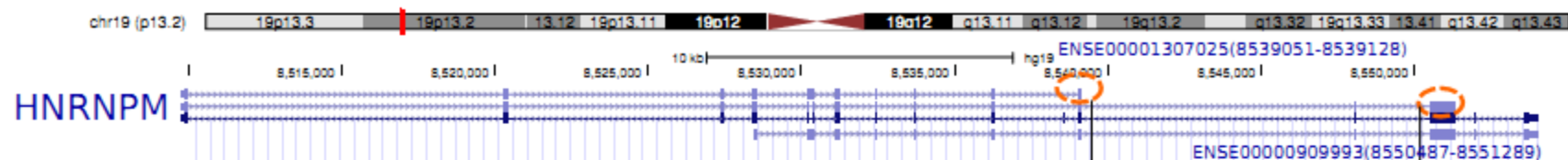
...CCGATCTCCTGTTGGAGCCGTACAACAAATACCGCTTCCTGTCCAATGGA
CACGTCACCATCCCCGGGCAGCAGGACAAGGACATGTTCCAGGAGACCA
TGGAGGCCATGAGGATTATGGGCATCCCAGAAGAGGAGCAAATGGTGCT
CTCCAGGAACATCCCCAGGCTCCAAGATGGCCCTGCAGAGCTCCTTCACT
TGTTGGAATGGGACAGTCCTCCAGCTTGGGCAGGCCTGTGACTTCCACCA
GGACTGTGCCCAGGGAGAAGATGAGAGCCAGATGTGCC...

DLLLEPYNKYRFLSNGHVTIPGQQDKDMFQ
ETMEAMRIMGIPEEQMVLSRNIPRLQDGP
AELLHLEWDSPPAWAGL



HNRNPM and HNRNPK

- Alternative splicing of HNRNPM was found all the three types of samples
- but the splicing of HNRNPK was only found in ADC
- **Multiple and Specific mRNA Processing Targets for the Major Human hnRNP, Julian et al, Mol Cell Biol, 2008**



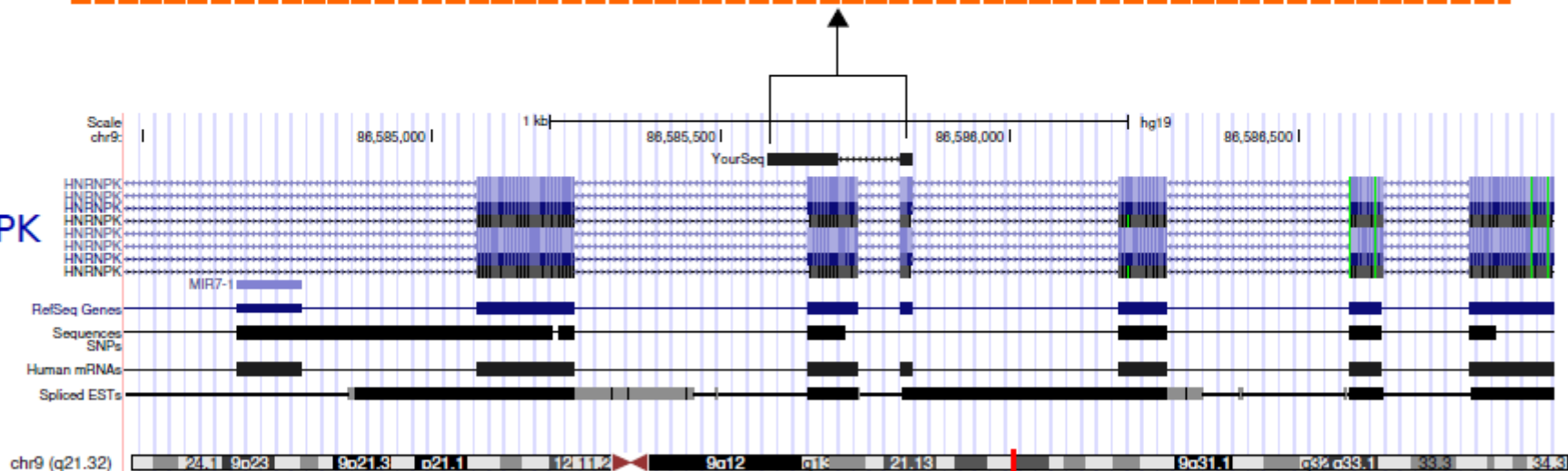
...GAATGGAGGGGCCCTTTGGTGGTGGTATGGAAAACATGGGTCGATTTGGAT
 CTGGGATGAACATGGGCAGGATAAATGGTGGAGGTGGAGGAAGCGTCCCTGG
 GATCGAGAGGATGGGTCCTGGCATTGACCGCCTCGGGGGTGCCGGCATGGAG
 CGCATGGGCGCGGGCCTGGGCCACGGCATGGATCGCGTGGGCTCCGAGATCG
 AGCGCATGGGCCTGGTCATGGACCGCATGGGCTCCGTGGAGCGCATGGGCTCC
 GGCATTGAGCGCATGGGCCCGCTGGGCCTCGACCACATGGCCTCCAGCATTGA
 GCGCATGGGCCAGACCATGGAGCGCATTGGCTCTGGCGTGGAGCGCATGGGT...

MEGPFGGGMENMGRFGSGMNMGRINGGGGGSVPGIERMGPIDR
 LGGAGMERMGAGLGHGMDRVGSEIERMGLVMDRMGSVERMGSGI
 ERMGPLGLDHMASSIERMGQTMERIGSGVERMGAGMGFGLERMAA
 PIDRVGQTIERMGSGVERMGPAIERMGLSMERMVPAGMGAGLERMG
 PVMDRMATGLERMGANNLERMGLERMGANSLERMGLERMGANS
 ERMGPAMGPALGAGIERMGLAMGGGGGASFDRAIEMERGNFGGSF
 AGSFGGAGGHAPGVARKACQIFVRN

GGSGYGDLGGPIITTQVTIPKDVSIFNTTR

GGTGGCTCCGGATATG GTGATCTTGGTGGACCTATTATTACTACACAAGTAACTA
TTCCCAAAGATGTAAGTATCTTTAATACTACCAGGAACATTTTATCACTTTTATGATT
ATTCCTCCTTTCTGTATATTTTTTAATTCAGAAGGTTTTAACAAAAATACACATTAG
GATTGAGGTTATGTTAATGGGCTTTAGTGAGCTGGGTTTTTCAGCTGTTTGAGTCTT
GTCAAGTGATCAGTGCTATTAATAAAAGTAGTTAAGTAGGTTTTGAGCCCTTAACTA
ACTAAGGGAAACATTAGTAAGTGTGACATAAATACATTATAACTCAAACCTTGACA
GGTTAGGGAGCGTTAGATCATCAGTTAAGATTCTGAATGAATAAAATTAATAAC...

HNRNPK





Conclusion

- For the purpose of identifying cancer fusion events, we constructed a cancer fusion peptide sequence database---CanProFu
- Applying mass spectrometry data from 40 non-small cell lung cancer(NSCLC) samples and 39 normal lung tissue controls to search in CanProFu, 20 fusion peptides and 4 splicing peptides were identified
- MYH9:ALK fusion peptide was newly found and only existed in NSCLC
- The CanProFu database and workflow in this work can be flexibly applied to other MS/MS based human cancer experiments to detect gene fusions as potential disease biomarkers and help improve understanding of the related cancer mechanism



Acknowledgments

SUN Han, Ph.D candidate

LI Jing, Ph.D candidate

LI Yixue, Ph.D, Professor

Funding: National Key Basic Research Program
2010CB912702