A new approach to identifying protein isoforms by using the cell line-specific protein sequence database

Yonsei Proteome Research Center

Seul-Ki Jeong, Chae-Yeon Kim and Young-Ki Paik
Yonsei Proteome Research Center

Chromosome 13

19th C-HPP Symposium, Santiago de Compostela, Spain

Introduction

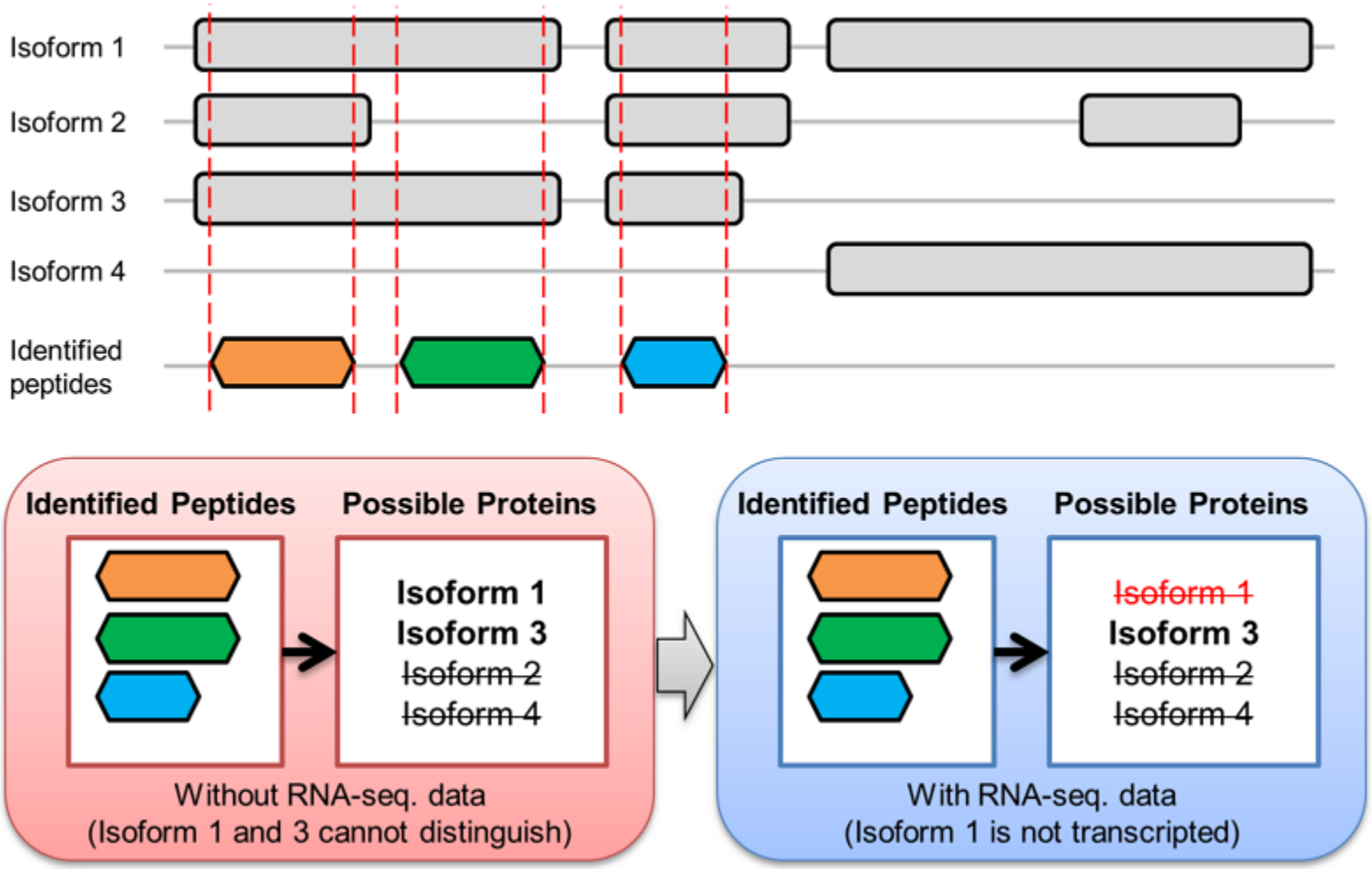
- **One of the goals of C-HPP is to map protein isoforms produced by alternative splicing.**
- Alternative splicing increases transcriptomic and protein diversity dramatically and contribute to a number of biological processes and diseases.
- Alternative splicing play important role in biological processes.
- **Alternative splicing variants (ASVs) are difficult to detect by proteomics due to their highly homologous nature.**
- Many studies aim to detect novel peptides of ASVs not protein itself.

Purposes

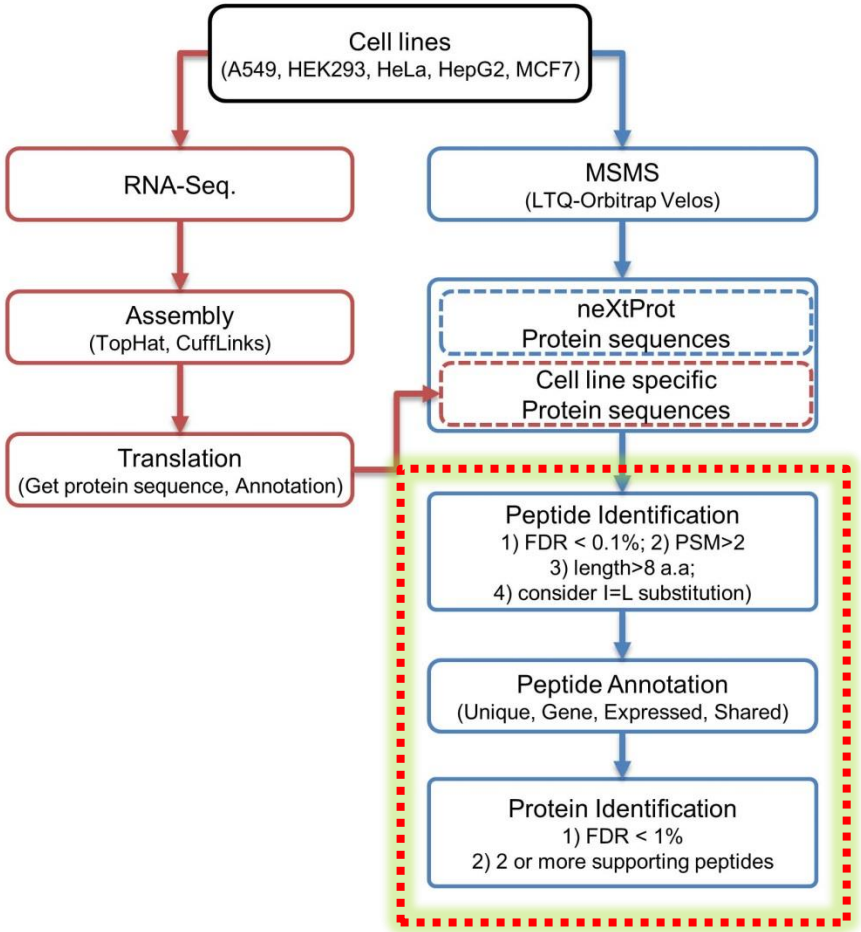
- Major purposes are to
  - Identify protein isoforms.
  - Identify cell-line specific isoforms and/or differently distributed isoforms present in cell lines.

- Rationale
  - Most protein-coding genes have a single dominant isoform depending on tissues or cell types (Ezkurdia I et al., *J Proteome Res*. **2015**, 14(4):1880-7)
  - With support of RNA-seq data, we can find dominant isoforms of target genes on target sample.

Scheme


Isoform 1
Isoform 2
Isoform 3
Isoform 4
Identified peptides
Identified Peptides
Possible Proteins
Isoform 1
Isoform 3
~~Isoform 2~~
~~Isoform 4~~
Without RNA-seq. data
(Isoform 1 and 3 cannot distinguish)
Identified Peptides
Possible Proteins
~~Isoform 1~~
Isoform 3
~~Isoform 2~~
~~Isoform 4~~
With RNA-seq. data
(Isoform 1 is not transcripted)

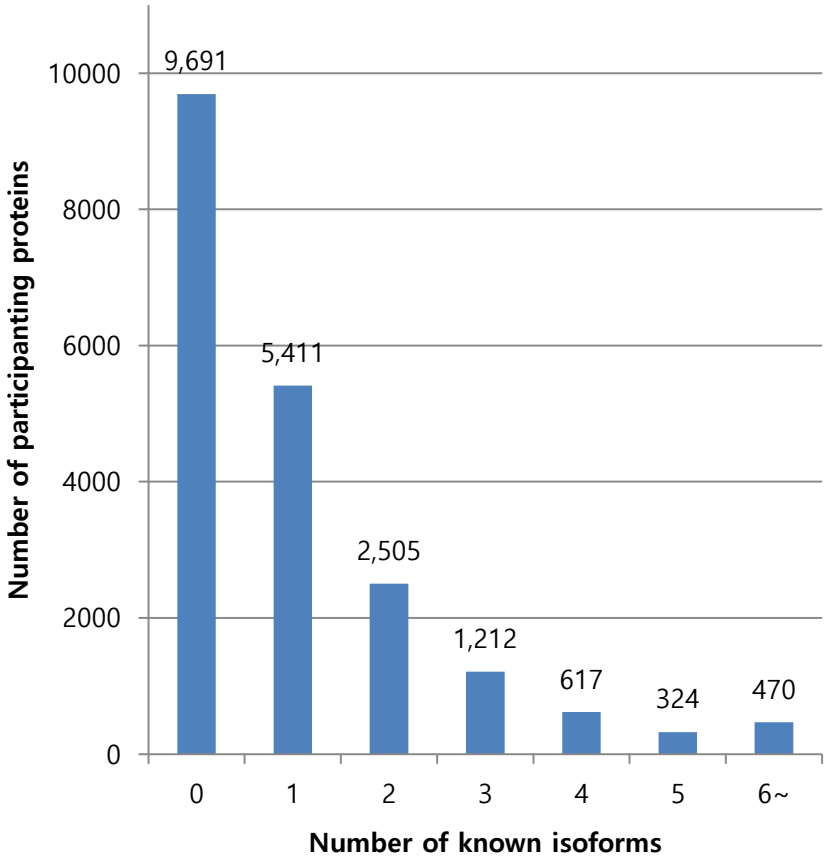
Workflow



- Proteomics and Genomics dataset of Five cell lines
- Peptide Identification
  - FDR less than 0.1% at peptide level
  - 3 or more PSMs
  - Peptide length is more than or equal nine amino acids.
  - Consider I=L substitution
- Peptide Annotation
  - Define four type of peptide groups
  - **U**: Uniquely-mapping peptide, only one protein/ isoform have this peptide
  - **G**: share in one protein group but not with other protein groups
  - **S**: Shared peptide, several proteins and protein groups have this peptide
  - **E**: shared only in one protein group and that protein group have only one expressed isoforms.
- Protein Identification
  - FDR less than 1% at protein level
  - Have at least two supporting peptides.
  - Supporting peptides are must in group U or E.

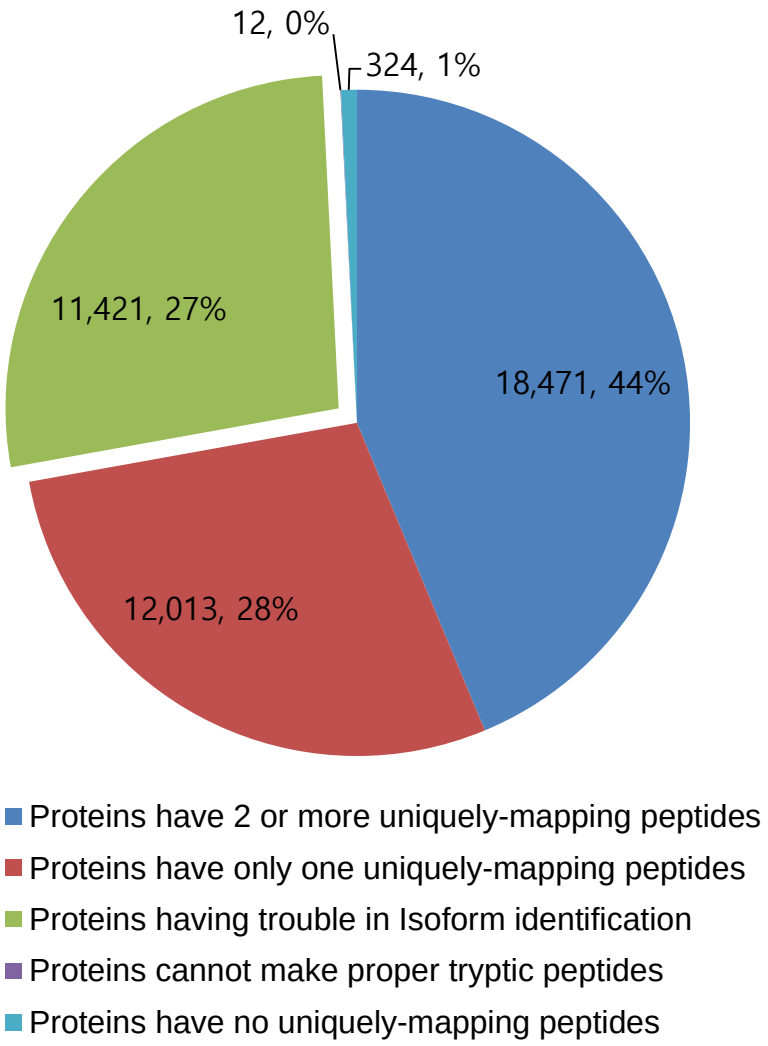
Number of Known Isoforms

- At present (neXtProt 2018-01-17 release)
  - There are 20,230 protein-coding genes with 10,539 protein-coding genes have at least one or more isoforms.
  - While 9,691 protein-coding genes have no known isoforms.
  - Together, there are at least 42,241 proteins that are encoded by human genes



Confidently Identifiable Isoforms

- Among those 42,241 proteins,
  - 44% of proteins are confidently identifiable (2 or more uniquely-mapping peptide with ≥9 a.a. in length)
  - 27% (45% of un-identifiable proteins) of proteins having trouble because of its highly homologous peptides composition with their isoforms.



Number of Expressed Proteins in Cell Lines

- Cell line specific transcript expression information.
- About 25 % of human proteins are expressed (FPKM>1.0) in selected target cell lines
- As expected, 91% of proteins that are included in protein-coding genes having multiple isoforms show only one isoform expressed
  - They could be useful for credible identification by transcript expression data

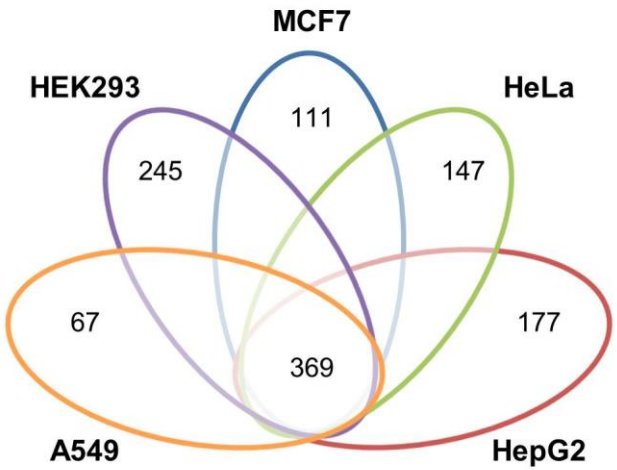
	neXtProt	A549	HEK293	HeLa	HepG2	MCF7
Human proteins (%)[a]	20,230	5,263 (26)	5,822 (29)	4,349 (22)	4,931 (24)	4,905 (24)
Protein-coding genes w/o Isoforms (%)[b]	9,691	2,187 (23)	2,446 (25)	1,875 (19)	2,072 (21)	2,028 (21)
Protein-coding genes with Isoforms (%)[c]	10,539	3,076 (29)	3,376 (32)	2,474 (23)	2,859 (27)	2,877 (27)
Only one isoform is expressed (%)[d]		2,783 (91)	3,063 (91)	2,294 (93)	2,589 (91)	2,617 (91)

Protein Identification

	A549	HEK293	HeLa	HepG2	MCF7
All PSM entries	83,217	102,753	100,806	89,415	85,994
All distinct peptides	36,150	40,204	40,208	36,479	42,910
Peptides 0.1% FDR, with 3 or more PSM	7,329	9,451	9,955	8,635	7,191
Proteins 1% FDR	2,129	2,885	2,561	2,364	2,062
Group I	572	739	742	687	559
Group Iue (Improved by expression information)	**55**	**87**	**67**	**68**	**54**
Group Ie (Improved by expression information)	**233**	**326**	**286**	**278**	**238**
Group O	407	577	434	419	389
Group Oe (Improved by expression information)	217	294	211	206	184
Group S	644	862	817	703	637
Proteins 1% FDR, with 2 or more supporting peptides (I+Iue+Ie)	860	1152	1095	1033	851

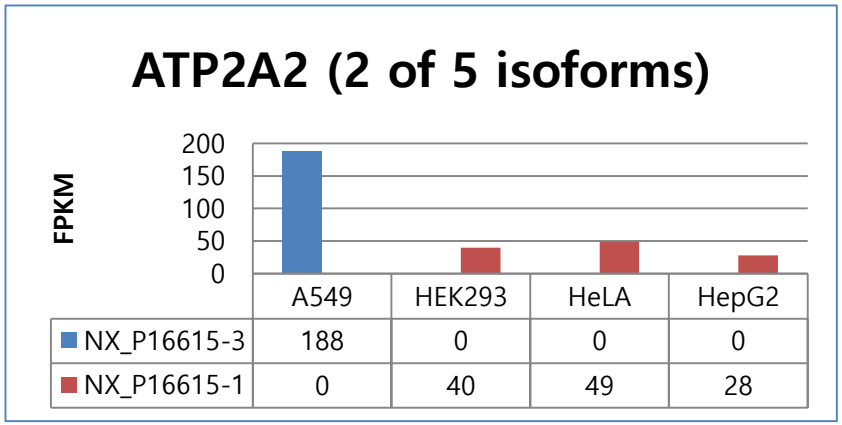
We identified **1,935** proteins from five cell lines.
Among 1,935 proteins, **841** proteins are further supported by transcript expression information

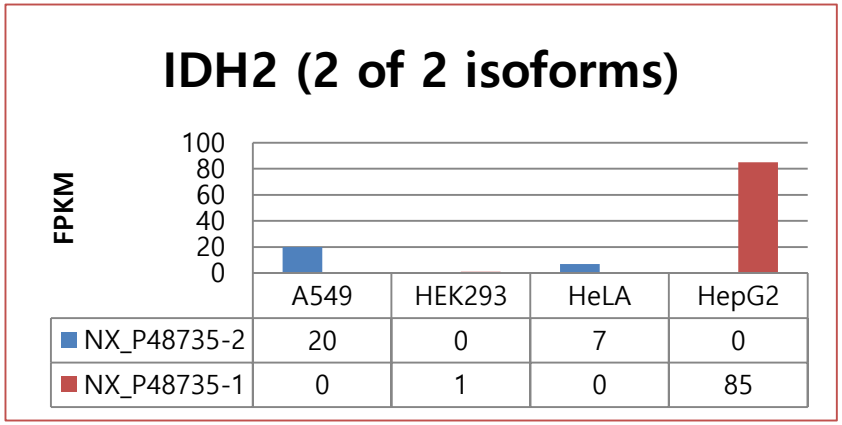
Common and Specific Proteins Identified from Five Cell Lines



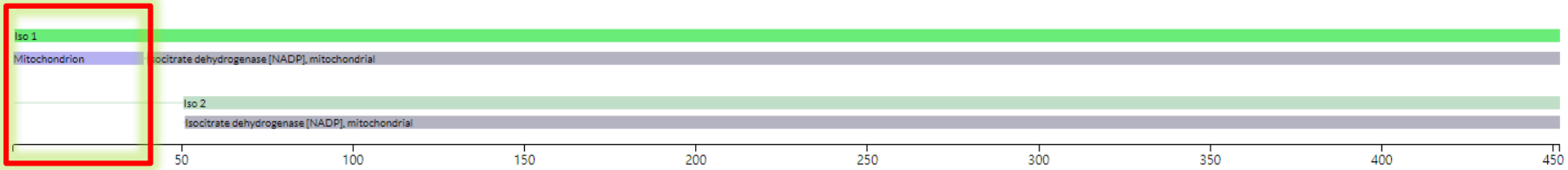
	Proteins
Only in each cell line	747
Commons in 2 cell lines	323
Commons in 3 cell lines	231
Commons in 4 cell lines	265
Commons in 5 cell lines	369

- So far, 19 cases of isoforms detected from 5 cell lines that have different isoform distribution. (e.g., ATP2A2, IDH2)





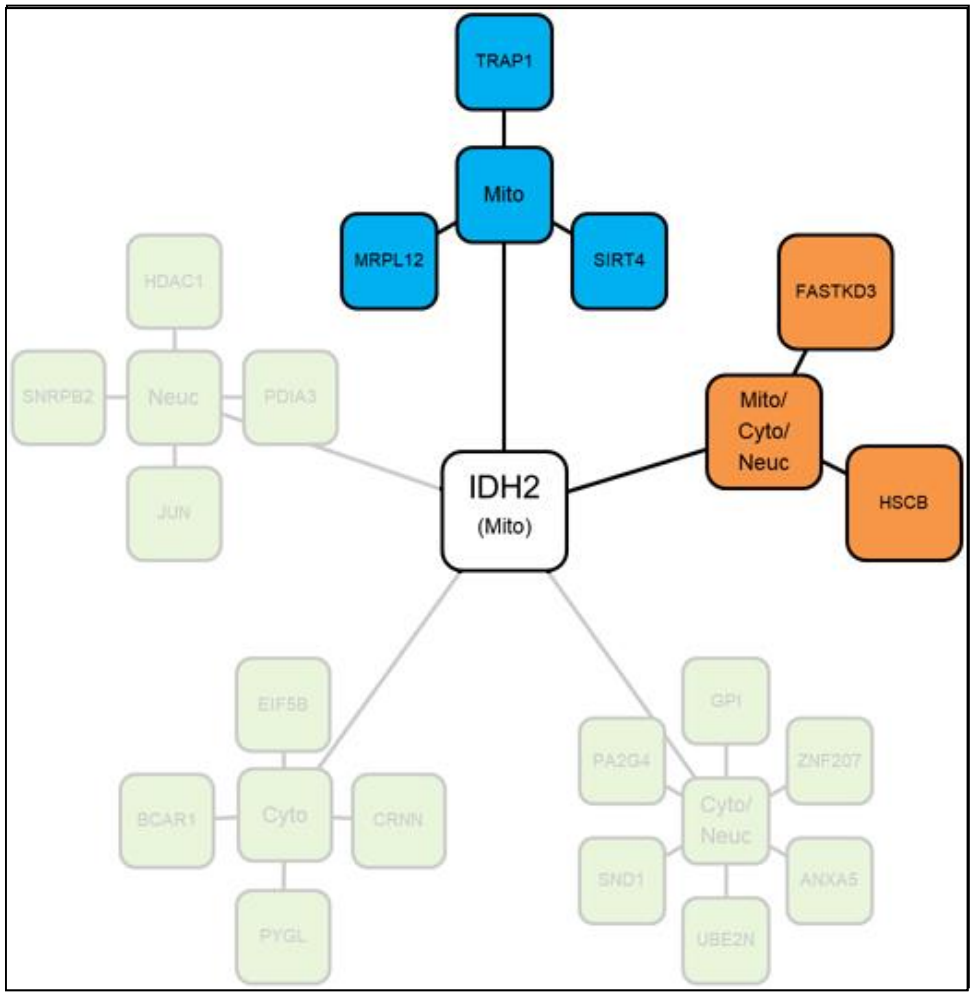
IDH2, Isocitrate dehydrogenase [NADH+]



- IDH2 has two known isoforms, *long* form (isoform **1**) and *short* form (isoform **2**)
- Short form lacks N-terminal transit sequence to mitochondria
- We think the existence of transit peptides could affect the localization of both isoforms of IDH2 and it may limit the interacting partners.

Interacting Partners of IDH2

- To estimate the effect of mitochondrial transit sequence, we analyzed the localization of 20 proteins which are known as interacting with IDH2.
  - 14 proteins are localized in cytoplasm or nucleus
  - And 3 proteins are localized in mitochondrion
- So we get crude clues that the existence of transit peptides could limit the interacting partners and could change biological activity.



Conclusions

- From the results of proteotypic peptide (by trypsin digestion) analysis for the neXtProt human protein sequences, **we figured out that 56% proteins are difficult to identify due to lack of suitable peptides**.
- The RNA-Seq data covered 25% of known proteome and 91% of protein-coding genes shows only one dominant isoform expression pattern.

- **Using this methods,**
  - **We were able to identify 1,935 distinct proteins** present in five cell lines by constructing cell-line specific protein sequence database under highly stringent identification condition.
  - **About a half of 1,935 identified proteins (841 proteins) were improved** with proper supporting peptides that are usually cannot be identified without transcript expression information.
  - Also we demonstrated 19 cases of differently expressed isoforms detected from 5 cell lines (e.g., IDH2 case).

Acknowledgment

- **Bioinformatics**
  - Seul-Ki Jeong (Ph. D), YPRC
  - Chae-Yeon Kim, YPRC

- **Dataset**
  - Uhlén M et. al, *Science*. **2015**, 347(6220):1260419
  - Geiger T et al., *MCP*. **2012**, 11(3):M111.014050

- **Project Director : Prof. Young-Ki Paik (Chair of C-HPP)**

This work was supported by a grant from the Korean Ministry of Health and welfare

THANK YOU

Appendix

- Protein group : group of isoforms from one gene
  - Proteins with same neXtProt accession.
  - e.g 1) NX_P23141-1, NX_P23141-2 and NX_P23141-3 are grouped as protein group NX_P23141.
  - Protein that have no known isoform also grouped as protein group
  - e.g 2) NX_A0AVK6-1 (and no known isoforms) is grouped as protein group NX_A0AVK6.