UNIVERSITÉ DE GENÈVE

SIB 15 YEARS Swiss Institute of Bioinformatics

neXtProt: The Human Protein Knowledge Platform in the Context of HPP

Amos Bairoch

September 16, 2013

HUPO 2013 Yokohama, Japan

neXtprot BETA

- **What**: a high-quality resource for human protein-centric information;
- **How**:
  - The foundation of neXtProt is the extensive compendium of continuously updated sequences and annotations provided by UniProtKB/Swiss-Prot;
  - We integrate, in the most appropriate way and with stringent quality criteria, data originating from many high-throughput data resources;
  - Currently this includes variants, PTMs, peptide identifications, protein/protein interactions, subcellular locations, but we want to also add protein/small molecules interactions, pathways/networks information, siRNA screen data, phylogenetic profiling, etc.

Gold, silver and bronze

- We have a three-tiered approach as to data quality:
  - Gold: data that we believe to be of a Swiss-(prot)-level quality (>=99% correct)
  - Silver: good data, but….. (>=95% correct).
  - Bronze: noisy or low quality data that is not imported in neXtProt;
- Quality classification is a dynamic process and is done whenever as possible with the active involvement of the data provider;
- Quality grading is done at the level of annotations. A neXtProt entry is neither gold nor silver but a particular annotation (a phosphosite for example) can be gold or silver.

What is not neXtProt?

- neXtProt is not a replacement for **UniProtKB/Swiss-Prot.** It is not universal in coverage. It is intended to provide knowledge pertinent to human proteins;
- neXtProt is not a sequence resource: it uses the sequence data curated in Swiss-Prot;
- neXtProt is not a proteomics repository. Protein identification data should be submitted to resources that participate to **ProteomeXchange**.

Why a gene-centric resource can't be used as the basis of a protein-centric analysis

Biological issues

- More than one gene can encode for identical proteins (ex: Histone H4 with 14 genes);
- As above but after processing. Signal/transit removal or other proteolytic event (ex: ubiquitin);
- More than one protein can be encoded by one gene; These are extreme cases of splicing where the shared exon(s) are non-coding;
- More than one protein encoded by one transcript due to bicistronic mRNAs;
- Some proteins can be encoded by transcripts arising from two adjacent genes;

- More than one transcript can encode for the same protein (they will differ in the non-coding regions);
- Proteins produced by somatic rearrangements (immunoglobulins) are not correctly represented;

Consensus human genome issues

- The consensus genome does not represent all genes (some individuals have additional copies of genes);
- Some genes can be either protein-coding or pseudogene depending on individuals but are arbitrarily assigned to one of the two categories;

Issues arising from annotation problems

- Ensembl is not in synch in the annotation of protein-coding genes with the efforts of the CCDS consortium;
- At each release Ensembl corrects errors but unfortunately introduce new ones…

How many protein-coding genes?

- We completely agree with recent reports by the HAVANA or NCBI genome annotation groups: that there are slightly less then 20’000 protein-coding genes;
- There are been various recent publications that report that there is a wealth of “new” small proteins that are “hidden” in:
  - The 5’ or 3’ regions of transcripts;
  - In the opposite direction as that of the main product of a transcript;
  - In regions of the genome shown by ENCODE to be transcribed but where no gene had yet been detected.
- Some of these proteins are probably real, but:
  - The lack of conservation of these potential proteins in other species is suspicious;
  - The absence of any genetic or biochemical characterization is also disturbing. If there are important why have they not turned out in any functional assays?
- So it would be nice to specifically hunt for these proteins but as *extraordinary claims require extraordinary evidence*, any such identification would require careful manual confirmation.

Some neXtProt stats

ENTRIES	Protein entries	**20,128**
	Isoforms (produced by splicing)	**39,325**
PROTEINS EXISTENCE	Entries whose protein(s) existence is based on evidence at protein level	**15,646**
	Entries whose protein(s) existence is based on evidence at transcript level	**3,570**
	Entries whose protein(s) existence is based on homology	**187**
	Entries whose protein(s) existence is based on a prediction (gene model)	**87**
	Entries whose protein(s) existence is uncertain	**638**
ANNOTATIONS	Identifiers	**794,822**
	Interactions	**55,923**
	Post-translational modifications (PTMs)	**96,366**
	Proteins with a disease annotation	**3,029**
	Proteins with an experimental 3D structure	**4,921**
	Proteins with expression based on ESTs	**14,719**
	Proteins with expression based on IHC	**12,437**
	Proteins with expression based on microarrays	**16,705**
	Variants (including disease mutations)	**854,404**

How many human proteins have been observed?

- UniProtKB and neXtProt use the same criteria to assign a protein existence category to entries;
- Five categories are defined:
  - “Evidence at protein level”
  - “Evidence at transcript level”
  - “Inferred from homology”
  - “Predicted”
  - “Uncertain”
- For an entry to be assigned “Evidence at protein level” does not necessarily means that one of the protein described by that entry has been identified by proteomics as there are other methodologies that prove that a protein exist (Edman, X-ray, Abs, etc);

How many human proteins have been observed?

- In UniProtKB/Swiss-Prot there are 13’659 protein entries assigned to “Evidence at protein level” (67%) while neXtProt has 15’646 protein entries (78%);
- This difference is due to the additional data integrated in neXtProt (specifically PTMs and peptide identifications);
- Another important number is that of entries assigned to “Uncertain”: 638 (3.2%). Their existence is quite dubious and are not expected to be found by HPP. So identification of such proteins should be further investigated.

HPP quest for the ‘missing proteins'

imental validation

otein level AND
ıfo

otein level AND other

otein level AND no
)

Proteins 1 to 10 of 208 show 10 summary | **details** export

☆ Dihydrolipoyl dehydrogenase, mitochondrial *(DLD)* [NX_P09622]
Lipoamide dehydrogenase is a component of the glycine cleavage system as well as of the alpha-ketoacid dehydrogenase complexes. Involved in the hyperactivation of spermatazoa during capacitation and in the spermatazoal acrosome reaction.
Gene location: 7q31.1 Isoforms: 1 Variants: 16 PTMs: 8
Disease: yes 3D structure: yes Proteomics: yes Tissue expression: yes Mutagenesis: no

☆ Isocitrate dehydrogenase [NADP], mitochondrial *(IDH2)* [NX_P48735]
Plays a role in intermediary metabolism and energy production. It may tightly associate or interact with the pyruvate dehydrogenase complex.
Gene location: 15q26.1 Isoforms: 1 Variants: 2 PTMs: 13
Disease: no 3D structure: no Proteomics: yes Tissue expression: yes Mutagenesis: no

☆ Aldehyde dehydrogenase, mitochondrial *(ALDH2)* [NX_P05091]
Cellular response to fatty acid. Cellular response to hormone stimulus. Liver development.
Gene location: 12q24.12 Isoforms: 1 Variants: 16 PTMs: 1
Disease: no 3D structure: yes Proteomics: yes Tissue expression: yes Mutagenesis: no

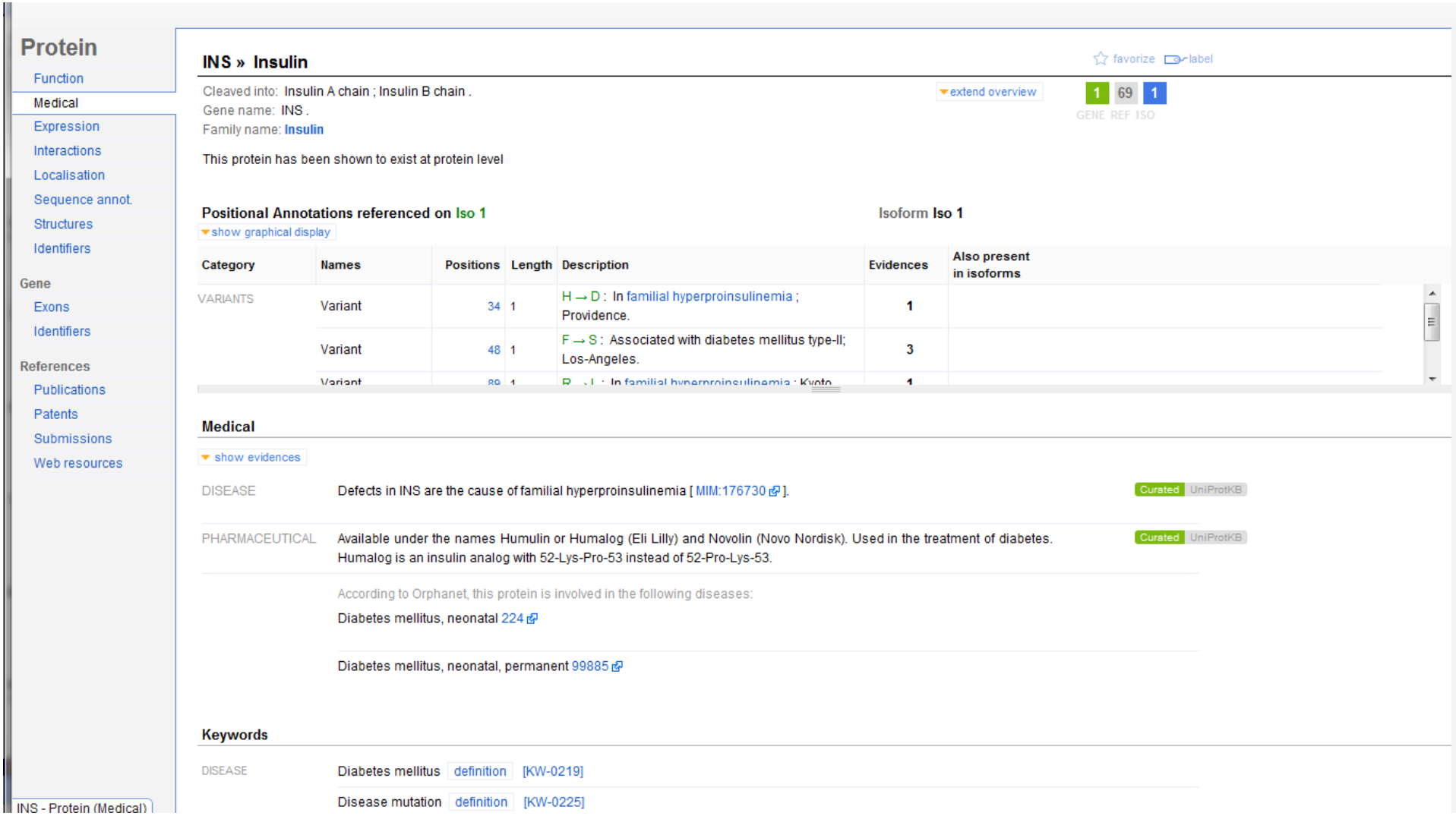
☆ Electron transfer flavoprotein subunit alpha, mitochondrial *(ETFA)* [NX_P13804]
The electron transfer flavoprotein serves as a specific electron acceptor for several dehydrogenases, including five acyl-CoA dehydrogenases, glutaryl-CoA and sarcosine dehydrogenase. [more]
Gene location: 15q24.2 Isoforms: 1 Variants: 10 PTMs: 1
Disease: yes 3D structure: yes Proteomics: yes Tissue expression: yes Mutagenesis: no

☆ Acyl-CoA dehydrogenase family member 9, mitochondrial *(ACAD9)* [NX_Q9H845]
Has a dehydrogenase activity on palmitoyl-CoA (C16:0) and stearoyl-CoA (C18:0). It is three times more active on palmitoyl-CoA then on stearoyl-CoA. Has little activity on octanoyl-CoA (C8:0), butyryl-CoA (C4:0) or isovaleryl-CoA (5:0).
Gene location: 3q21.3 Isoforms: 1 Variants: 5 PTMs: 2
Disease: yes 3D structure: no Proteomics: yes Tissue expression: yes Mutagenesis: no

☆ Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial *(PDHA1)* [NX_P08559]
The pyruvate dehydrogenase complex catalyzes the overall conversion of pyruvate to acetyl-CoA and CO(2). It contains multiple copies of three enzymatic components: pyruvate dehydrogenase (E1), dihydrolipoamide acetyltransferase (E2) and lipoamide dehydrogenase (E3).
Gene location: Xp22.12 Isoforms: 1 Variants: 31 PTMs: 11
Disease: yes 3D structure: yes Proteomics: yes Tissue expression: yes Mutagenesis: no

☆ Aconitate hydratase, mitochondrial *(ACO2)* [NX_Q99798]
Catalyzes the isomerization of citrate to isocitrate via cis-aconitate (By similarity).
Gene location: 22q13.2 Isoforms: 1 Variants: 23 PTMs: 5
Disease: no 3D structure: no Proteomics: yes Tissue expression: yes Mutagenesis: no

A variety of views for a single protein



The sequence viewer

VAV1 » Proto-oncogene vav

favorite label

Gene name: VAV1

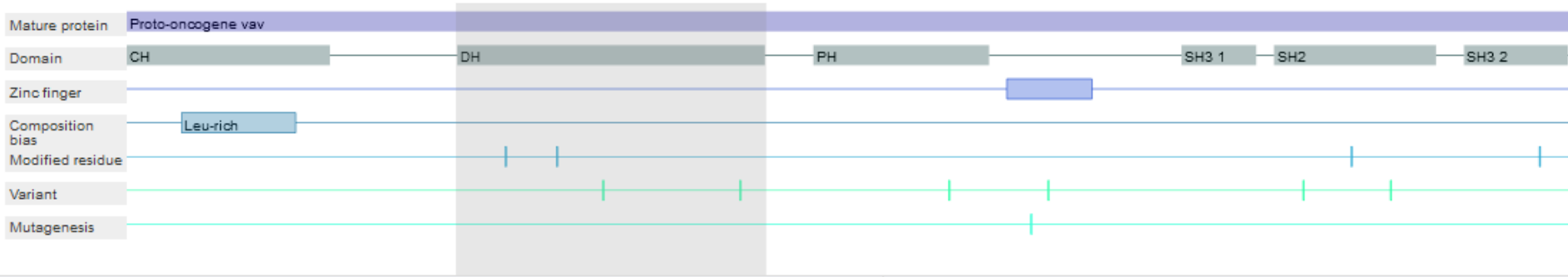
extend overview

1 81 1
GENE REF ISO

This protein has been shown to exist at protein level

Displayed isoform: Iso 1

Processing Region Modified residue Variant Conflict All/None



Name	Position	Length	Description	Evidence
Domain	194 - 373	180	DH	UniProtKB
Domain	402 - 504	103	PH	UniProtKB
Domain	617 - 660	44	SH3 1	UniProtKB
Domain	671 - 765	95	SH2	UniProtKB
Domain	782 - 842	61	SH3 2	UniProtKB
Zinc finger	515 - 564	50	Phorbol-ester/DAG-type	UniProtKB
Composition bias	33 - 99	67	Leu-rich	UniProtKB
Modified residue Modified residue	222	1	N6-acetyllysine	1 UniProtKB

Isoform **Iso 1** 845 aa, Mass: 98314 Da, pI: 6.2

view FASTA | BLAST sequence | BLAST selection

```
  1 MELWRQCTHW LIQCRVLPPS HRVTWDGAQV CELAQALRDG VLLCQLLNNL
 51 LPHAINLREV NLRPQMSQFL CLKNIRTFLS TCCEKFGLKR SELFEAFDLF
101 DVQDFGKVIY TLSALSWTPI AQNRGIMPFP TEEESVGDED IYSGLSDQID
151 DTVEEDEDLY DCVENEEAEG DEIYEDLMRS EPVSMPPKMT EYDKRCCCLR
201 EIQQTEEKYT DTLGSIQQHF LKPLQRFLKP QDIEIIFINI EDLLRVHTHF
251 LKEMKEALGT PGAANLYQVF IKYKERFLVY GRYCSQVESA SKHLDRVAAA
301 REDVQMKLEE CSQRANNGRF TLRDLLMVPM QRVLKYHLLL QELVKHTQEA
351 MEKENLRLAL DAMRDLAQCV NEVKRDNETL RQITNFQLSI ENLDQSLAHY
401 GRPKIDGELK ITSVERRSKM DRYAFLLDKA LLICKRRGDS YDLKDFVNLH
451 SFQVRDDSSG DRDNKKWSHM FLLIEDQGAQ GYELFFKTRE LKKKWMEQFE
501 MAISNIYPEN ATANGHDFQM FSFEETTSCK ACQMLLRGTF YQGYRCHRCR
551 ASAHKECLGR VPPCGRHGQD FPGTMKKDKL HRRAQDKKRN ELGLPKMEVF
601 QEYYGLPPPP GAIGPFLRLN PGDIVELTKA EAEQNWWEGR NTSTNEIGWF
651 PCNRVKPYVH GPPQDLSVHL WYAGPMERAG AESILANRSD GTFLVRQRVK
701 DAAEFAISIK YNVEVKHIKI MTAEGLYRIT EKKAFRGLTE LVEFYQQNSL
751 KDCFKSLDTT LQFPFKEPEK RTISRPAVGS TKYFGTAKAR YDFCARDRSE
801 LSLKEGDIIK ILNKKGQQGW WRGEIYGRVG WFPANYVEED YSEYC
```

Information at the genomic level

Gene information

Chromosomal location: 16p13.3
Orientation: plus strand
Ensembl: ENSG00000140992

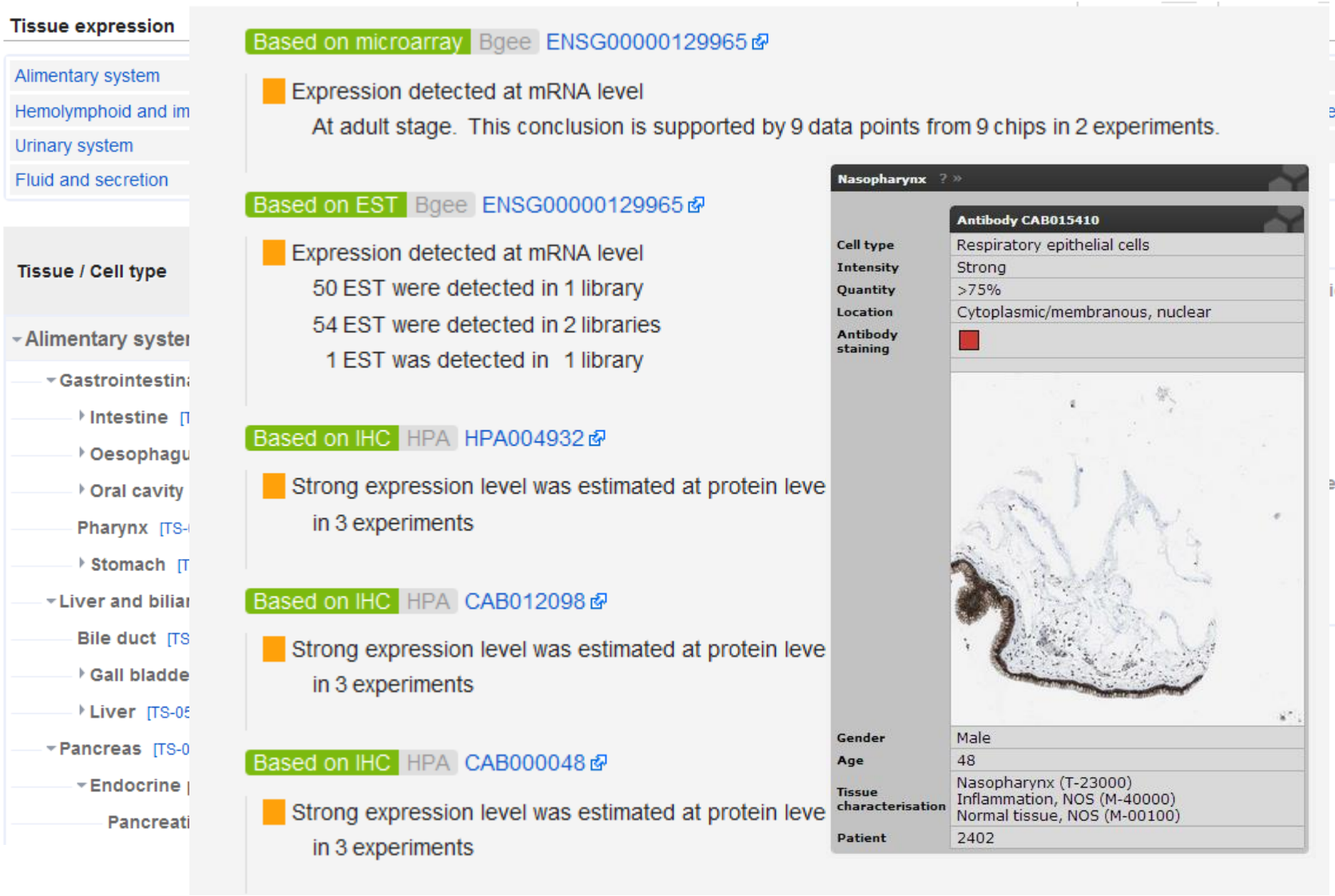
Ensembl

The gene codes for 4 isoforms
Coding positions: from 2588114 to 2647765 [length: 59652 bp]

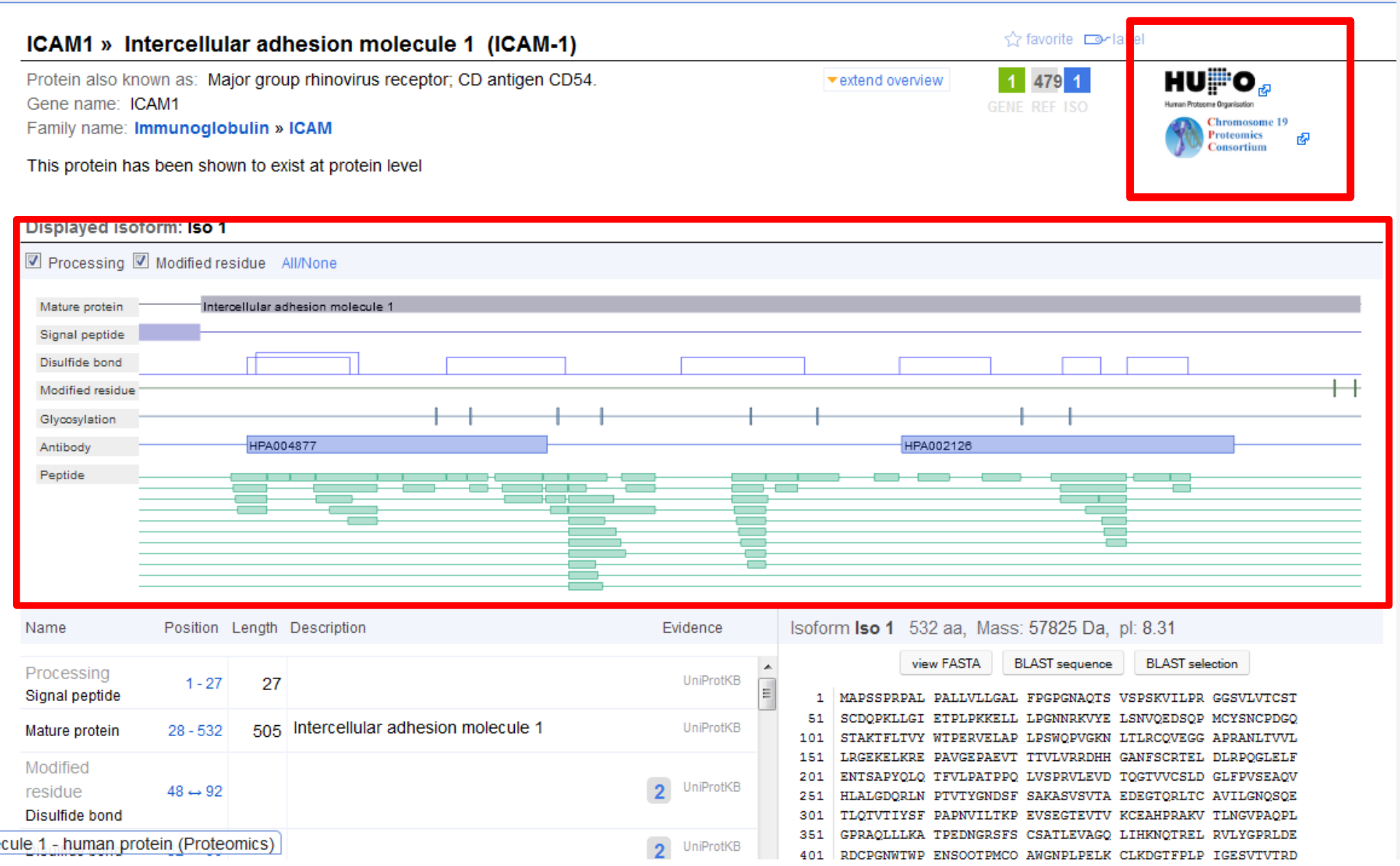
Exons

Identifier	Position on gene	Length	Coding for Iso 1 ENST00000342085	Coding for Iso 2 ENST00000342085 show transcripts (1)	Coding for Iso 3	Coding for Iso 4 ENST00000268673
ENSE00001944688	1 - 173	173	Met 1 - Leu 8			
ENSE00001867541	40 - 173	134				Met 1 - Leu 8
ENSE00002298089	70 - 173	104			Met 1 - Leu 8	
ENSE00001640744	19740 - 20000	261	Tyr 9 - Thr 95	Met 1 - Thr 45		Tyr 9 - Thr 95
ENSE00001620065	23517 - 23559	43	Val 96 - Ile 110	Val 46 - Ile 60	Val 96 - Ile 110	Val 96 - Thr 110
ENSE00001909937	23808 - 23945	138	Ile 110 - Tyr 156	Ile 60 - Tyr 106	Ile 110 - Tyr 156	
ENSE00001854073	27590 - 27734	145	Tyr 156 - Arg 204	Tyr 106 - Arg 154	Tyr 156 - Arg 204	
ENSE00001705748	28393 - 28490	98	Arg 204 - Ala 237	Arg 154 - Ala 187	Arg 204 - Ala 237	
ENSE00001811853	39462 - 39537	76	Ala 237 - Ser 262	Ala 187 - Ser 212		Thr 110 - Ser 135
ENSE00001943440	43332 - 43400	69	Ser 262 - Gly 285	Ser 212 - Gly 235		Ser 135 - Gly 158
ENSE00001810733	43644 - 43740	97	Gly 285 - Leu 317	Gly 235 - Leu 267		Gly 158 - Leu 190
ENSE00001859498	45449 - 45622	174	Val 318 - Asn 375	Val 268 - Asn 325	Val 292 - Asn 349	Val 191 - Asn 248
ENSE00000946017	48713 - 48930	218	Tyr 376 - Trp 448	Tyr 326 - Trp 398	Tyr 350 - Trp 422	Tyr 249 - Trp 321
ENSE00000946018	57830 - 57887	58	Trp 448 - Lys 467	Trp 398 - Lys 417	Trp 422 - Lys 441	Trp 321 - Lys 340
ENSE00000946019	59160 - 59312	153	Gly 468 - Thr 518	Gly 418 - Thr 468	Gly 442 - Thr 492	Gly 341 - Thr 391
ENSE00002320580	59688 - 59944	257			Pro 493 - Gln 530	
ENSE00001505218	59688 - 65225	5538	Pro 519 - Gln 556	Pro 469 - Gln 506		Pro 392 - Gln 429

Expression data at mRNA and protein levels



The proteomics view



Peptide identifications

- All human peptides from the latest release (Aug 2013) of PeptideAtlas;
- Sets linked with PTMs;
- Carapito et al mitochondrial N-terminome project;
- Sanchez BD-HPP diabetes project pancreatic islet set.

PTMs

Evidence 4: Inferred from Experiment neXtProt

System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation.
Rigbolt K.T., Prokhorova T.A., Akimov V., Henningsen J., Johansen P.T., Kratchmarova I., Kassem M., Mann M., Olsen J.V., Blagoev B.
Sci Signal **4**, rs3-rs3 (2011) [Full text:10.1126/scisignal.2001570] [PubMed:21406692]

Show abstract

Identification, phosphorylation and N-acetylation of embryonic stem cell proteins

Hide experimental details

Detection method
Mass spectrometry Nano LC-MS/MS.
Cell line
HUES9 [CVCL_0057].
Sample preparation
Protein reduction, alkylation, followed by digestion with endoproteinase Lys-C and trypsin. Peptide fractionation by Strong Cation eXchange column (SCX) and phosphopeptides enrichment by titanium dioxide (TiO2) chromatography.
Instrument/platform
Nanoscale C18 HPLC coupled online to a LTQ FT Ultra, LTQ Orbitrap or LTQ Orbitrap mass spectrometer equipped with a nanoelectrospray source (Proxeon).
Data analysis procedure
Protein database: human IPI release 3.37 concatenated with known contaminants and reversed sequences of all entries. Software: MaxQuant version 1.0.12.25; Mascot version 2.2. Maximum missed cleavages: 3. Mass tolerance for fragment ion: 0.5 Da. Fixed modification: Cys carbamidomethylation. Variable modifications: Met oxidation; N-terminal acetylation; pyro-glutamate for N-terminal

sets of PTMs

ed N-glycosylation,
on, sumolyation

en added;

wiss-Prot recently
high-throughput
v years back and
quality sites.

Subcellular localization data

Endoplasmic reticulum definition [SL-0095] silver

2 refs

EXP GFP-cDNA@EMBL

Evidence 1: Inferred from Experiment GFP-cDNA@EMBL

hlcc3_51a15b

DKFZ GFP-cDNA localisation project

Hide experimental details

Description

Transient overexpression of human open reading frames as N- and C-terminal fusions with CFP and YFP.

Detection method

Fluorescence microscopy.

Cell line

Monkey Vero cells [CVCL_0059].

Treatment/physiological state

Overexpression of cDNA clones under the CMV promoter.

Sample preparation

Live cells.

Instrument/platform

Leica DM/IRBE microscope with 63x NA 1.4PL objective and custom designed CFP or YFP filters.

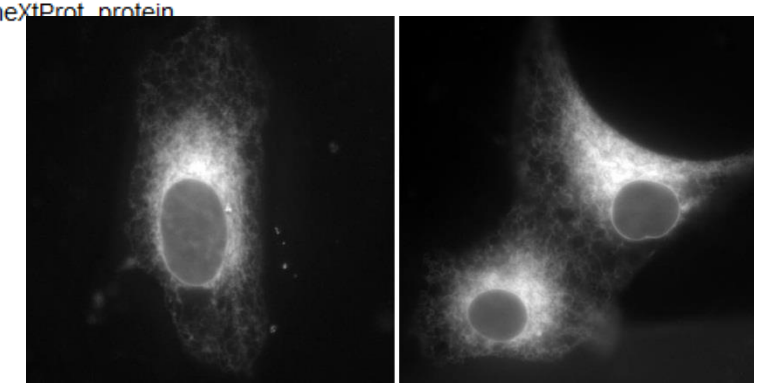
Data processing by neXtProt

Exclusion of partial cDNAs (236 clones). Clones are aligned to neXtProt protein sequences for assignment to entries.

Data confidence documentation

SILVER. Full length clones (680 isoforms for 667 proteins).

y

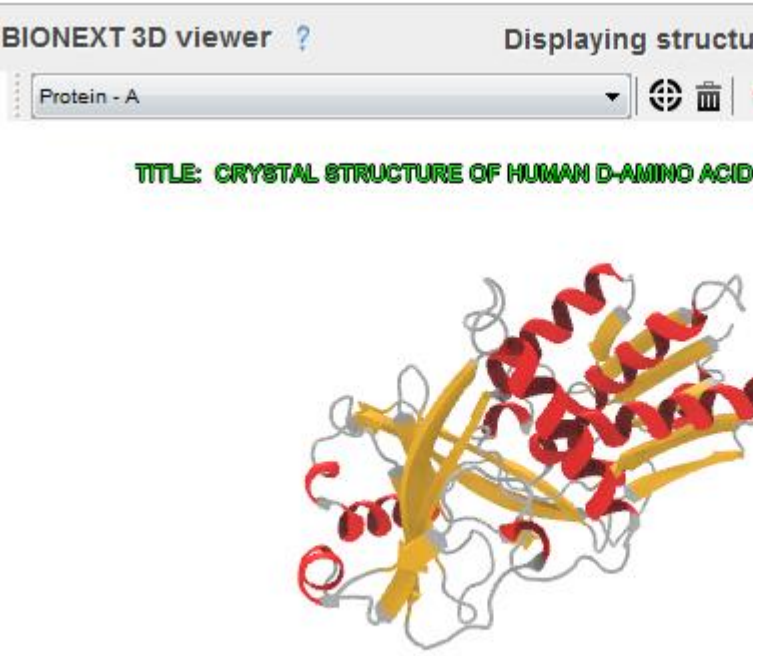


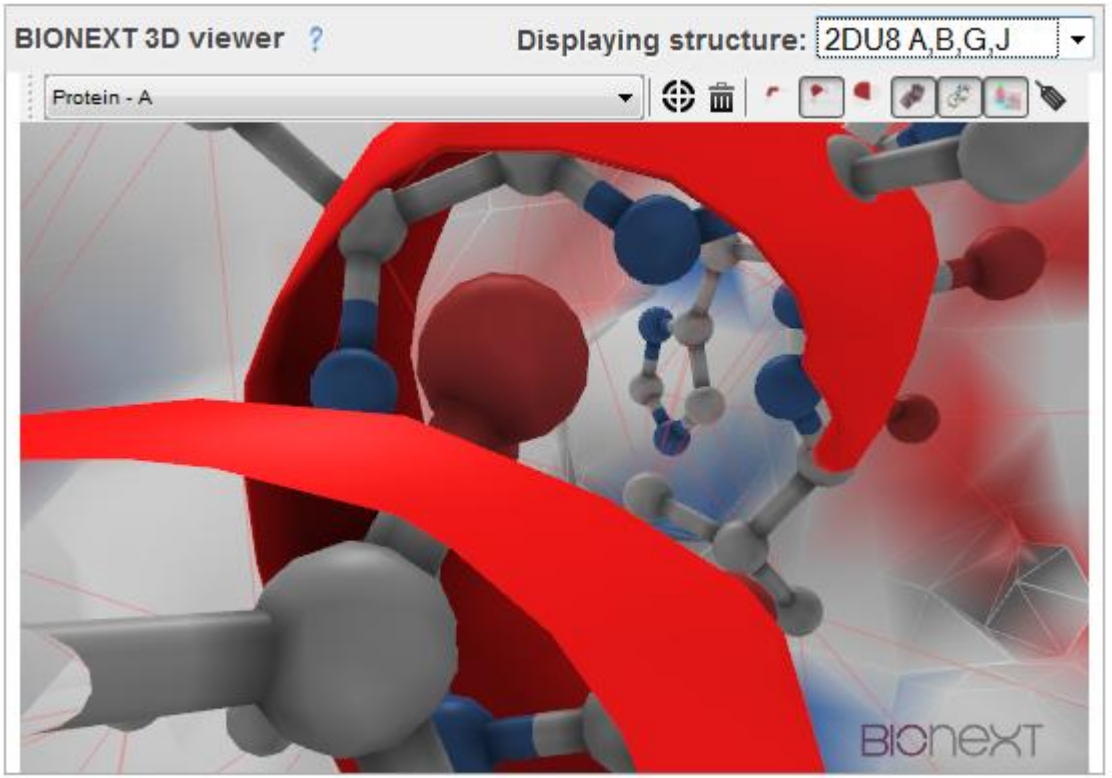
Variants

- We now have almost 850’000 protein variants in neXtProt;
- We complemented the 68’000 from UniProtKB (Gold) with a wealth of variants from dbSNP and COSMIC (Silver);
- This means an average of 40 variants per human protein and this will continue to grow as we enter the era of personalized genome sequences.

3D viewer

- We have integrated a 3D structure visualisation tool (BioWiz from BIONEXT);
- We plan to use it to display position-specific annotations (such as PTMs) in the context of the structure.





Data export

- Export of data both in XML and in PEFF formats;
- neXtProt is the first resource to offer support to the PSI PEFF format;
  - This enriched FASTA format allows search engines and other tools to easily and consistently access data essential to the success of HPP, namely sequence variations and PTMs;
  - Unfortunately while MS identification vendors requested this format, they do not yet support it!

Download by FTP

- At ftp.nextprot.org
- To obtain downloads in XML or PEFF;
- These files are also available per chromosome as well as ‘report’ files

```
Description: Chromosome 17 report
Name:        nextprot_chromosome_17
Release:     2011-08-23

-------------------------------------------------------------------------------

This file lists all neXtProt entries on chromosome 17
Total number of genes: 1182

-------------------------------------------------------------------------------------------------------------------------------
Gene      neXtProt    Chromosomal  Start     Stop      Protein          Prote-  Ab   3D   Dise-  Iso  Var  PTMs  Description
name      AC          position     position  position  evidence         omics                ase
_______________________________________________________________________________________________________________________________
DOC2B     NX_Q14184   17p13.3       6007      31427    protein level      no     no   no   no     1     0     0   Double C2-like domain-contain
RPH3AL    NX_Q9UNE2   17p13.3      62293     202888    protein level      no     yes  no   no     2     7     0   Rab effector Noc2
C17orf97  NX_Q6ZQX7   17p13.3     260118     264367    transcript level   no     yes  no   no     4     8     0   Uncharacterized protein C17or
FAM101B   NX_Q8N5W9   17p13.3     289769     295730    protein level      yes    no   no   no     1     0     1   Protein FAM101B
VPS53     NX_Q5VIR6   17p13.3     411908     618096    protein level      no     yes  no   no     3     5     4   Vacuolar protein sorting-asso
FAM57A    NX_Q8TBR7   17p13.3     635847     646074    protein level      no     no   no   no     2     3     0   Protein FAM57A
GEMIN4    NX_P57678   17p13.3     647661     655501    protein level      no     no   no   no     1    30     2   Component of gems 4
GLOD4     NX_Q9HC38   17p13.3     662550     686505    protein level      yes    yes  no   no     3     5     1   Glyoxalase domain-containing
RNMTL1    NX_Q9HC36   17p13.3     685513     695749    protein level      yes    yes  no   no     1     9     0   RNA methyltransferase-like pr
NXN       NX_Q6DKJ4   17p13.3     702581     883010    protein level      no     yes  no   no     2     1     0   Nucleoredoxin
TIMM22    NX_Q9Y584   17p13.3     900357     905388    protein level      yes    no   no   no     1     8     0   Mitochondrial import inner me
ABR       NX_Q12979   17p13.3     906758    1090616    protein level      no     no   no   no     2    11     0   Active breakpoint cluster reg
BHLHA9    NX_Q7RTU4   17p13.3    1173853    1174754    homology           no     no   no   no     1     0     0   Class A basic helix-loop-heli
TUSC5     NX_Q8IXB3   17p13.3    1182957    1204281    transcript level   no     no   no   no     1    11     0   Tumor suppressor candidate 5
YWHAE     NX_P62258   17p13.3    1247566    1303505    protein level      yes    yes  yes  no     2     6     8   14-3-3 protein epsilon
CRK       NX_P46108   17p13.3    1323983    1359552    protein level      no     yes  yes  no     2     7     9   Adapter molecule crk
MYO1C     NX_O00159   17p13.3    1367480    1395995    protein level      yes    yes  no   no     3    26     7   Myosin-Ic
INPP5K    NX_Q9BT40   17p13.3    1397872    1420182    protein level      no     yes  no   no     2    17     0   Inositol polyphosphate 5-phos
PITPNA    NX_Q00169   17p13.3    1421287    1466110    protein level      no     yes  yes  no     1     3     2   Phosphatidylinositol transfer
SLC43A2   NX_Q8N370   17p13.3    1477666    1532180    protein level      no     yes  no   no     3     9     5   Large neutral amino acids tra
SCARF1    NX_Q14162   17p13.3    1537152    1549083    protein level      no     yes  no   no     5    21    23   Scavenger receptor class F me
```

Programmatic access

- We have developed a first release of an API to allow third-party software tools to make use of the data in neXtProt;
- You can currently access PTMs, variants, subcellular localisations and expression both at entry («gene») or isoform («splice») level;
- We will soon increase the scope of the API to allow all neXtProt data to be programatically accessible;
- This REST API is available at:

http://www.nextprot.org/rest/

What we want to do in the near future

- Deploy an advanced search capability so that you can query very precisely neXtProt. Example:

  Give me all proteins from chr18 with 2 TM regions that are highly expressed in liver (HPA) and which have proteomics identification in PeptideAtlas

- Load additional sets of PTMs, peptide identifications, subcellular locations and variants;
- Tools for protein lists analysis.

neXtProt team at the SIB

- **Content:**
  - Coordinator: Pascale Gaudet
  - Biocurators: Guislaine Argoud-Puy, Aurore Britan, Jonas Cicenas, Isabelle Cusin, Paula Duek, Nevila Nouspikel and Ying Zhang (from T. Yamamoto's group)
- **Quality assurance:**
  - Monique Zahn
- **Software:**
  - Coordinator: Pierre-André Michel
  - Developers: Olivier Evalet, Alain Gateau, Anne Gleizes, Mario Pereira, Daniel Teixeira
- **Directed by:**
  - Amos Bairoch and Lydie Lane