Knowledge Base pillar Committee (KBpC) Update

Eric Deutsch
Institute for Systems Biology

Bioinformatics Hub Recap

- Open all week (SMTW 8:30-5:30) (Sunday at UCD)
- Hub program was online early and linked to from HUPO program
- Hub program dynamically updated during the congress

- Some of the lively discussions:
  - Quality Control in proteomics – tools and standards
  - Spectral clustering
  - Statistics in proteomics – decoy database, entrapment databases, protein grouping
  - Immunopeptidomics
  - Confidence in PTM localization
  - Universal Spectrum Identifier
  - Proteogenomics
  - Real-time proteomics

Timeline for KB releases & SI next year?

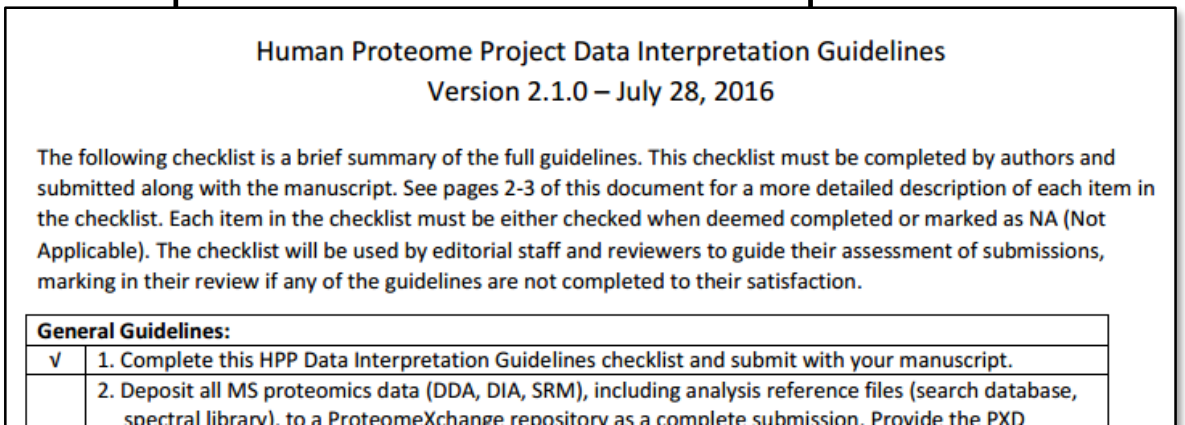
- For this 2017 Special Issue:
  - 2017-01 PeptideAtlas release
  - 2017-02 neXtProt release
  - 2017-05-01 Initial JPR deadline
  - 2017-10-01 Final JPR acceptance deadline
- Next year 2018:
  - What shall the schedule be?
  - PeptideAtlas and neXtProt are aiming for same release schedule
  - There is rather little time to alter course at this point, but…
  - A final decision this week would be very welcome

HPP Guidelines 2.1 - 2016

- After initial design at HUPO 2015, Guidelines 2.0 released for 2016

- Manuscript written for the Special Issue discussing guidelines at length
- Triggered more minors clarifications
- Version 2.0.5 aligned with manuscript and manuscript submitted

- One reviewer prompted one substantive change
- And other minor clarifications
- Released version 2.1 with the publication of the manuscript
- Stable since then



Human Proteome Project Data Interpretation Guidelines
Version 2.1.0 – July 28, 2016

The following checklist is a brief summary of the full guidelines. This checklist must be completed by authors and submitted along with the manuscript. See pages 2-3 of this document for a more detailed description of each item in the checklist. Each item in the checklist must be either checked when deemed completed or marked as NA (Not Applicable). The checklist will be used by editorial staff and reviewers to guide their assessment of submissions, marking in their review if any of the guidelines are not completed to their satisfaction.

General Guidelines:	
√	1. Complete this HPP Data Interpretation Guidelines checklist and submit with your manuscript.
	2. Deposit all MS proteomics data (DDA, DIA, SRM), including analysis reference files (search database, spectral library), to a ProteomeXchange repository as a complete submission. Provide the PXD

How did version 2.1 fare for SI 2017?

- Checklists were submitted for all manuscripts

- But not always with care the first time
- Some items were not checked
- Substantial numbers of items were checked without actually being done

- However, by final review most papers complied reasonably with most guidelines

What should we do differently in 2018?

- Guidelines are basically good, and stability is encouraged
- But in order to improve, some possible enhancements:

1. Encourage/require use of Universal Spectrum Identifier for key spectra backing extraordinary claims (MPs)
2. Require not just a check mark, but also page # / location
3. Clarification of guidelines for DIA/SWATH datasets
4. Continue to allow “candidate detections” for non-compliant evidence?
5. Clarify proteins that are allowable exceptions to the 2 non-nested ≥9 AA peptides rule
6. Do we need guidelines for PTMs??

Universal Spectrum Identifier

- It is a widespread problem that requests for “annotated spectrum evidence” result in PDF screenshots or other renderings that resist close inspection or re-analysis

- Further, if full raw data are available, a reanalysis may not reproduce the finding, but without a spectrum identifier, it is hard to know which spectrum was the supposed evidence

- Define and implement widespread repository support for:
  - mzspec:PXD002145//HeLa45_20_160423/scan/14321

Universal Spectrum Identifier:

mzspec:PXD003294::00603_F01_P004608_B00F_A00_R1:15220:SSLLDVLAAR/2 FIND

Spectrum for SSLLDVLAAR +2

Ions:

a ☑ 1+ ☐ 2+ ☐ 3+
b ☑ 1+ ☑ 2+ ☐ 3+
c ☐ 1+ ☐ 2+ ☐ 3+
x ☐ 1+ ☐ 2+ ☐ 3+
y ☑ 1+ ☑ 2+ ☐ 3+
z ☐ 1+ ☐ 2+ ☐ 3+

[Deselect All]

Neutral Loss:

☑ NH_3 (*)
☑ H_2O (o)
☐ H_3PO_4 (p)
☑ **Immonium ions**
☐ **Reporter ions**

Mass Type:
◉ **Mono** ○ **Avg**

Mass Tol: 0.01

Update

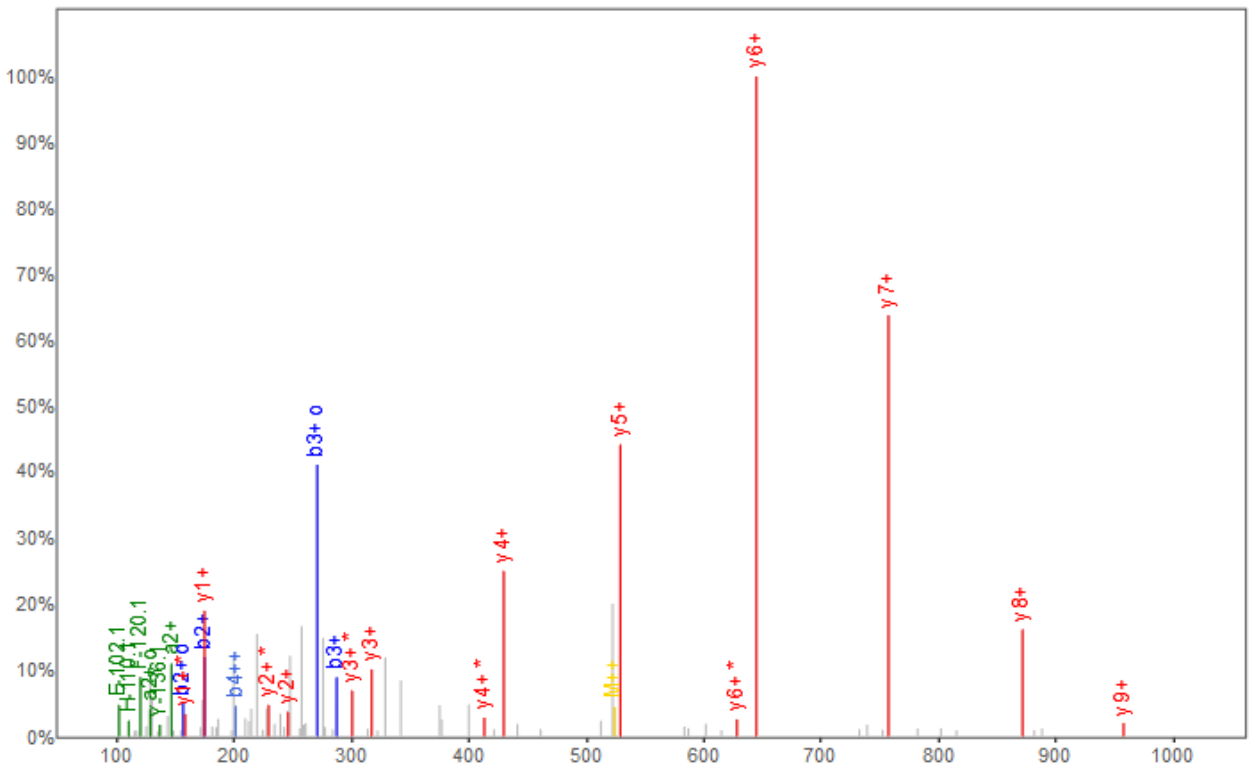
Peak Assignment:
◉ **Most Intense**
○ **Nearest Match**
☑ **Peak Detect**

Peak Labels:
◉ **Ion** ○ **m/z**
○ **None**

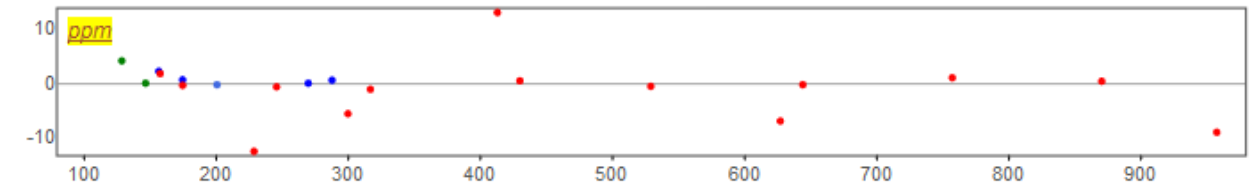
Width: 650

SSLLDVLAAR, MH+ 1044.6048, m/z 522.8060

File: 00603_F01_P004608_B00F_A00_R1.15220.15220.2, Scan: 15220, Exp. m/z: 522.80601896688, Charge: 2



Click and drag in the plot to zoom X: ☑ Y: ☐ Zoom Out Print ☐ Enable tooltip ☑ Plot mass error



a+	b+
60.0444	88.03
147.0764	175.0
260.1605	288.1
373.2445	401.2
488.2715	516.2
587.3399	615.3
700.4240	728.4
771.4611	799.4
842.4982	870.4

[Click] to move tabl

Checklist → Page numbers

Human Proteome Project Data Interpretation Guidelines
Version 2.1.0 – July 28, 2016

The following checklist is a brief summary of the full guidelines. This checklist must be completed by authors and submitted along with the manuscript. See pages 2-3 of this document for a more detailed description of each item in the checklist. Each item in the checklist must be either checked when deemed completed or marked as NA (Not Applicable). The checklist will be used by editorial staff and reviewers to guide their assessment of submissions, marking in their review if any of the guidelines are not completed to their satisfaction.

General Guidelines:	
√	1. Complete this HPP Data Interpretation Guidelines checklist and submit with your manuscript.
Abstract	2. Deposit all MS proteomics data (DDA, DIA, SRM), including analysis reference files (search database, spectral library), to a ProteomeXchange repository as a complete submission. Provide the PXD identifier(s) in the manuscript abstract and reviewer login credentials.
Pg 5	3. Use the most recent version of the neXtProt reference proteome for all informatics analyses, particularly with respect to potential missing proteins.
Pg 8	4. Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels.
Supp Table 2	5. Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected true positives and false positives at each level.
	6. Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR.
	7. Recognize that the protein-level FDR is an estimate based on several imperfect assumptions, and present the FDR with appropriate precision.
	8. Acknowledge that not all proteins surviving the threshold are "confidently identified".
	9. If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result.
Guidelines for extraordinary detection claims (e.g., missing proteins, novel coding elements)	
	10. Present "extraordinary detection claims" based on DDA mass spectrometry with high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra.
	11. Consider alternate explanations of PSMs that appear to indicate extraordinary results.
	12. Present high mass-accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the

Clarification for DIA/SWATH

If DIA/SWATH is analyzed like SRM (trace extraction via spectral libraries):

- SRM guidelines apply
  - Display traces
  - Heavy labeled reference peptides to demonstrate coelution and intensity pattern match
  - Need spectral library

If DIA/SWATH is analyzed like shotgun (DIA-Umpire, DISCO):

- Shotgun guidelines apply
  - Display spectra
  - Heavy labeled reference peptides to demonstrate intensity pattern match

Allow “candidate detections”?

- There is still widespread reporting of evidence that does not quite meet the guidelines
- Do we allow this labelled as “candidate detections”?

Exceptions to 2 peptide rule?

- Current rules require 2 distinct non-nested ≥9 AA peptides
- But guidelines allow: When only weaker evidence can be obtained, “justify that other peptides cannot be expected”
- This needs clarification

Exceptions to 2 peptide rule?

- Example: Q8N688
- DEFB123 Q8N688-1 Beta-defensin 123 isoform Iso 1 PE=2
- Options for trypsin are bleak. But missed cleavages would be okay..

Sequence Display Mode: Trypsin

MK LLLLTLTVLLLLSQLTPGGTQR CWNLYGK CR YR CSK K ER VYVYCINNK MCCVKPK YQPK ER WWPF

Annotated Variants from Swiss Prot (see neXtProt annotations)

Type	Num	Start	End	Info
Signal	1	1	20	
Chain	1	21	67	Beta-defensin 123

Annotations in Sequence Context:

```
                    10        20        30        40        50        60
           ---------|---------|---------|---------|---------|---------|------
  Primary: MKLLLLTLTVLLLLSQLTPGGTQRCWNLYGKCRYRCSKKERVYVYCINNKMCCVKPKYQPKERWWPF
 Signal_1: MKLLLLTLTVLLLLSQLTPG-------------------------------------------------
  Chain_1: -------------------GTQRCWNLYGKCRYRCSKKERVYVYCINNKMCCVKPKYQPKERWWPF
TrypticSites: -K-------------------R------K-R-R--KK-R--------K------K---K-R----
```

Exceptions to 2 peptide rule?

- Example: Q8N688
- DEFB123 Q8N688-1 Beta-defensin 123 isoform Iso 1 PE=2
- Or what about chymotrypsin?

Sequence Display Mode: Chymotrypsin

MKLLLLTLTVLLLLSQLTPGGTQRCW NLY GKCRY RCSKKERVY VY CINNKMCCVKPKY QPKERW W PF

Annotated Variants from Swiss Prot (see neXtProt annotations)

Type	Num	Start	End	Info
Signal	1	1	20	
Chain	1	21	67	Beta-defensin 123

Annotations in Sequence Context:

```
                    10        20        30        40        50        60
           ---------|---------|---------|---------|---------|---------|-------
Primary  : MKLLLLTLTVLLLLSQLTPGGTQRCWNLYGKCRYRCSKKERVYVYCINNKMCCVKPKYQPKERWWPF
Signal_1 : MKLLLLTLTVLLLLSQLTPG----------------------------------------------
Chain_1  : -------------------GTQRCWNLYGKCRYRCSKKERVYVYCINNKMCCVKPKYQPKERWWPF
```

Exceptions to 2 peptide rule?

- Example: Q8N688
- DEFB123 Q8N688-1 Beta-defensin 123 isoform Iso 1 PE=2

- Manually declare this one good enough?
- Or hold out for better evidence that meets the rules?

- Opinions will surely vary
- Who will decide?

Exceptions to 2 peptide rule?

- Example: P0DMU9
- CT45A10 P0DMU9-1 Cancer/testis antigen family 45 member A10 isoform Iso 1 PE=2
- Part of a big family for identical and near-identical proteins
- Do we insist on the 2 peptide rule, although it is hard

Exceptions to 2 peptide rule?

- Example: P0DMU9
- CT45A10 P0DMU9-1 Cancer/testis antigen family 45 member A10 isoform Iso 1 PE=2
- Part of a big family for identical and near-identical proteins
- Do we insist on the 2 peptide rule, although it is hard

- Opinions will surely vary
- Who will decide?

```
P0DMU8  : MTDKTEKVAVDPETVFKRPRECDSPSYQKRQRMALLARKQGAGDSLIAGSAMSKEKKLMTGHAIPPSQLDSQIDDFTGFSKDGMMQ
P0DMV1  : MTDKTEKVAVDPETVFKRPRECDSPSYQKRQRMALLARKQGAGDSLIAGSAMSKEKKLMTGHAIPPSQLDSQIDDFTGFSKDRMMQ
Q5HYN5  : MTDKTEKVAVDPETVFKRPRECDSPSYQKRQRMALLARKQGAGDSLIAGSAMSKAKKLMTGHAIPPSQLDSQIDDFTGFSKDRMMQ
P0DMU9  : MTDKTEKVAVDPETVFKRPRECDSPSYQKRQRMALLARKQGAGDSLIAGSAMSKEKKLMTGHAIPPSQLDSQIDDFTGFSKDGMMQ
P0DMU6  : MTDKTEKVAVDPETVFKRPRECDSPSYQKRQRMALLARKQGAGDSLIAGSAMSKAKKLMTGHAIPPSQLDSQIDDFTGFSKDRMMQ
consensus: ************************************************** ************************** ***
```

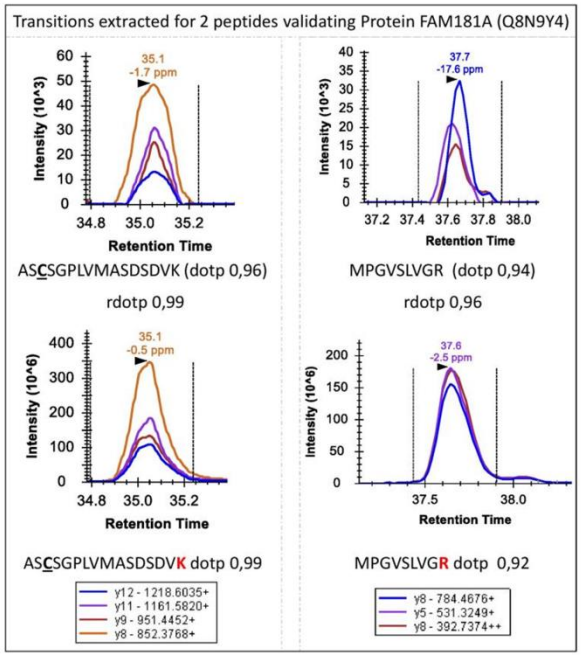
In addition to the existing HPP guidelines, new validation guidelines are needed for:

- Targeted MS data
- Ab-based data

(Others types of data at protein level ?)

Validating protein existence using targeted MS data

- What to capture? Data, metadata?
- How to represent the information?
- Quality filtering?

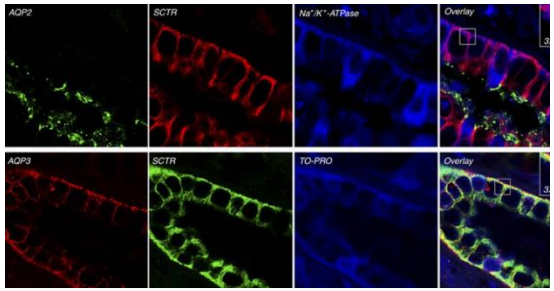


Carapito et al, *JPR* 2017
12 proteins from chr 2 and 14 confirmed by SRM, >= 2 peptides

We need to build a C-HPP approved pipeline

Validating protein existence using Ab-based data

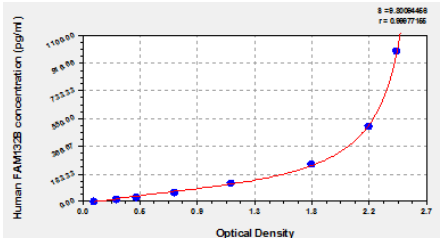
Immunohistochemistry

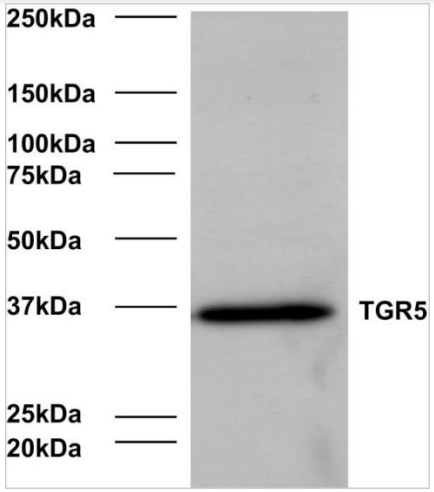


Immunocytochemistry

Ab-based data

ELISA



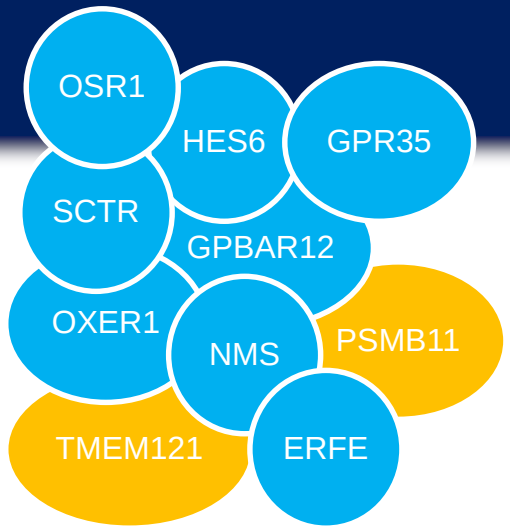


Western blot

Radioimmunoassay…

Antibody specificity has to be carefully checked, using siRNA, independent methods, consistency with RNA expression data, independent antibodies, competition with peptide against which the antibody was generated…

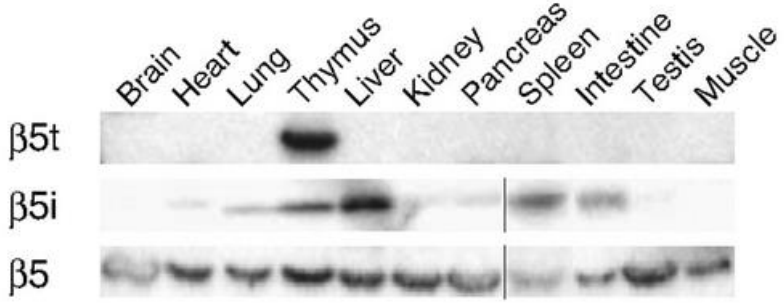
Papers not always/rarely perform those controls.



A new pipeline was beta-tested with

- 8 proteins from chromosome 2
- 2 from chromosome 14

Example : PSMB11



Expressed at protein level in thymus. Not detected in brain, heart, lung, liver, kidney, pancreas, spleen, lymph node, intestine, Peyer's patch, testis, muscle, skin. Expressed mainly in the thymus cortex (mainly in thymic epithelial cells and also in some cortical dendritic cells). Detected in neonatal and adult thymus cortex. • Gold

Evidence 1: western blot evidence neXtProt • Gold

Exclusive expression of proteasome subunit {beta}5t in the human thymic cortex.

Tomaru U., Ishizu A., Murata S., Miyatake Y., Suzuki S., Takahashi S., Kazamaki T., Ohara J., Baba T., Iwasaki S., Fugo K., Otsuka N., Tanaka K., Kasahara M.

Blood, **113**, 5186 - 5191 (2009) [Full text: 10.1182/blood-2008-11-187633] [PubMed: 19289856]

Acknowledgements

Knowledge Base pillar Committee:

Eric Deutsch (USA) (Chair)
Lydie Lane (Switzerland) (Co-Chair)
Henning Hermjakob (UK) (Co-Chair)
Kalle von Feilitzen (Sweden)
Seul-Ki Jeong (Korea)
Lennart Martens (Belgium)
Weimin Zhu (China)

Bioinformatics Hub Organizers:

Andy Jones (UK)
Henning Hermjakob (UK)
Lennart Martens (Belgium)
Juan Antonio Vizcaíno (UK)
Nuno Bandeira (USA)
Yasset Perez-Riverol (UK)
Mathias Walzer (Germany)
Yves Vandenbrouck (France)

Guidelines Development:

Gil Omenn (USA)
Chris Overall (Canada)
Robert Moritz (USA)
Jenny Van Eyk (USA)
Mark Baker (Australia)
Young-Ki Paik (Korea)
Susan Weintraub (USA)
Lydie Lane (Switzerland)
Lennart Martens (Belgium)
Yves Vandenbrouck (France)
Ulrike Kusebauch (USA)
William Hancock (USA)
Henning Hermjakob (UK)
Ruedi Aebersold (Switzerland)