



Human Proteome Organisation

The Human Proteome Project Launch

***Sydney HUPO2010
23 September 2010***

Agenda

Pierre Legrain

Where do we stand today ? What are we aiming at?

Ruedi Aebersold

MS-based HPP: Current developments and next phase for the MS pillar

Mathias Uhlén

Ab-based HPP: Long term objectives and 5 year milestones

Amos Bairoch

HPP Knowledge base: Building on present achievements

Eric Deutsch

In depth cross analyses of HUPO Initiatives collections

Young Ki Paik

Chromosome-based systematic proteome survey

Gil Omenn

General Discussion

The Human proteome Project Launch

Creation of the HPP Working Group

Decision taken by HUPO Executive committee in October 2009

The creation of an international working group for a “Human Proteome Project” (HPP) follows HUPO commitment to coordinate the efforts of gathering information, ideas and proposals and to work toward international consensus and a final proposal for a “Human Proteome Project”.

This mandate was voiced by the international scientific community, major scientific journals, industry representatives and funding agencies from around the world.

The Human Proteome Project is a huge undertaking. The Working Group built on the foundation of major progress in mass spectrometry, protein capture knowledge base and open sharing of proteome datasets.

The HPP Working Group

Building up the working group

Individually contacted scientists

Ruedi Aebersold, Amos Bairoch, Laura Beretta, Christoph Borchers, Eric Deutsch, Bill Hancock, Denis Hochstrasser, Gil Omenn, Young-Ki Paik, Salvatore Sechi, Mike Snyder, Sudhir Srivastava, Cathy Wu, Tadashi Yamamoto,

Additional key players

Kumar Bala, Patrik Kolar, Henry Rodriguez

HUPO past presidents and president-elect

Rolf Apweiler, John Bergeron, Cathy Costello, Sam Hanash,

Additional experts who contributed to the work

Steve Carr, Bruno Domon, Fuchu He, Mike Taussig, Mathias Uhlén

Three Workshops

Seattle January 25, 2010 - Kick Off meeting

A short history of the HPP vision was presented.

Key issues were addressed, such as updates on technologies and methodologies; quality of data; databases and data access; biological questions and biomedical challenges

**A HPP web page was created on HUPO web site
www.hupo.org/research/hpp**

Three Workshops

Montreal May 10, 2010

Emphasis on:

- the definition of terms, the need for standards and references for protein identification and characterization
- database systems
- the link between HPP and on-going large scale scientific initiatives
- coordination of national efforts.

Several subgroups constituted to address components of the HPP:

- mass-spectrometry-based proteomics for protein signatures
- antibody-based proteomics for protein expression/localization
- data format and interconnected databases
- HPP pilot projects.

Sydney September 19, 2010

Open discussion on the HPP

- Do we need it?
- Added value compared to present human proteomics?
- Is it possible?
- Can we capture important data
- from studies of other species?

Decision taken

Long term vision for the HPP

Key milestones at 5 years

How to make HPP desirable outside the proteomics community into the wider scientific community and beyond

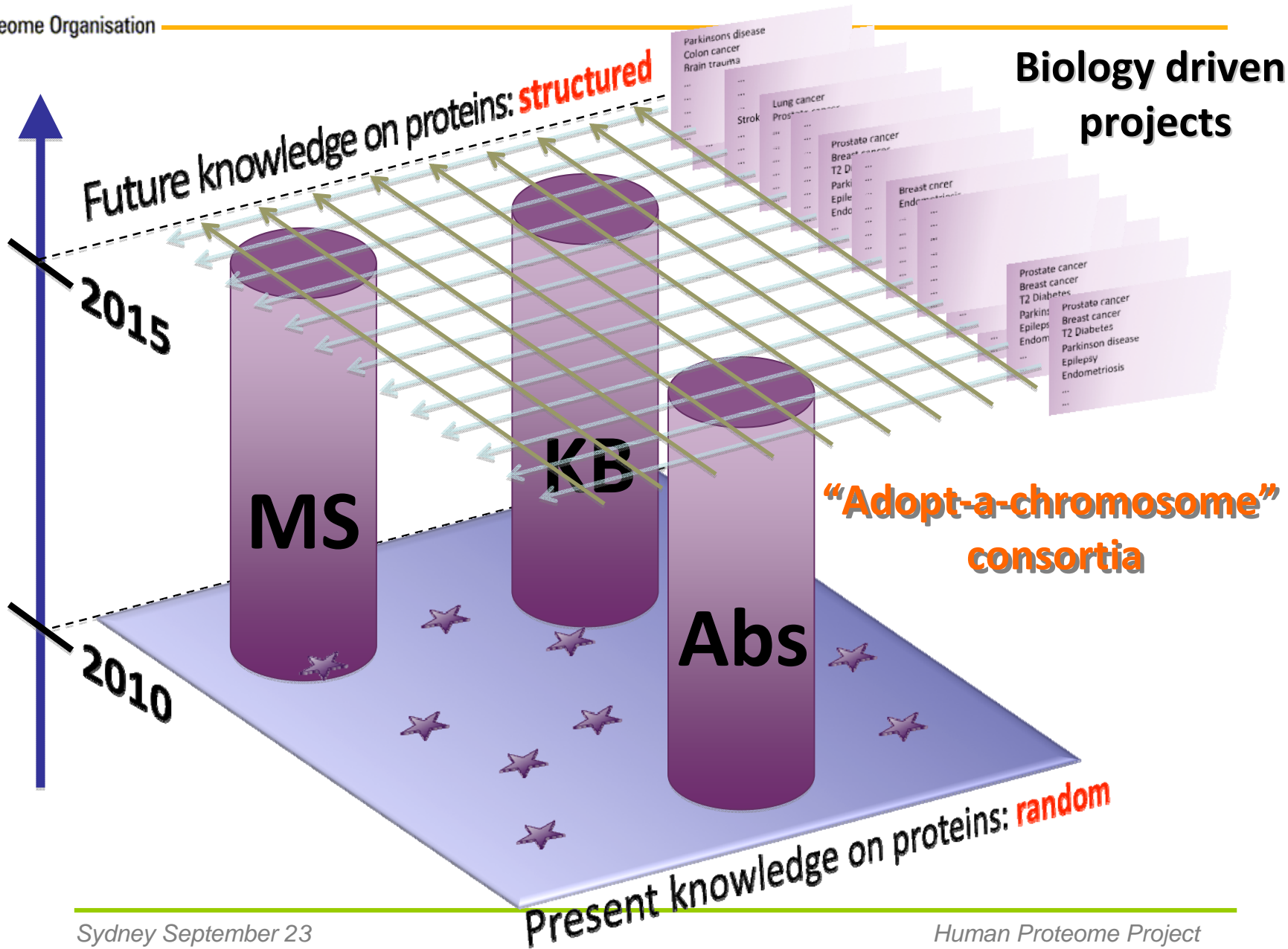


The Mission of the HPP

HPP will deliver a **comprehensive map** of the human proteins in their biological context.

HPP will **provide tools** for the scientific community that will allow each scientist to design experiments in a better way, as the Human Genome Project did.

HPP will **inspire**, beyond the scientific community, other stakeholders for diagnosis, prevention, therapy and cure of diseases and improved health worldwide.



MS-based HPP: Current developments and next phase for the MS pillar

Ruedi Aebersold

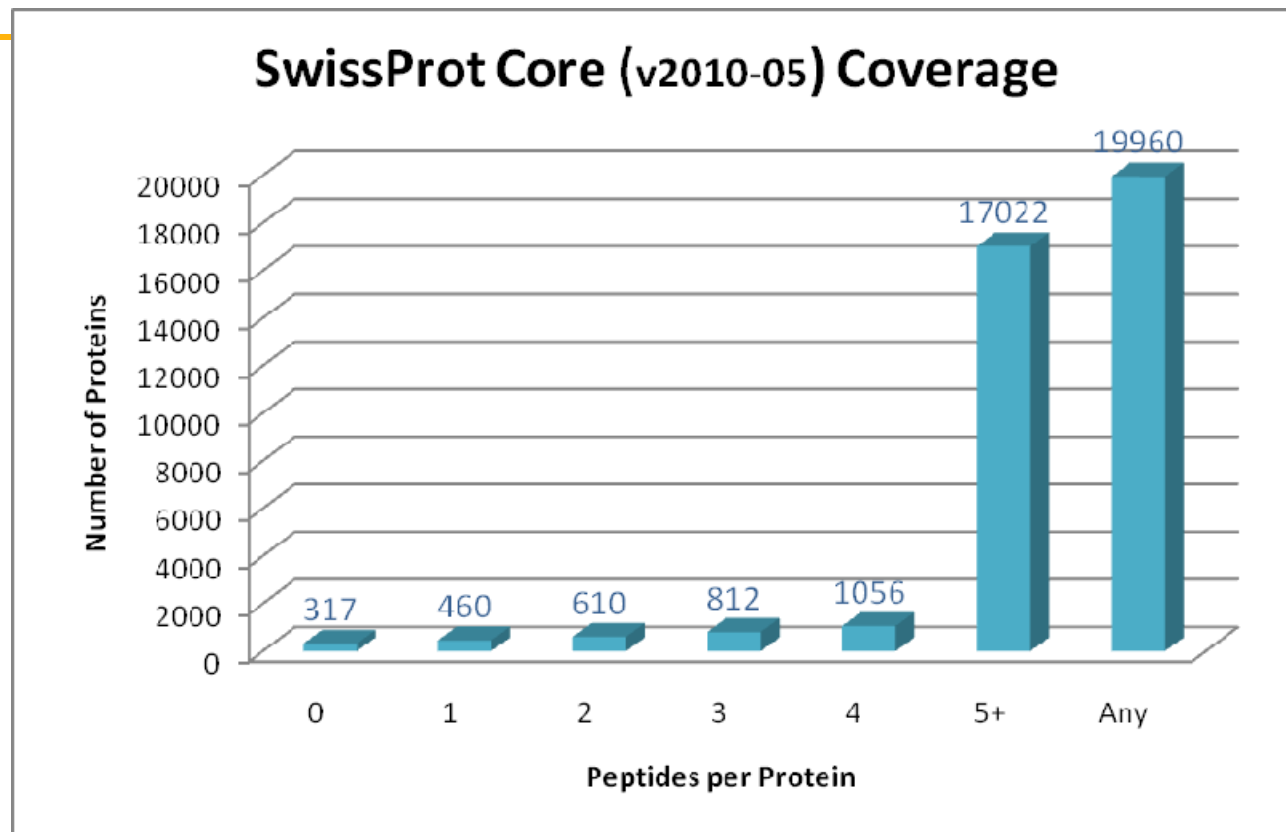
The Human Proteome Project HAS to ENABLE an infinite number of quantitative proteome measurements:

- affinity reagents and robust assays: antibodies**
- MS assays and robust assays: Selected reaction monitoring (SRM)**
- Informational resources (databases, programs)**

We report significant advance towards these stated goals:

- **Completion of reference spectra database for all yeast proteins**
- **Completion of reference spectra database for all (Uniprot) human proteins**
- **New technology to use the maps for quantitative proteomic measurement**

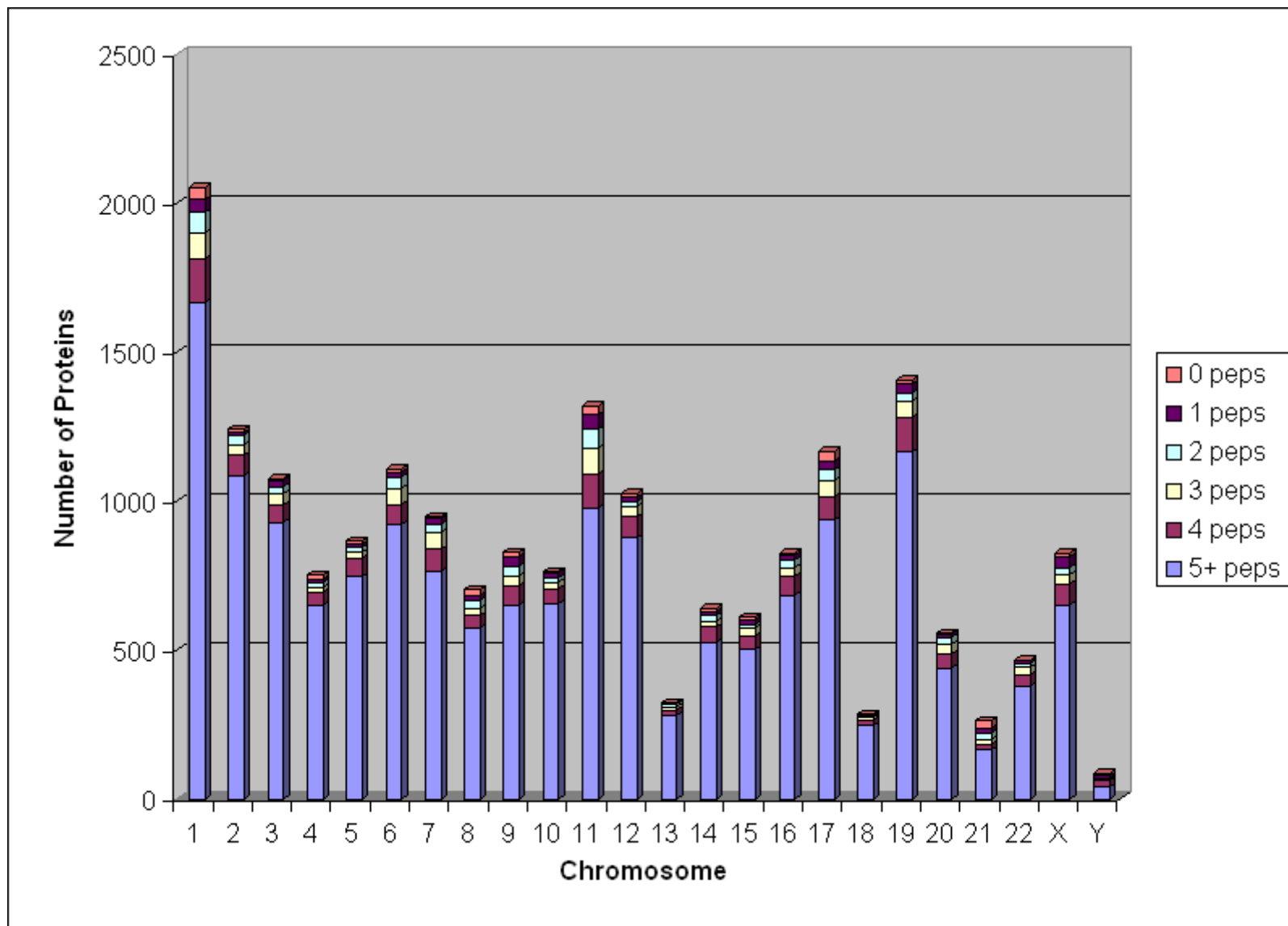
Human Proteome SRM Coverage

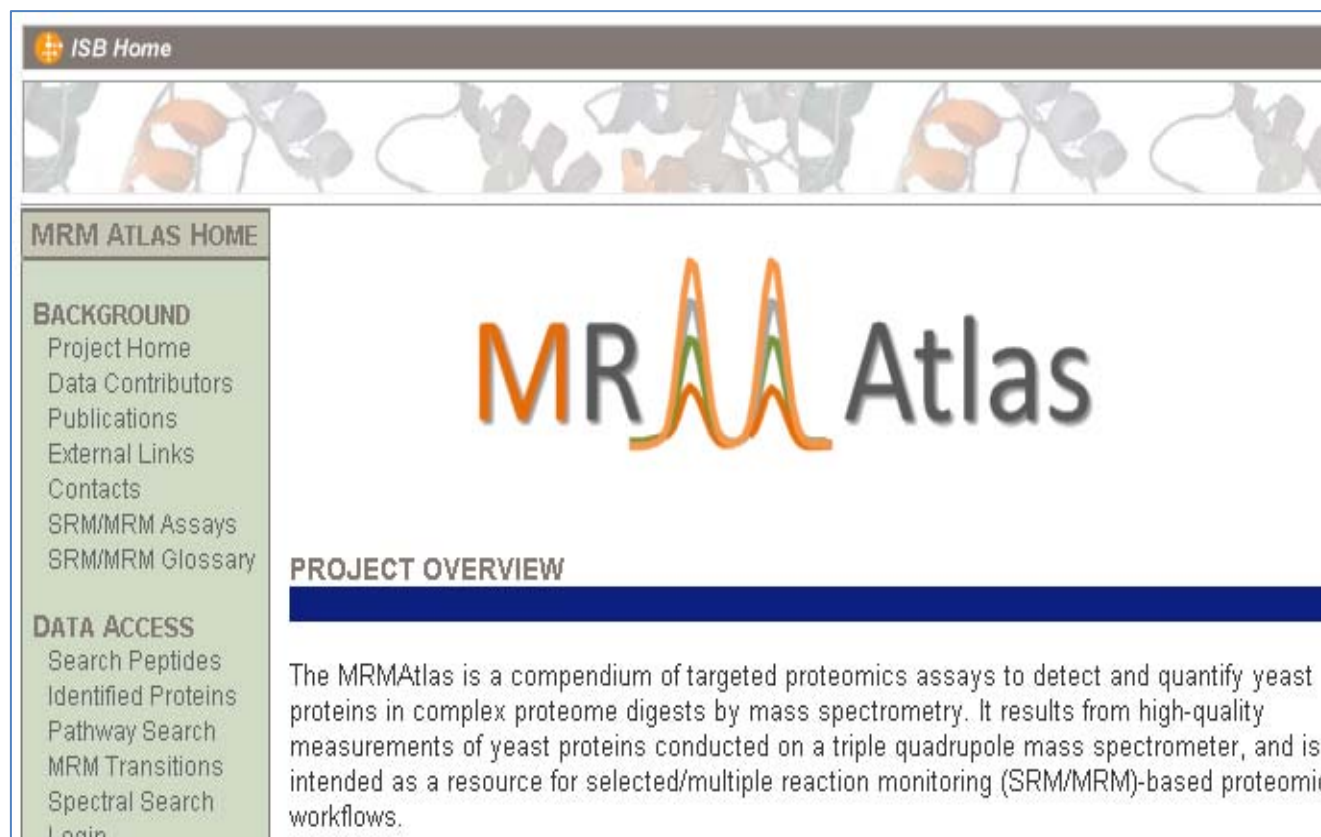


Coverage with peptides as of August 2010

- 170.000 peptides total
 - >10.000 N-glycosites (all transmembrane and secreted proteins)
 - 2726 SNPS with frequency >30%
 - >10.000 splice forms

Distribution of Peptide/Protein SRM coverage by chromosome





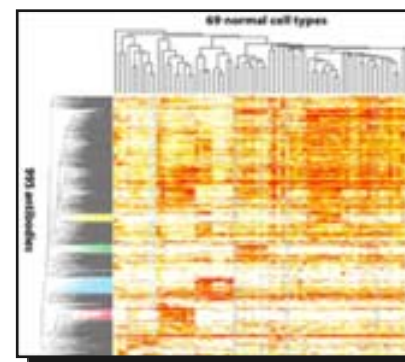
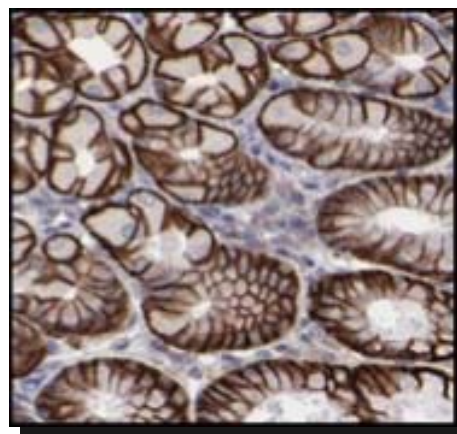
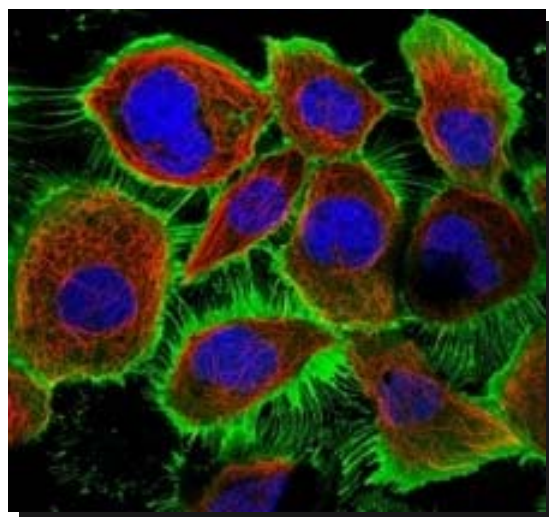
The screenshot shows the MRM Atlas website interface. At the top, there is a navigation bar with "ISB Home" and a decorative banner with molecular structures. Below the banner is a sidebar menu titled "MRM ATLAS HOME" containing sections for "BACKGROUND" (Project Home, Data Contributors, Publications, External Links, Contacts, SRM/MRM Assays, SRM/MRM Glossary) and "DATA ACCESS" (Search Peptides, Identified Proteins, Pathway Search, MRM Transitions, Spectral Search, Login). The main content area features the "MRM Atlas" logo, which includes a stylized mass spectrum. Below the logo is a "PROJECT OVERVIEW" section with a blue header bar and a paragraph of text: "The MRMatlas is a compendium of targeted proteomics assays to detect and quantify yeast proteins in complex proteome digests by mass spectrometry. It results from high-quality measurements of yeast proteins conducted on a triple quadrupole mass spectrometer, and is intended as a resource for selected/multiple reaction monitoring (SRM/MRM)-based proteomic workflows."

Picotti et al Nature Methods 2007
Picotti et al , Nature Methods, 2010

Next phase(s) of MS pillar

- **Make data from first phase publicly accessible**
- **Extend SRM library approach**
 - **PTM's**
 - **SNP's**
 - **Splices**
- **Generate database for registering SRM data**
- **Extend SRM resources to other species**
- **Create additional resources at Uniprot level**
 - **GFP libraries**
 - **Isotopic reference molecules**
- **Create links to antibody and informatics pillars**

Contribution to the HPP by antibody-based proteomics



Mathias Uhlen

A gene-centric proteome project

Map the human proteins (gene by gene):

- Molecular (isoforms)
- Subcellular (localization)
- Cells, tissues and organs (expression)
- Interaction networks
- Plasma/serum (abundance)

Use all available technology platforms:

- Mass spectrometry
- Antibodies (immunotechnologies)
- Gene fusions (GFP)

For each protein

Type of profile	Description	Technology platform
Molecular	Molecular weight, isoforms, modifications	WB, immuno-capture, MS
Subcellular	A localization index in selected cell lines	IF (confocal), GFP-fusions, organell-specific MS
Cell	Expression patterns across a panel of human cell lines	IF (confocal), IHC, whole-proteome MS, RNA expression
Tissue	Expression patterns across all major tissues and organs	IHC (tissue micro array), whole-proteome MS, RNA expression
Plasma/serum	Concentration of each protein across age and gender	Immunoassays, MS (targeted)

Antibody-based proteomics involved in all areas

All data available to the public (with no IPR restrictions)

For each protein

Type of profile	Description	Technology platform
Molecular	Change in isoforms	WB, immuno-capture, MS
Subcellular	Change in localization pattern	IF (confocal), GFP-fusions, organell-specific MS
Cell and tissues	Changed in quantitative expression pattern	IF (confocal), IHC, whole-proteome MS, RNA expression
Plasma/serum	Change in protein or isoform concentration	Immunoassays, MS

Disease specific projects

- Investigator-driven
- Possible to seek IPR-protection

Physical resource

Resource	Description	Comment
cDNAs	Full-length proteins (splice variants) and various gene fusions	Including over-expressed cell lysates
Antibodies	At least two antibodies with non-overlapping epitopes	Preferably monoclonals
Peptides	At least two independent synthetic peptides or recombinant PrESTs	Cleavable peptides shown to be detectable by MS
siRNA	At least two independent molecules	Validated by RT-PCR and antibodies

- **All resources available to the public (with no IPR restrictions)**
- **Decentralized resource with international coordination**
- **Academic and commercial partnerships**

Long Term Objectives of HPP (2020)

- **A complete atlas of expression and subcellular localization of proteins in human cells, tissues and organs**
- **Paired renewable antibodies to a representative protein from every human gene**
- **A mixture of monoclonal and recombinant binders**
- **A subcellular localization index covering all human genes and most isoforms**
- **Tissue profiles covering all major tissues and organs**
- **A plasma/serum atlas across gender and age (all major isoforms)**

- **At least one antibody to a representative protein from every gene**
- **Achieved by contribution from both academic and commercial providers**
- **Technology to generate recombinant binders in a systematic manner**
- **Validation of all “HPP antibodies” with new technologies (such as siRNA)**
- **A draft version of subcellular localization covering all human genes**
- **A draft version of tissue profiles covering all major tissues and organs**
- **A plasma/serum atlas across gender and age (including major isoforms)**
- **A partial “parts-list” of the major protein isoforms from every gene (achieved by combining immuno- and tag-capture with MS characterization)**

Building on present achievements

Amos Bairoch

Main point

- **As far as possible, HPP will use resources that already exist**
- **Some of these resources will be «extended» to cater with specific needs of HPP**
- **Specific developments will take place that will be targeted in integrating data and knowledge concerning human proteins**

Existing resources 1/2

- **Sequences (including PTMs, SAPs, splice isoforms, etc.): *UniProtKB/Swiss-Prot* human section**
- **For identification purpose, participating labs should make use of the proposed HUPO PSI Extended FASTA format (*PEFF*)**
- **Identification data should be deposited in one of the proteomics repositories (*PeptideAtlas*, *PRIDE*, *Peptidome*, etc.) that abides to the standards that are being developed in the framework of the EU-funded *ProteomeExchange* program**

- **Quantitative proteomics: SRM spectra based on synthetic peptides will be stored in the *SRMAtlas***
- **Antibodies: Human Proteomics Atlas (*HPA*) and *Antibodypedia***

Almost existing resource!

- **neXtProt (beta.nextprot.org) is being developed at SIB with the goal of providing an integrative knowledge platform on human proteins**
- **In term of proteomics-derived information, it already includes:**
 - **PTMs from many large-scale and targeted identification studies**
 - **Tissue and organ expression data from the Human Protein Atlas**
- **It should soon (beginning of 2011) include deep integrative links to the reference spectra in SRMAtlas.**

- Protein
- Function
- Medical
- Expression
- Interactions
- Localisation

- Proteomics
- Structures
- Identifiers

- Gene
- Exons
- Identifiers
- References
- Publications
- Patents
- Submissions
- Web resources

- Isoforms (2)**
- (de)select all
- Iso 1
- Iso 2
- Apply selection

FNDC3A » Fibronectin type-III domain-containing protein 3A

[★ favorite](#) [🏷 label](#)

Protein also known as: Human gene expressed in odontoblasts .

Gene name: FNDC3A .

Family name: **FNDC3**

One or more isoforms of this protein have been shown to exist at protein level

[extend overview](#) **1** **15** **2**
GENE REF ISO

Positional Annotations referenced on Iso 1

Isoform **Iso 1** 1198 aa, Mass: 131852 Da, pI: 6.29

Proteomics Topology Domains/regions Modified residues Variants All/None Actions: [FASTA](#) , [Blast](#): [full sequence](#) [on selection](#)



```

51  LSPQQLTAE QYVDSSES STEELVES LGEYLDNS YANVYVQA
101 PEFHPGSHV LHRSPHPLP GFIPVPTMP PPRHMYSV TGAGDMITQY
151 MPQYQSSQVY GDVDAHSTHG RSNFRDERSS KTYERLQKKL KDRQGTQKDK
201 MSSPPSPQK CPSPINEHNG LIKQIAGGI NTGSAKIKSG KKGKGTQVDT
251 EIEEKDEETK AFEALLSNIV KPVASDIQAR TVVLTWSPPS SLINGETDES
301 SVPELYGYEV LISSTGKDGK YKSVYVGEET NITLNDLKPA MDYHAKVQAE
351 YNSIKGTPSE AEIFTTSLCE PDIPNPPRIA NRTKNSLTQ WKAPSDNGSK
401 IQNFVLEWDE GKGNGEFCQC YMGSQKPKI TKLSPAMGCK FRLSARNDYG
451 TSGFSEEVLV YTSGCAPSMP ASPVLTKAGI TWLSLQWSKP SGTSPDEGIS
501 YILEMEEETS GYGFKPKYDG EDLAYTVKNL RRSTKYKPKV IAYNSEGKSN
551 PSEVVEFTTC PDKPGIPVKP SVRGIHSHS FKITWDPKDK NGGATINKYV
601 VEMAEGSNGN KWEMIYSGAT REHLCDRLNP GCFYRLRVYC ISDGGQSAVS
651 ESLLVQTPAV PPGPCLPRL QGRPKAKEIQ LRWGPPVDG GSPISCYSVE
701 MSPIEKDEPR EVYQGSEVEK TVSSLLPGKT YSFRRLAANK MGFPGPSEKC
751 DITTAPGPPD QCKPPQVTCR SATCAQVNWV VPLSNGTDVT EYRLEWGGVE
801 GSMQICYCGP GLSYEIKGLS PATTYICRVQ ALSVVGAGPF SEVVACVTFP
851 SVPGIVTCLQ EISDDEIENP HYSPTCLAI SWEKPCDHGS EILAYSIDFG
901 DKQSLTVGKV TSYIINNLQP DTTYRIRIQA LNSLGAGPFS HMIKLTQKPL
951
    
```

[hide graphical display](#)

Category	Names	Positions	Length	Description	Evidences	Also present in isoforms
PROTEOMICS	Antibody	142 - 274	133	HPA008927 (HPA ↗)		2
	Peptide	224 - 236	13	Liver HUPO Project (PeptideAtlas ↗)		2
	Peptide	224 - 236	13	SRM (SRMAtlas ↗)		2
	Peptide	261 - 271	11	Liver HUPO Project (PeptideAtlas ↗)		2
	Peptide	261 - 271	11	SRM (SRMAtlas ↗)		2
	Peptide	518 - 528	11	Liver HUPO Project (PeptideAtlas ↗)		2
	Peptide	599 - 610	12	SRM (SRMAtlas ↗)		2
	Peptide	628 - 635	8	SRM (SRMAtlas ↗)		2
	Antibody	776 - 910	135	HPA012825 (HPA ↗)		2
	Peptide	957 - 968	12	SRM (SRMAtlas ↗)		2

The HPP Portal

- **A web portal will be developed that aims to be the central focal point of HPP in terms of publicizing the project, its goals and its results**
- **It should be populated by appropriate documents, time line, and links to the participating laboratories, funding agencies and major proteomic resources and initiatives**
- **It should provide deep links to the data and knowledge resources that are described in the preceding slides. One should avoid duplication of such data storage and the HPP portal should be as “primary data” light as possible**
- **It should also provide access to proteomics identification and analysis tools**
- **To engage the participation of the broad scientific community, the design and development of the portal should be driven by user requirements..**

Cross-Analysis of HUPO Initiative Data Collections

**Eric Deutsch, Terry Farrah, Laura Beretta,
Tadashi Yamamoto, Gilbert Omenn**

Comparing 4 Human Data Collections

Liver – L. Beretta, Fred Hutchinson Cancer Research Center, contribution to the HUPO Liver Proteome Project

Plasma – G. Omenn, HUPO Plasma Proteome Project, Univ. of Michigan and numerous investigators

Kidney – T. Yamamoto, Niigata University, Japan and HUPO Kidney-Urine Proteome Project investigators

Urine – T. Yamamoto; Y-A Goo and P. von Haller, Univ. of Washington

Bioinformatics – Terry Farrah and Eric Deutsch, Institute for Systems Biology

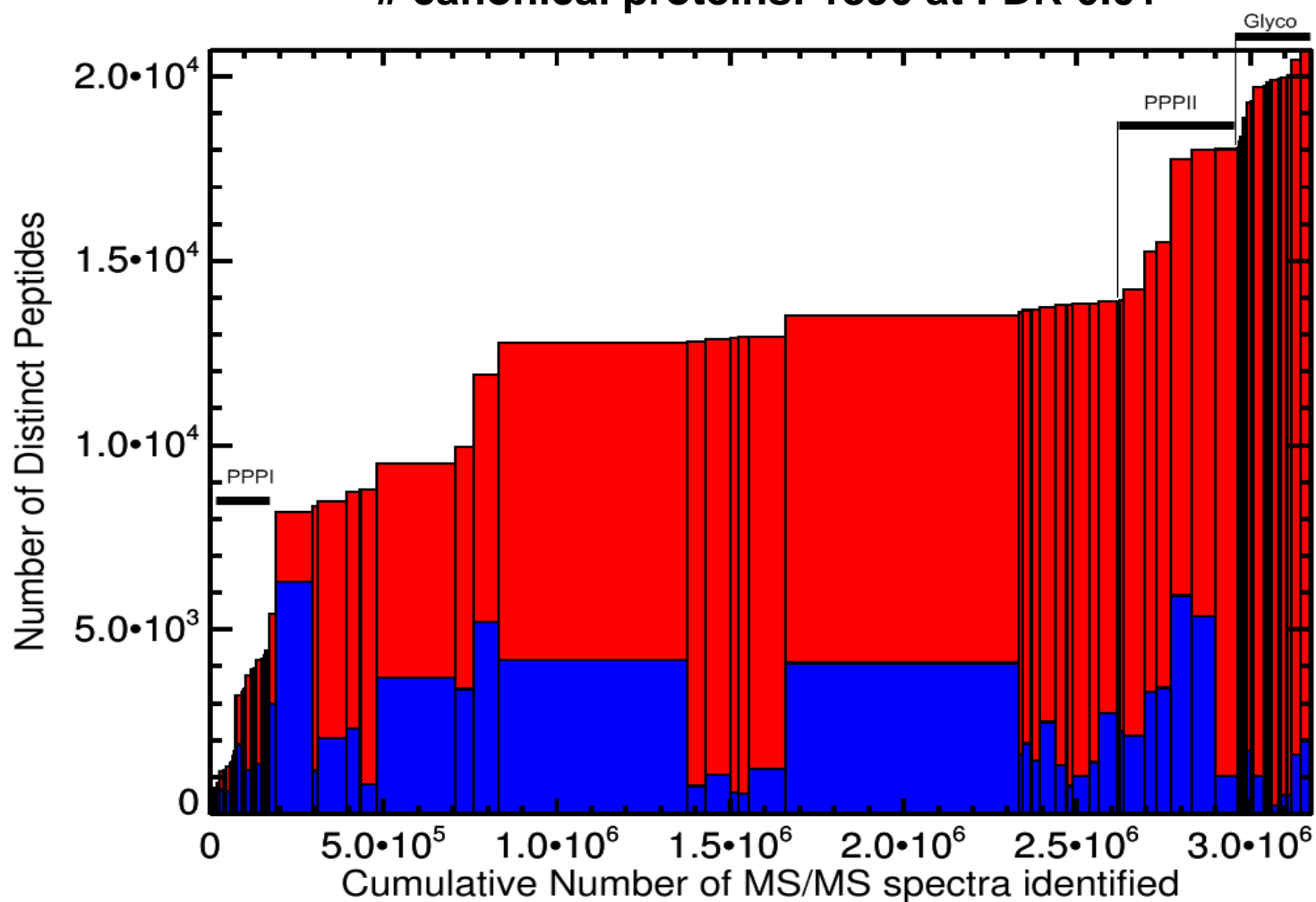
through PeptideAtlas

Plasma atlas – # expts: 91

peptide-spectrum matches: 3,172,759

distinct peptides: 20,679

canonical proteins: 1890 at FDR 0.01



Total counts and global FDRs in the 4 Atlases

<u>Atlas</u>	<u># expts</u>	<u># PSMs</u>	<u># distinct peptides</u>	<u># Canonical Proteins</u>	<u>PSM global FDR</u>	<u>peptide global FDR</u>	<u>protein global FDR</u>
liver	2	554711	44710	5940	0.03%	0.1%	1.0%
plasma	91	3172759	20679	1890	0.002%	0.1%	1.0%
kidney	8	314514	8101	1380	0.02%	0.2%	1.0%
urine	8	41128	6799	1259	0.07%	0.3%	1.0%

201 Proteins Common to all Atlases (excluding keratins and immunoglobulins)

		per mil of total PSMs for atlas				
		liver	plasma	kidney	urine	description
12 highest average normalized PSM counts	P69905	3.63	3.06	3.55	27.11	Hemoglobin subunit alpha
	P02753	0.12	27.16	0.05	0.78	Retinol-binding protein 4
	P02760	0.05	4.39	0.04	15.71	Protein AMBP
	P08670	2.19	0.01	17.09	0.15	Vimentin
	P02675	0.15	12.67	0.23	3.45	Fibrinogen beta chain
	P02042	3.04	1.95	2.46	5.62	Hemoglobin subunit delta
	P02679	0.20	9.54	0.16	1.63	Fibrinogen gamma chain
	P05090	0.01	6.35	<0.01	4.33	Apolipoprotein D
	P41222	0.01	1.20	0.02	8.61	Prostaglandin-H2 D-isomerase
	P10909	0.27	3.27	0.45	2.53	Clusterin
	P13987	0.01	0.16	0.13	5.52	CD59 glycoprotein
	P04004	0.08	3.40	0.91	1.26	Vitronectin
12 lowest average normalized PSM counts	P15121	0.013	0.007	0.003	0.073	Aldose reductase
	Q96G03	0.027	<0.001	0.019	0.049	Phosphoglucomutase-2
	P00492	0.025	0.002	0.016	0.049	Hypoxanthine-guanine phosphoribosyltransferase
	O15144	0.038	0.003	0.025	0.024	Actin-related protein 2/3 complex subunit 2
	O75608	0.018	0.001	0.013	0.049	Acyl-protein thioesterase 1
	P48637	0.040	0.007	0.003	0.024	Glutathione synthetase
	Q14019	0.031	0.001	0.006	0.024	Coactosin-like protein
	Q15185	0.018	0.003	0.003	0.024	Prostaglandin E synthase
	Q9Y624	0.013	0.001	0.003	0.024	Junctional adhesion molecule A
	P07711	0.007	0.004	0.003	0.024	Cathepsin L1
	P07738	0.002	0.005	0.003	0.024	Bisphosphoglycerate mutase
	P06865	0.005	<0.001	0.003	0.024	Beta-hexosaminidase subunit alpha

Approximate abundances

Proteins Unique to Each Atlas

Canonical protein sequences (non-Ig, non-keratin) unique to each atlas relative to other three			
Atlas	total canonicals in atlas	# Unique (% total)	per cent of unique that are secreted *
liver	5897	3619 (61%)	4.8%
plasma	1879	349 (19%)	67%
kidney	1321	137 (9%)	20%
urine	1189	147 (12%)	47%

* Secreted according to annotation in Swiss-Prot, which may be either experimentally determined or predicted.

- **Data generated on different platforms can be compared after collective analysis with the same rigorous bioinformatics tools**
- **Collective analysis of large datasets improves the quality of individual analyses via spectral library creation/search**
- **Creates a baseline compendium of MS/MS observations to begin the Human Proteome Project**

Chromosome-based proteome survey

Young-Ki Paik

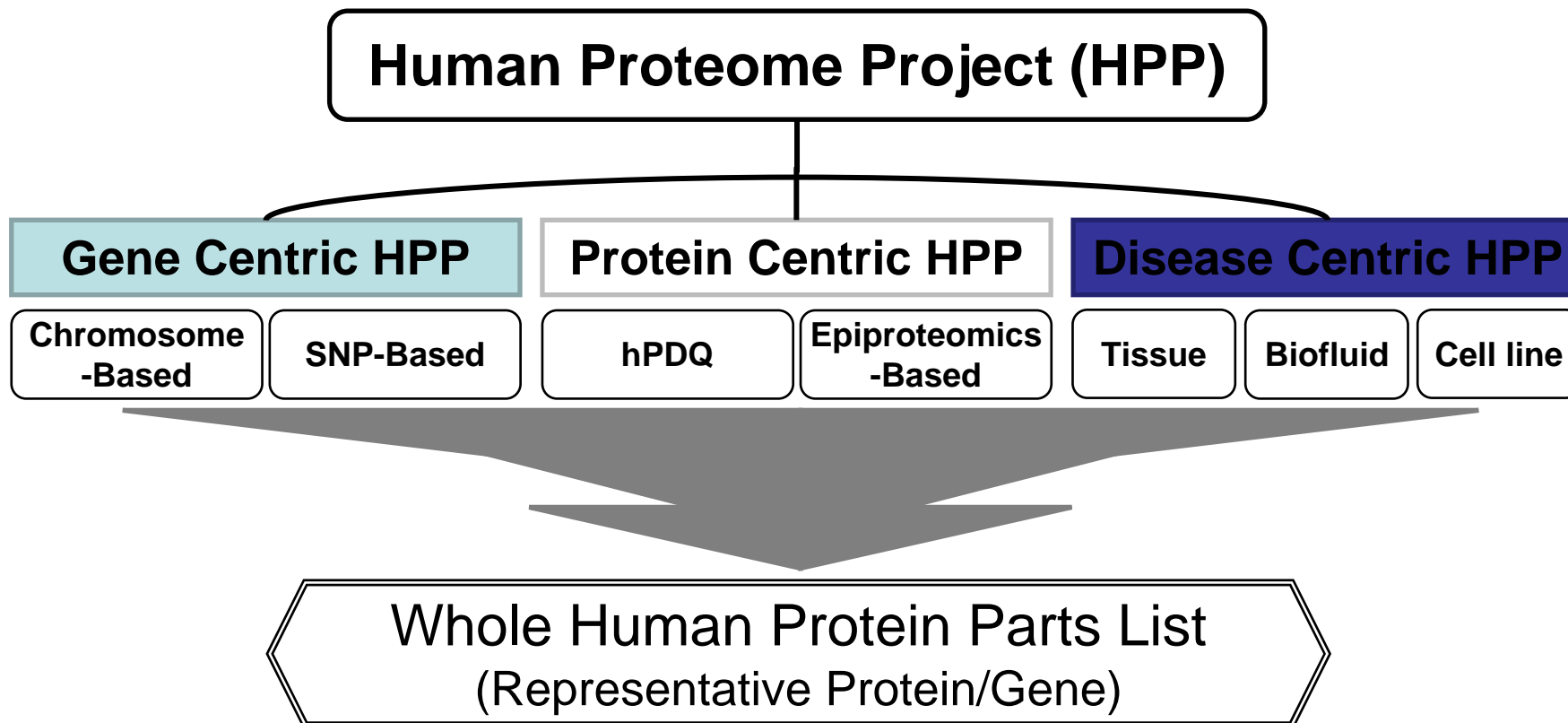
**On Behalf of 9 Chromosome-based
HPP Working Groups**

HGP vs. HPP

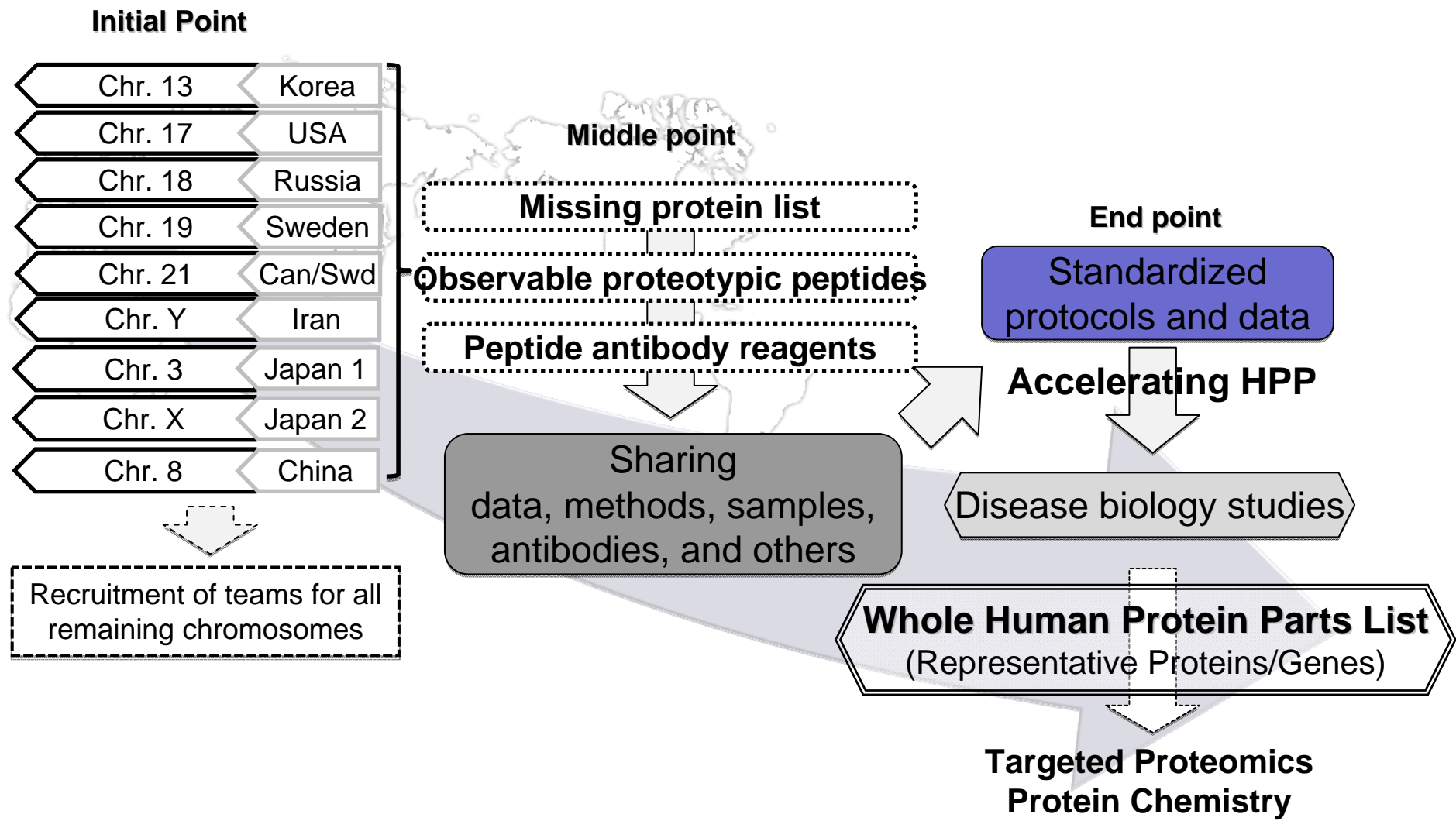
	Human Genome Project*	Human Proteome Project
Major Achievement	<ul style="list-style-type: none"> I. Genetic map II. DNA sequence III. Gene model IV. SNP/EST 	<ul style="list-style-type: none"> I. Proteome map/Interaction networks II. Representative proteins III. Confirmed/refined gene model IV. PTMs
Technologies Improved	<ul style="list-style-type: none"> I. Sequencing technologies II. Genome informatics III. Capillary electrophoresis 	<ul style="list-style-type: none"> I. MS, MS/MS technologies II. Proteome informatics III. Separation techniques IV. PTM analysis methods
Byproducts	<ul style="list-style-type: none"> I. High-throughput oligo-nucleotide synthesis II. cDNA libraries III. DNA microarray IV. Scale-up of two-hybrid mapping 	<ul style="list-style-type: none"> I. High-throughput oligo-peptide synthesis II. Standard proteins III. Antibody libraries IV. Profile/Spectral libraries V. Protein microarray / Antibody array VI. IP, Y2H, IHC... VII. High-throughput PPI
Derived Fields	<ul style="list-style-type: none"> I. Comparative genomics II. Functional genomics III. Genome dynamics IV. Epigenomics <p>Transcriptome/Spliceome Structurome Interactome Reactome/Metabolom Proteome</p>	<ul style="list-style-type: none"> I. Comparative proteomics II. Functional proteomics III. Proteome dynamics IV. Epiproteomics V. Proteogenomics <p>Transcriptome/Spliceome Structurome Interactome Reactome/Metabolom Proteome</p>
Personalized study	<ul style="list-style-type: none"> I. Each person's genetic profile (profile contains ¹gene expression level and ²SNP or mutation) 	<ul style="list-style-type: none"> I. Each person's proteomic profile II. Tissue/organ specific proteomic profile III. Normal/disease specific proteomic profile IV. Age/Developmental stage specific proteomic profile (profile contains ¹protein expression level, ²PTM change, ³interaction partner change, ⁴splicing variant, ⁵isoforms and also ⁶cSNP or mutation)

* Collins FS, Morgan M, Patrinos A., Science. 2003 Apr 11;300(5617):286-90

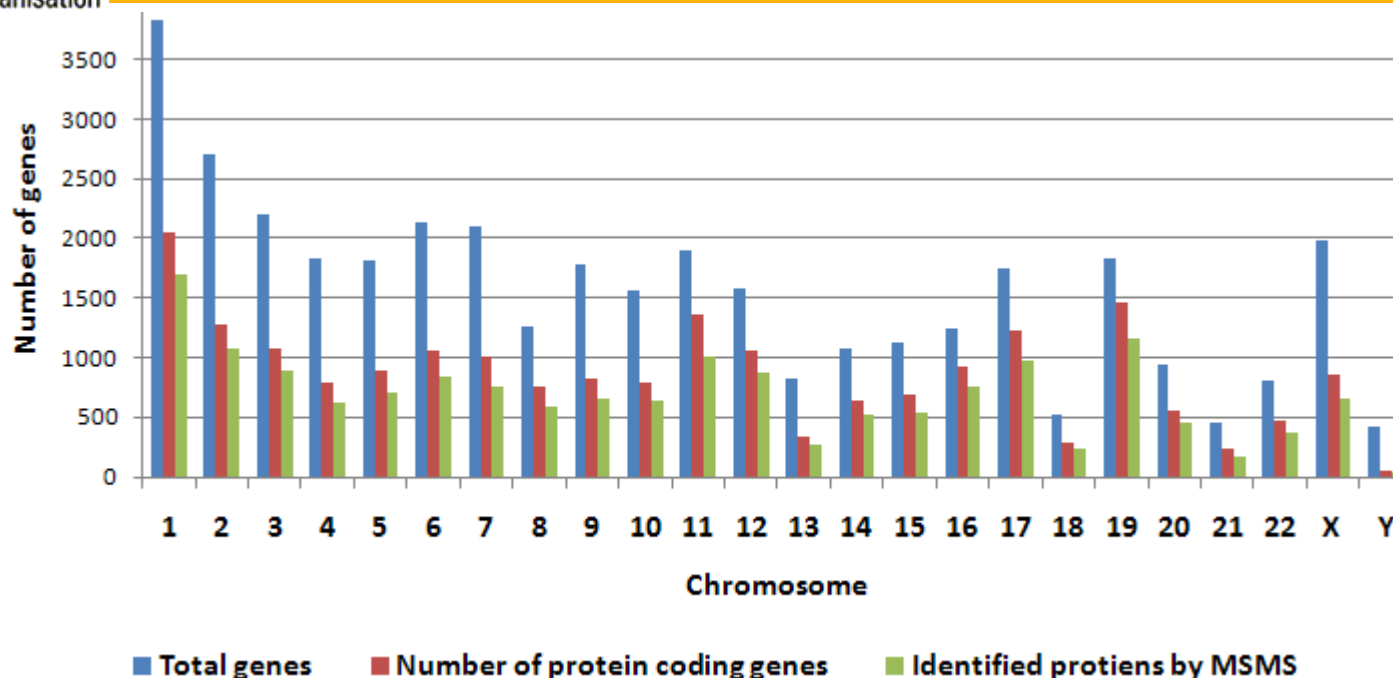
Big picture of HPP



Chromosome-based HPP Pipeline



Current status



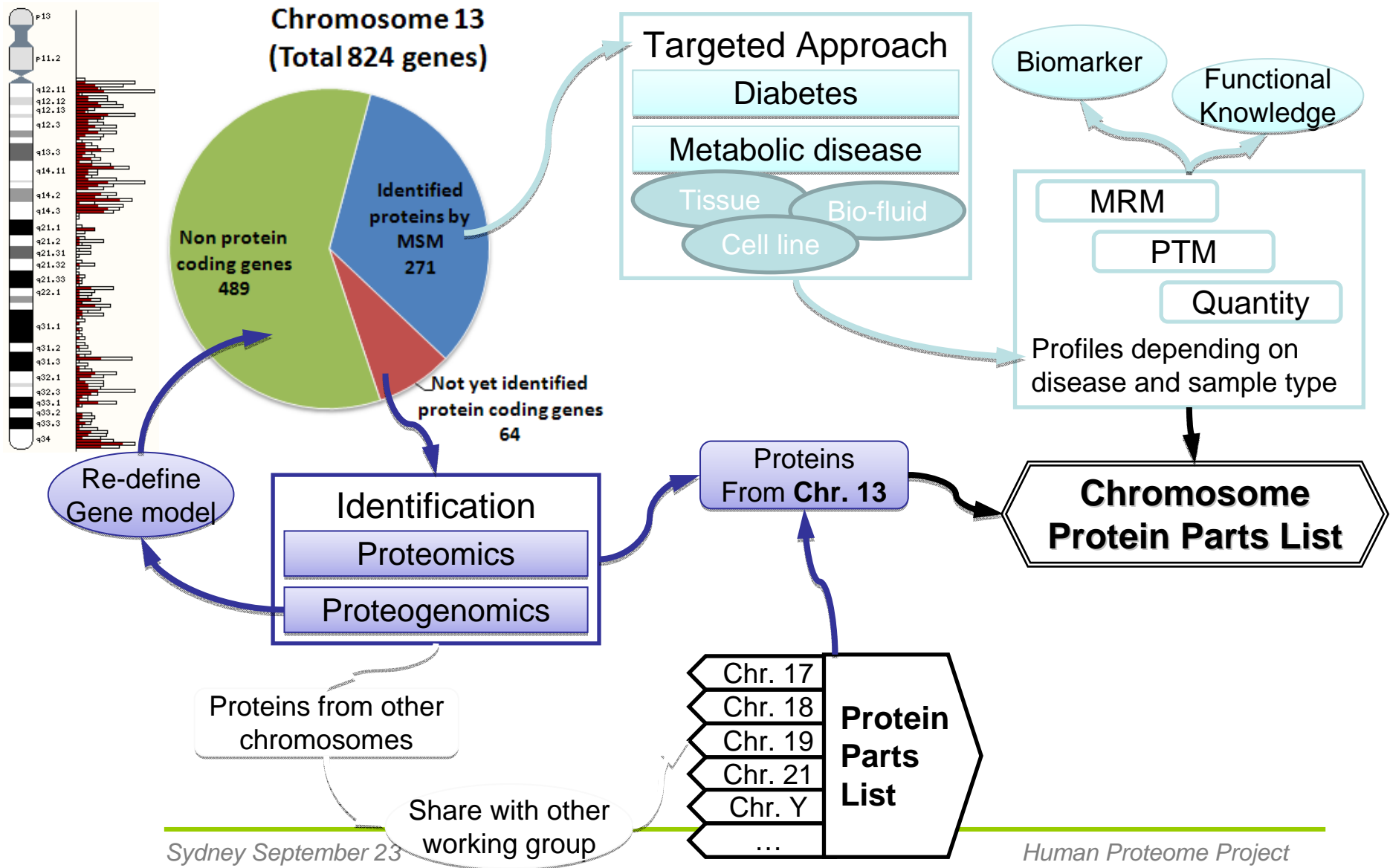
Gene types	Chromosome																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Pseudogene	1099	887	688	676	586	722	740	209	696	486	249	183	316	133	116	91	249	63	164	201	141	232	761	315
miRNA	134	115	99	92	83	81	90	80	69	64	63	72	42	92	78	52	61	32	110	57	16	31	128	15
rRNA	66	40	29	24	25	26	24	28	19	32	24	27	16	10	13	32	15	13	13	15	5	5	22	7
snRNA	221	161	138	120	106	111	90	86	66	87	74	106	45	65	63	53	80	51	29	46	21	23	85	17
snoRNA	145	117	87	56	61	73	76	52	51	56	76	62	34	97	136	58	71	36	31	37	19	23	64	3
miscRNA	106	93	77	71	68	67	70	42	55	56	53	69	36	46	39	34	46	25	15	34	8	23	52	2
Protein coding	2050	1283	1078	788	889	1053	1008	762	817	787	1358	1056	335	640	687	922	1220	294	1467	559	242	469	862	60
Total	3821	2696	2196	1827	1818	2133	2098	1259	1773	1568	1897	1575	824	1083	1132	1242	1742	514	1829	949	452	806	1974	419
GPMdb	1704	1075	884	623	707	846	750	581	661	635	1014	878	271	516	541	751	981	238	1162	453	175	368	657	36
%	83.1	83.8	82	79.1	79.5	80.3	74.4	76.2	80.9	80.7	74.7	83.1	80.9	80.6	78.7	81.5	80.4	81	79.2	81	72.3	78.5	76.2	60

Number of genes for each chromosome, protein coding genes and proteins identified.

(Depending on Ensemble, May 2010 (v59) and GPM db)

Extracted From GPM DB

Strategy of Chr-based HPP working group (e.g., Chr 13)





Human Proteome Organisation

Gil Omenn

General Discussion

"Protein Maps and Interactions in Human Body"

- **It is the time to start the "Human Proteome Project" by international and national collaborations.**
- **HUPO encourages/urges each national public funding agency to identify their preferred pathway to support aspects of this extraordinary coordinated Project.**