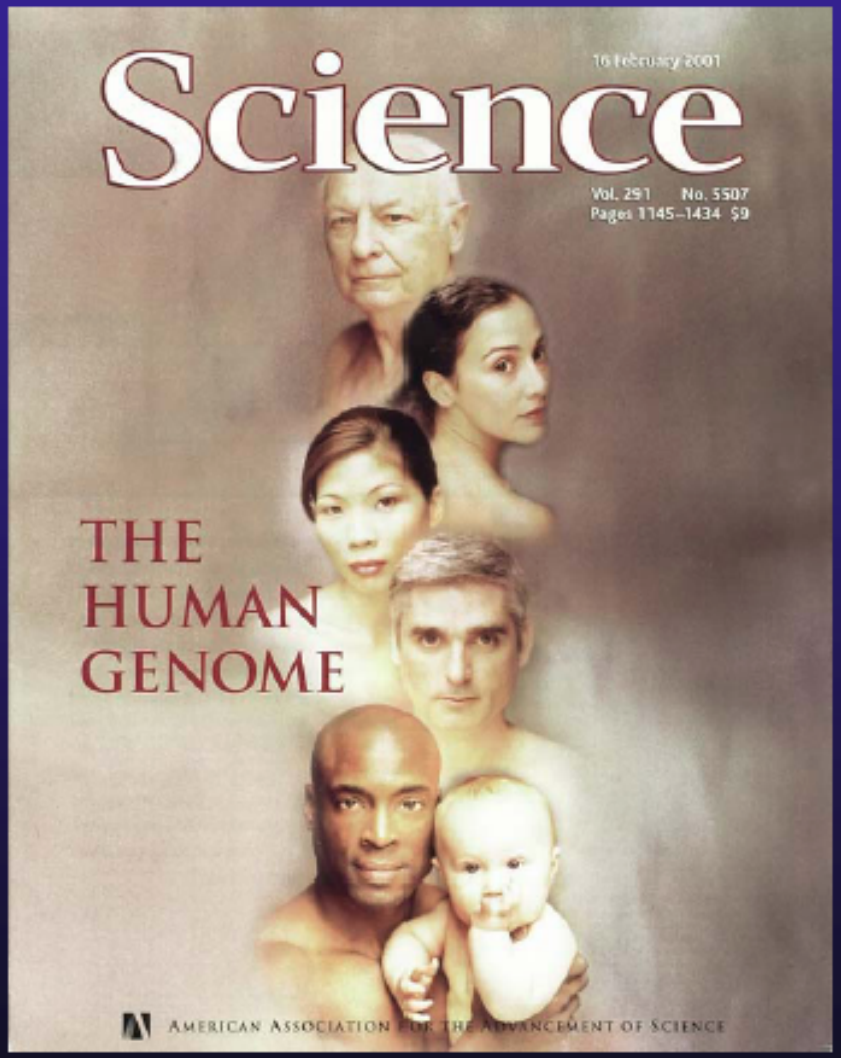
Tracking the Human Serum/Plasma Proteome: The Plasma PeptideAtlas and Contributions to the Emerging Human Proteome Project

Gilbert S. Omenn, M.D., Ph.D.

Director, Center for Computational Medicine and Bioinformatics
University of Michigan
HUPO Vice-President, Initiatives Chair, US HUPO President

British Society for Proteome Research
EBI/Hinxton, 13July, 2011

Near-Completion of Human Genome Sequence, Feb 2001



FINANCIAL TIMES WEDNESDAY FEBRUARY 21 2001

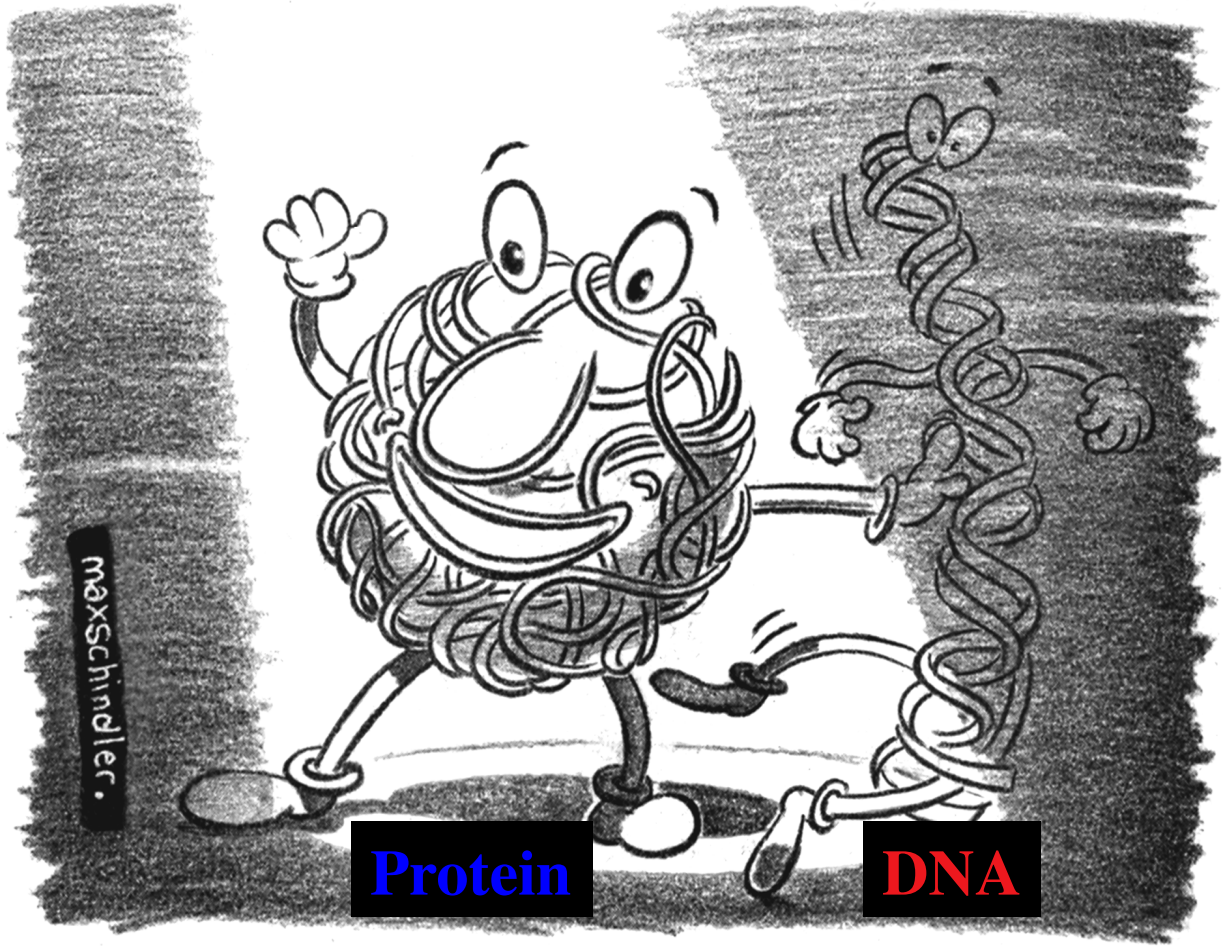
INSIDE TRACK

Strategy, Management, Technology & People

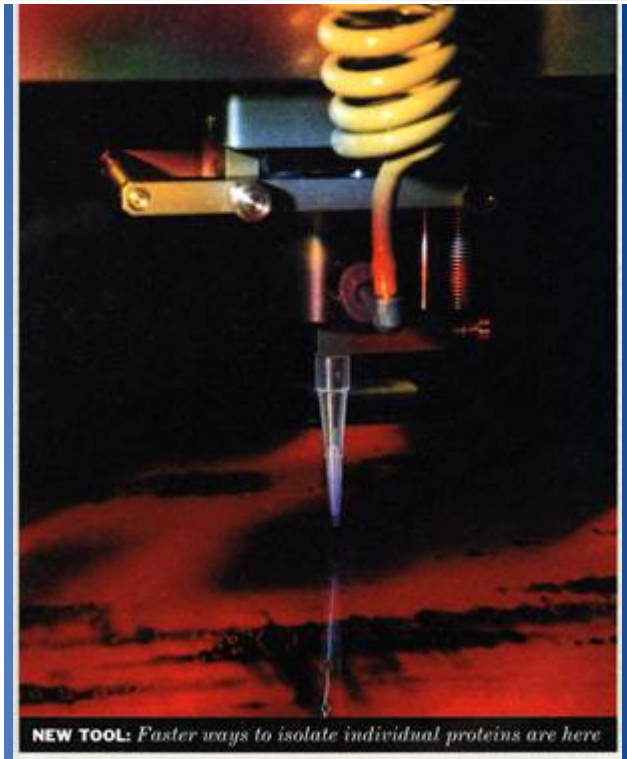
PROTEOMICS

Searching for the real stuff of life

The discovery that humans have fewer genes than expected has thrust proteins into the research spotlight, says Victoria Griffith



www.businessweek.com
BusinessWeek
APRIL 10, 2000
A PUBLICATION OF THE McGRAW-HILL COMPANIES



NEW TOOL: *Faster ways to isolate individual proteins are here*

BIOTECH'S NEXT HOLY GRAIL

Now, companies are racing to decipher the human protein set

Challenges of the Plasma/Serum Proteome

- Mass dominated by albumin and then a couple dozen other high abundance proteins
- Primary plasma proteins with important functions—osmotic pressure, coagulation and anti-coagulation, lipid transport, complement, immunoglobulins
- Variable release of blood cell or platelet proteins
- Most complex specimen, with a very long tail of low abundance proteins of tissue origin
- Extreme dynamic range
- Minimally invasive sample, but pre-analytical variation in techniques and conditions complicates results

HUPO Human Plasma Proteome Project

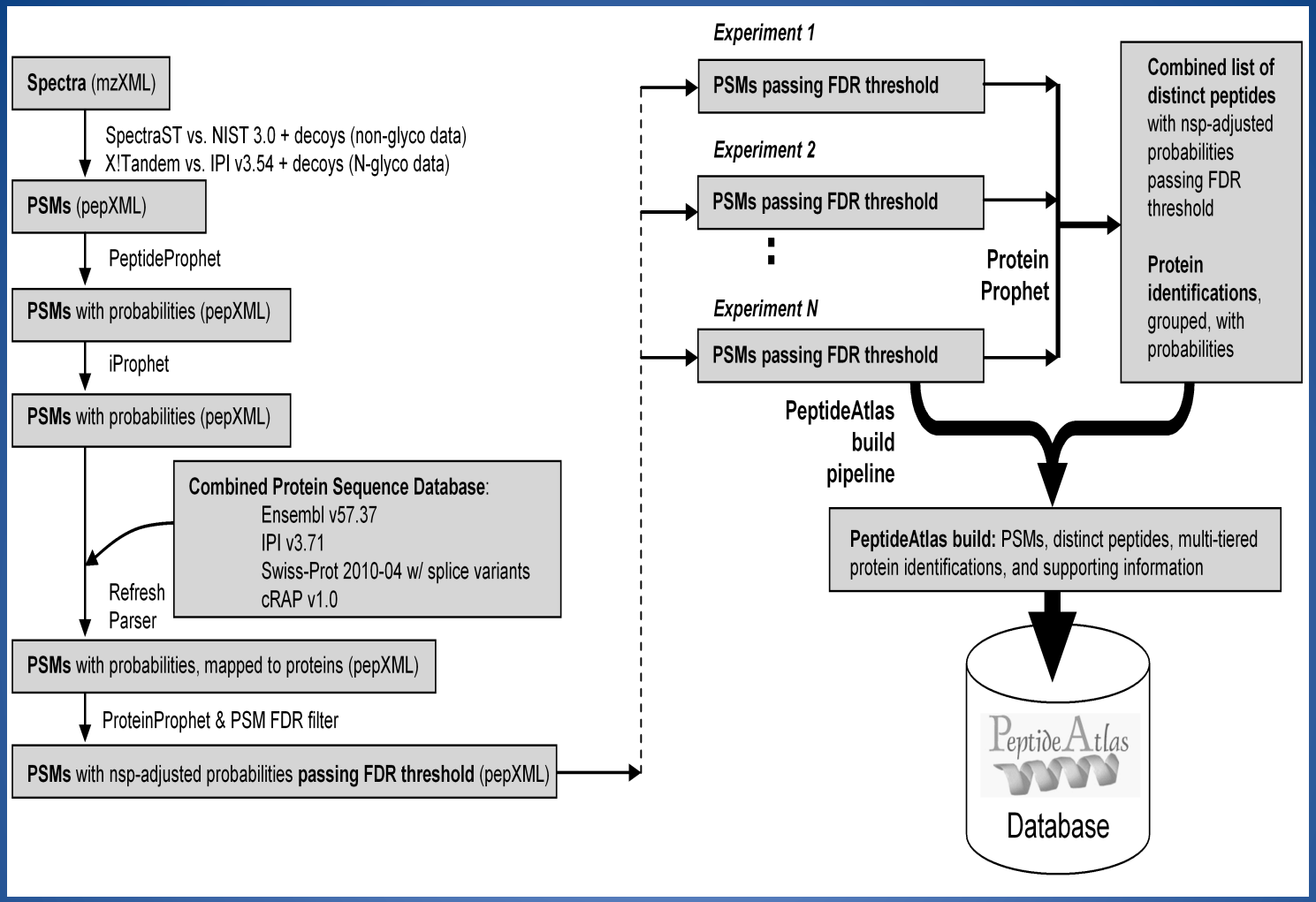
- HPPP: First HUPO Initiative, 2002
- Produced 17 reference specimens, distributed to 55 laboratories worldwide
- Raised funding for small grants program + PSI
- "Exploring the Human Plasma Proteome"—Proteomics 2005/Wiley 2006: 3020 proteins with 2 or more peptide IDs, tiered approach, from 1274
- Stringent analysis, Nat Biotech 2006: 889 IDs
- Additional datasets now captured and combined in PeptideAtlas with uniform analysis

Progressive Plasma PeptideAtlases

- The partial Peptide Atlas for the HPPP-1 in 2005 appears to have had a peptide FDR of about 12 percent.
- The 2007 Plasma Peptide Atlas contained 27,801 peptides mapping to 2738 non-redundant proteins; the PSM FDR was 1 to 3 orders of magnitude higher than the present build (0.0002).
- The 2010-11 Peptide Atlas has fewer peptide and protein IDs, fulfilling much more stringent criteria: protein FDR 0.01, peptide FDR 0.0016, PSM 0.0002

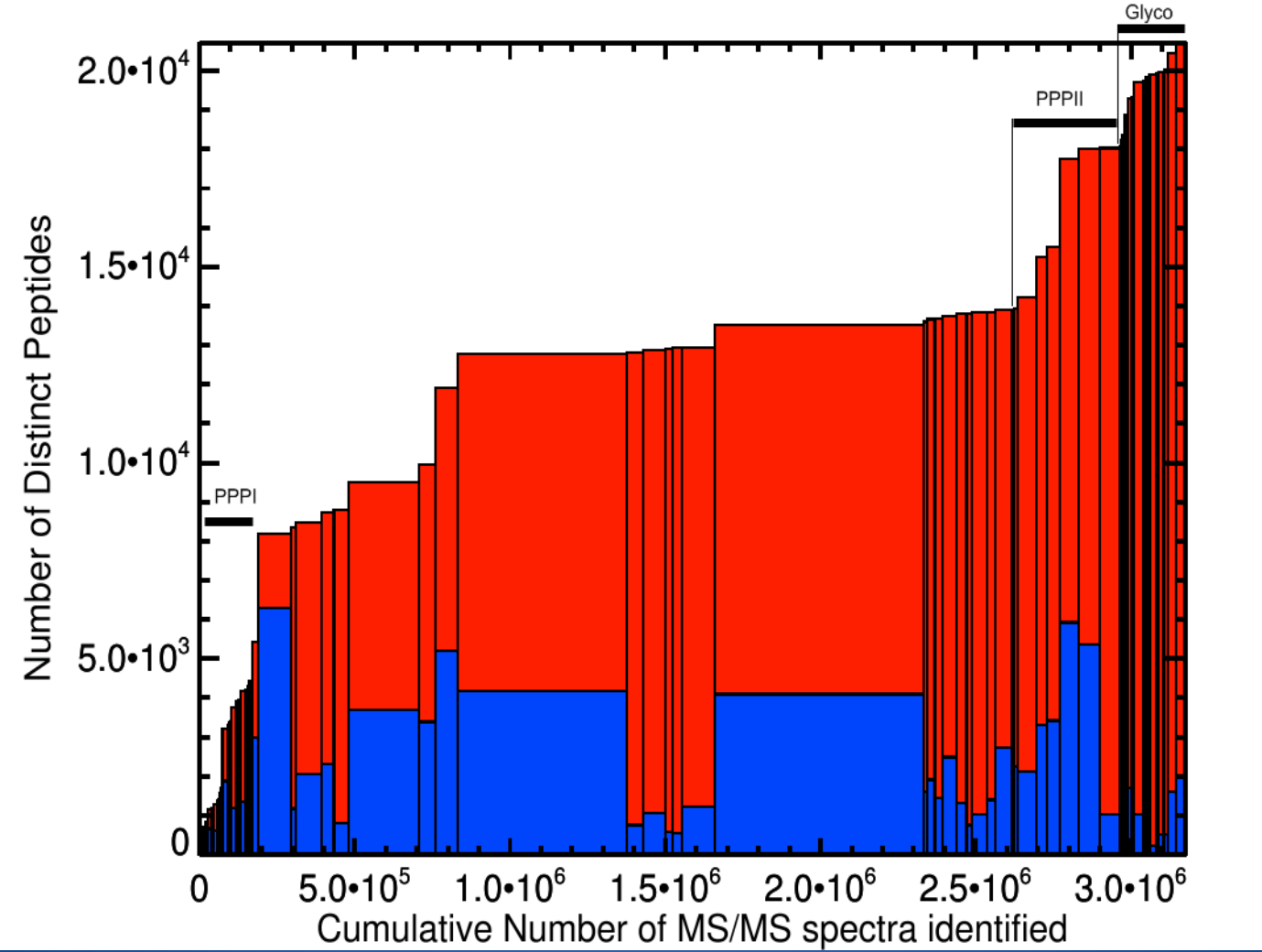
Enhancements to PeptideAtlas

- Searching of spectral library
- Addition of iProphet
- Much more stringent FDR cutoff for PSM
- Use of Mayu for decoy-estimated protein FDR
- Many more datasets; others yet to be added



PeptideAtlas Build from peptide mass spectra: Search, analysis, and validation steps for each LC-MS/MS experiment [Farrah et al, Mol Cell Proteomics 2011]

Human Plasma PeptideAtlas – 91 expts; 3,172,759 peptide-spectrum matches; 20,679 distinct peptides at FDR 0.0016; 1929 canonical proteins at FDR 0.01



Datasets

- PPP-1 contributed 38% of the present canonical proteins.
- Next large accessions gave shallow increase.
- Datasets with extensive fractionation and more accurate instruments gave a big jump upward.
- Likewise, the N-glyco-enriched datasets gave a sharp rise.

Rationale for Cedar: A Multi-Tiered Protein Identification Scheme for Shotgun Proteomics

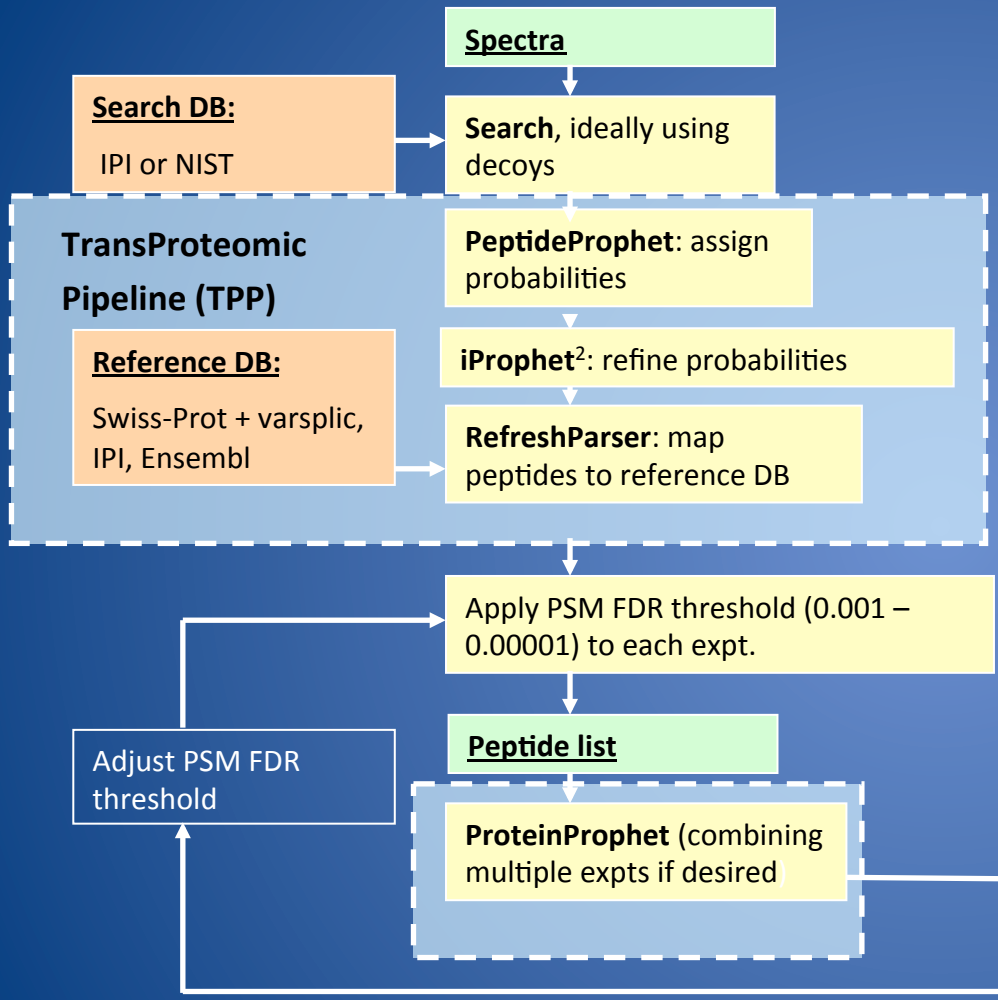
Producing a definitive set of protein identifications from a set of peptide identifications is a continuing challenge in interpreting shotgun proteomics data. To date, there is no standard method.

Mass spectrometry is an inherently incomplete sampling of proteins in a specimen. A cutoff for balance of false-positives and false-negatives must be chosen. There is no standard FDR.

For estimating the number of truly distinct proteins detected, a highly non-redundant protein identification set is preferable.

For comparison with a non-redundant list for another proteome, or selection of peptides for selected reaction monitoring (SRM) experiment design, redundancy is preferable.

Cedar Architecture[1]



Protein identification classification steps

Preliminary protein identification list (***exhaustive set***), organized into sets of indistinguishable protein sequences (share exact same observed peps) and further separated into *protein groups,* with some labeled ***subsumed***

Within each set of indistinguishable protein sequences:

select one with preferred accession for ***distinguishable*** set.

from each cluster of identical sequences, select one with preferred accession for ***sequence-unique*** set.

Among non-subsumed distinguishables, find ***ntt-subsumed***

Among remaining non-subsumed, determine one or more ***canonicals*** for each protein group using 80% peptide sharing threshold. Label others ***possibly-distinguished***

Protein FDR ~ 1% for canonical set

no

yes

Among canonicals + possibly distinguished, iteratively find ***covering set***.

1.Farrah, et al., A High Confidence Human Plasma Proteome Reference Set with Estimated Concentrations in the PeptideAtlas, MCP 2011

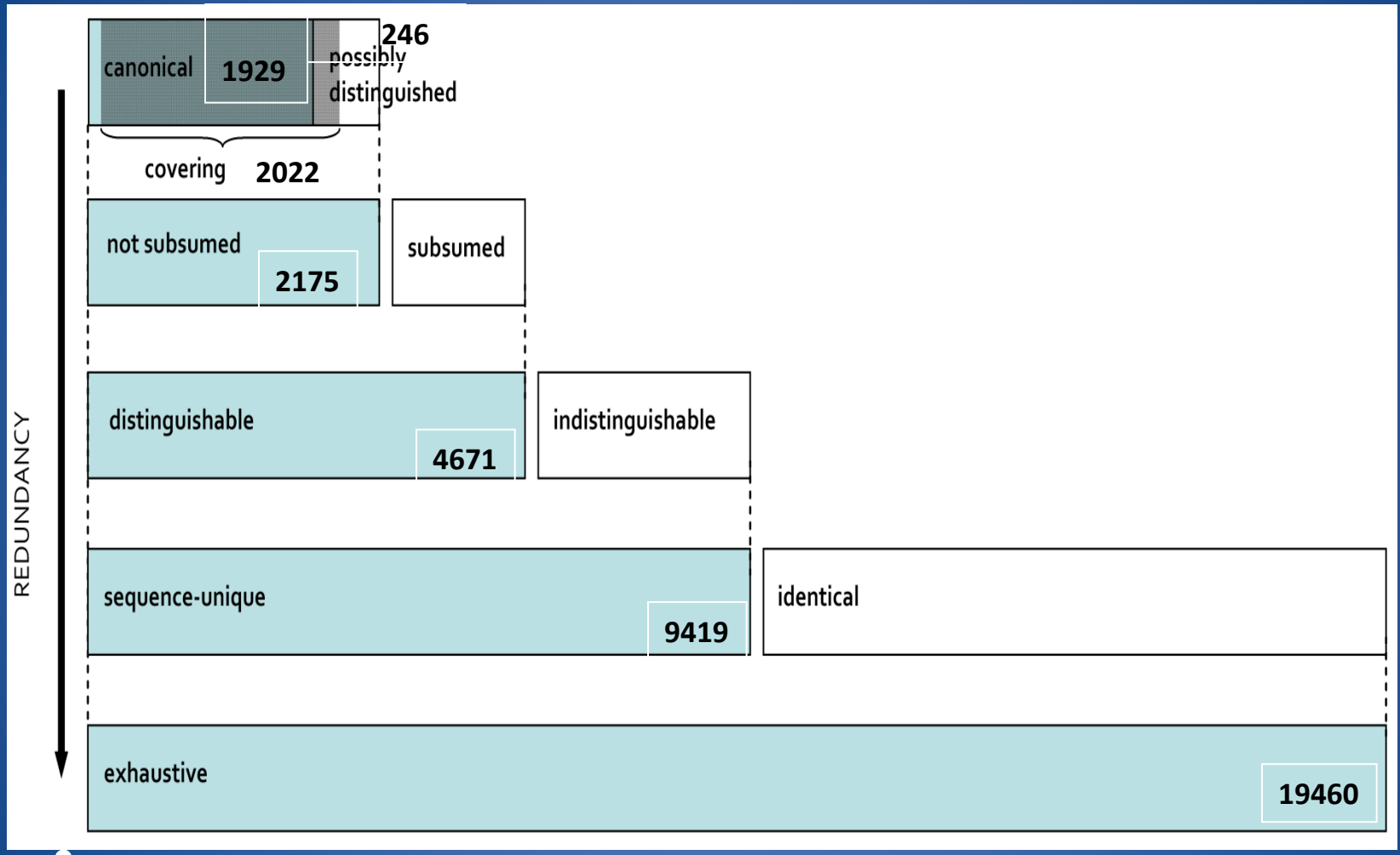
2.Shteynberg, et al., iProphet: Improved Statistical Validation of Peptide Identifications in Shotgun Proteomics, submitted to MCP.

iProphet: A New Feature of TPP

Models five additional properties, adjusting peptide probabilities accordingly:

1. Number of sibling searches
2. Number of replicate spectra
3. Number of sibling experiments
4. Number of sibling ions
5. Number of sibling mass modifications

Multi-Tiered Protein Identifications

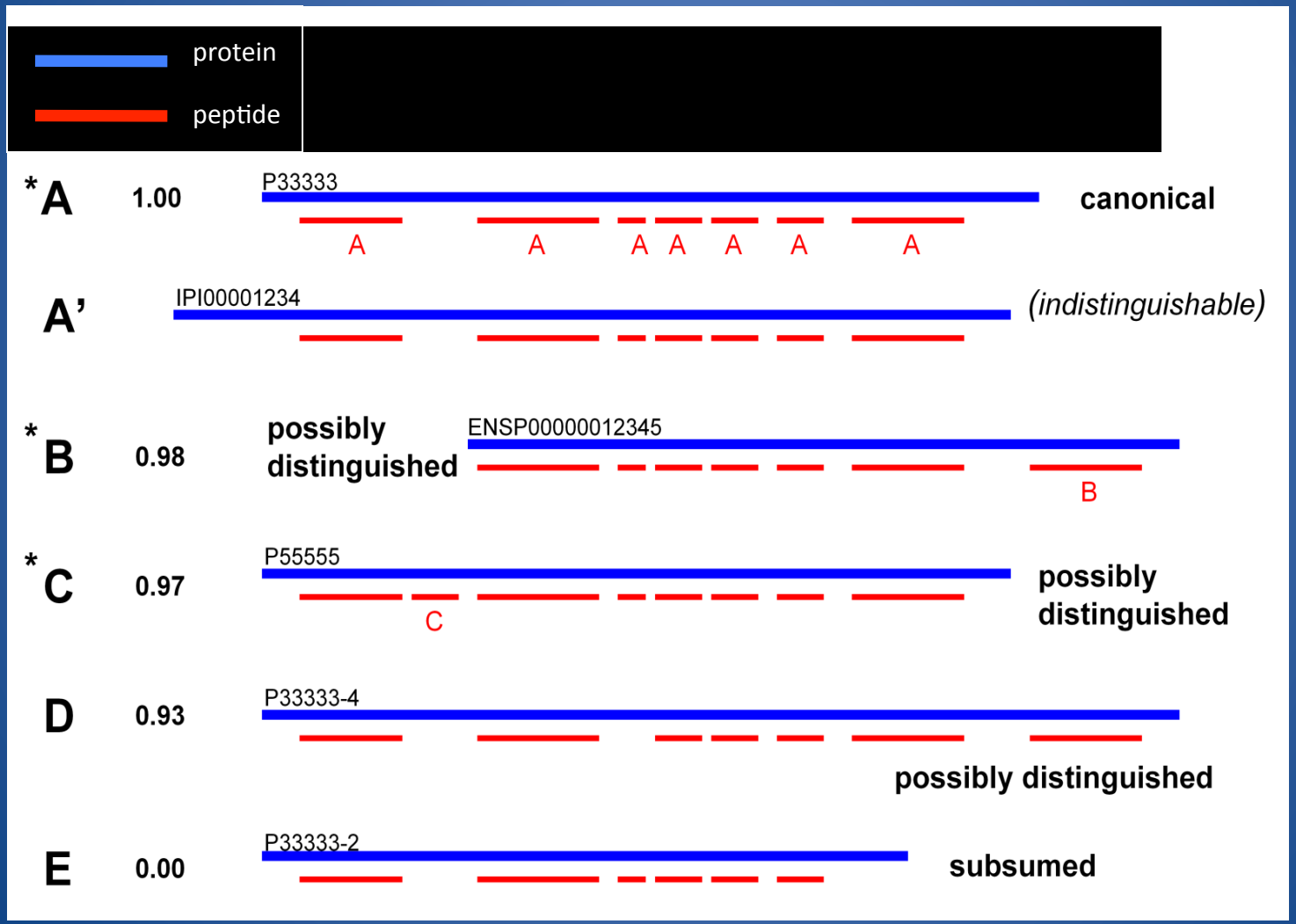


Canonical set *is conservative list;* **Exhaustive set** *is comprehensive list*

Covering set *provides protein identification for each peptide*

Illustration of Terminology at Peptide Level

using a hypothetical ProteinProphet protein group

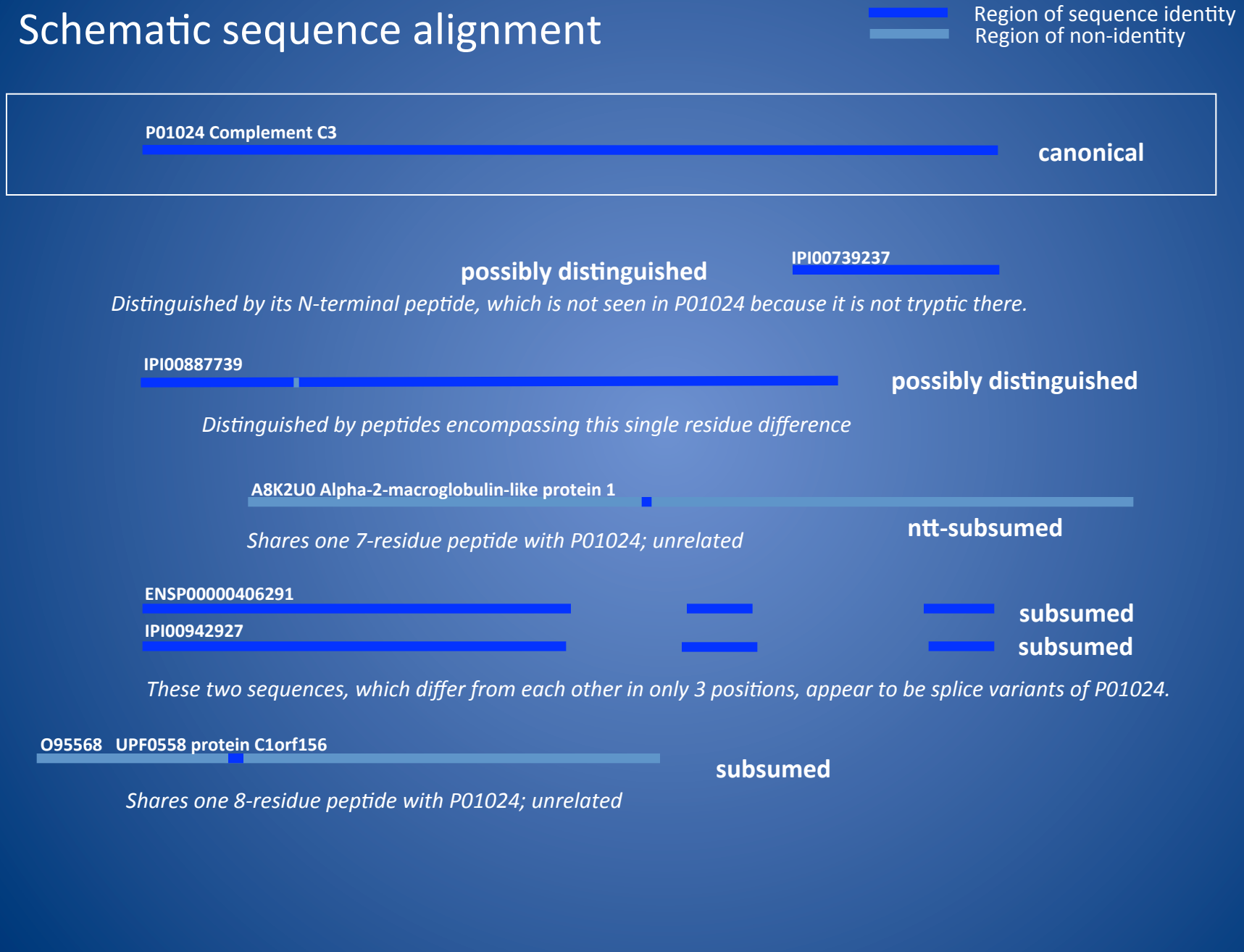


Example: Complement C3 Protein Group in Human Plasma PeptideAtlas

Biosequence Name	Covering Set	Presence Level	Protein Prophet Prob	N Obs	N Distinct Peptides	Estimated ng/ml	Uncertainty ng/ml	Redundancy Relationship	Redundant With Respect To	Protein Group	Seq Uniq Prots in Grp	Protein Group Seq Alignmt	
P01024	C	canonical	1.000	83136	448	750000	3x			P01024	8		CO3_HUMAN Complement C3 OS=Homo sapiens GN=C3 PE=1 SV=2
ENSP00000245907								identical	P01024	P01024	8		pep:known chromosome:GRCh37:19:6677846:6720693:-1 gene:ENSG00000
IPI00783987								identical	P01024	P01024	8		IPI:IPI00783987.2|SWISS-PROT:P01024|ENSEMBL:ENSP00000245907|REFSE
IPI00739237	C	possibly distinguished	1.000	15924	95					P01024	8		IPI:IPI00739237.1|REFSEQ:XP_001724196 Tax_Id=9606 Gene_Symbol=LOCE
IPI00887739	C	possibly distinguished	1.000	67443	358					P01024	8		IPI:IPI00887739.3|REFSEQ:XP_001719515 Tax_Id=9606 Gene_Symbol=LOC1
ENSP00000299698								indistinguishable	A8K2U0	P01024	8		pep:known chromosome:GRCh37:12:8975217:9029379:1 gene:ENSG00000
IPI00419215								identical	A8K2U0	P01024	8		IPI:IPI00419215.5|SWISS-PROT:A8K2U0 Tax_Id=9606 Gene_Symbol=A2ML1
IPI00956115								identical	ENSP00000299698	P01024	8		IPI:IPI00956115.1|TREMBL:B3KVV6;B5MDD1;B7Z7V4;D3DUV3;Q6ZWK7|ENS
A8K2U0		NTT subsumed	0.002	113	1				P01024	P01024	8		A2ML1_HUMAN Alpha-2-macroglobulin-like protein 1 OS=Homo sapiens GN=
ENSP00000307077								identical	O95568	P01024	8		pep:known chromosome:GRCh37:1:169761670:169763830:-1 gene:ENSG0
ENSP00000307975								identical	O95568	P01024	8		pep:known chromosome:GRCh37:1:169761670:169764107:-1 gene:ENSG0
IPI00009815								identical	O95568	P01024	8		IPI:IPI00009815.1|SWISS-PROT:O95568|ENSEMBL:ENSP00000307077;ENSP
ENSP00000406291		subsumed	0.000	59027	291				P01024	P01024	8		pep:known chromosome:GRCh37:19:6677846:6720662:-1 gene:ENSG00000
IPI00942927		subsumed	0.000	58618	296				P01024	P01024	8		IPI:IPI00942927.1|TREMBL:B4DR57;B4E216;Q6LDJ0|ENSEMBL:ENSP0000040
O95568		subsumed	0.000	1	1				P01024	P01024	8		CA156_HUMAN UPF0558 protein C1orf156 OS=Homo sapiens GN=C1orf156

Complement C3 group:

Schematic sequence alignment



Inclusions and Exclusions

- Of the 1929 canonicals, 1642 are supported by >1 PSM and 1313 are supported by >1 distinct peptide.
- Whether to include any single-PSM “one-hit wonders”, is a frequent decision. Given our stringent parameters, we performed manual validation on all single-PSM IDs. We excluded 70 which failed to fully meet our criteria, while retaining 287 of the original 357.

Sensitivity Analysis: FDR

- We introduced the Mayu decoy-based algorithm for FDR.
- Mayu improves accuracy by taking into consideration size of the dataset, number of tryptic peptides in each protein, and proteome coverage.
- Using 1% vs 5% protein FDR: 340 likely correct identifications are excluded, in order to avoid including 96 false-identifications. Of course, we don't know which is which.

	Fraction of identifications containing N-glycosite motif	
	Non-glyco PeptideAtlas 69 datasets	N-Glyco PeptideAtlas 22 datasets
Distinct peptides	3.9%	53%
Canonical proteins	72%	90%

Prevalence of N-glycosite motif in the component builds of the Human Plasma PeptideAtlas, showing enrichment of less abundant proteins in the N-Glyco-PeptideAtlas.

Comparison with Other Plasma Datasets

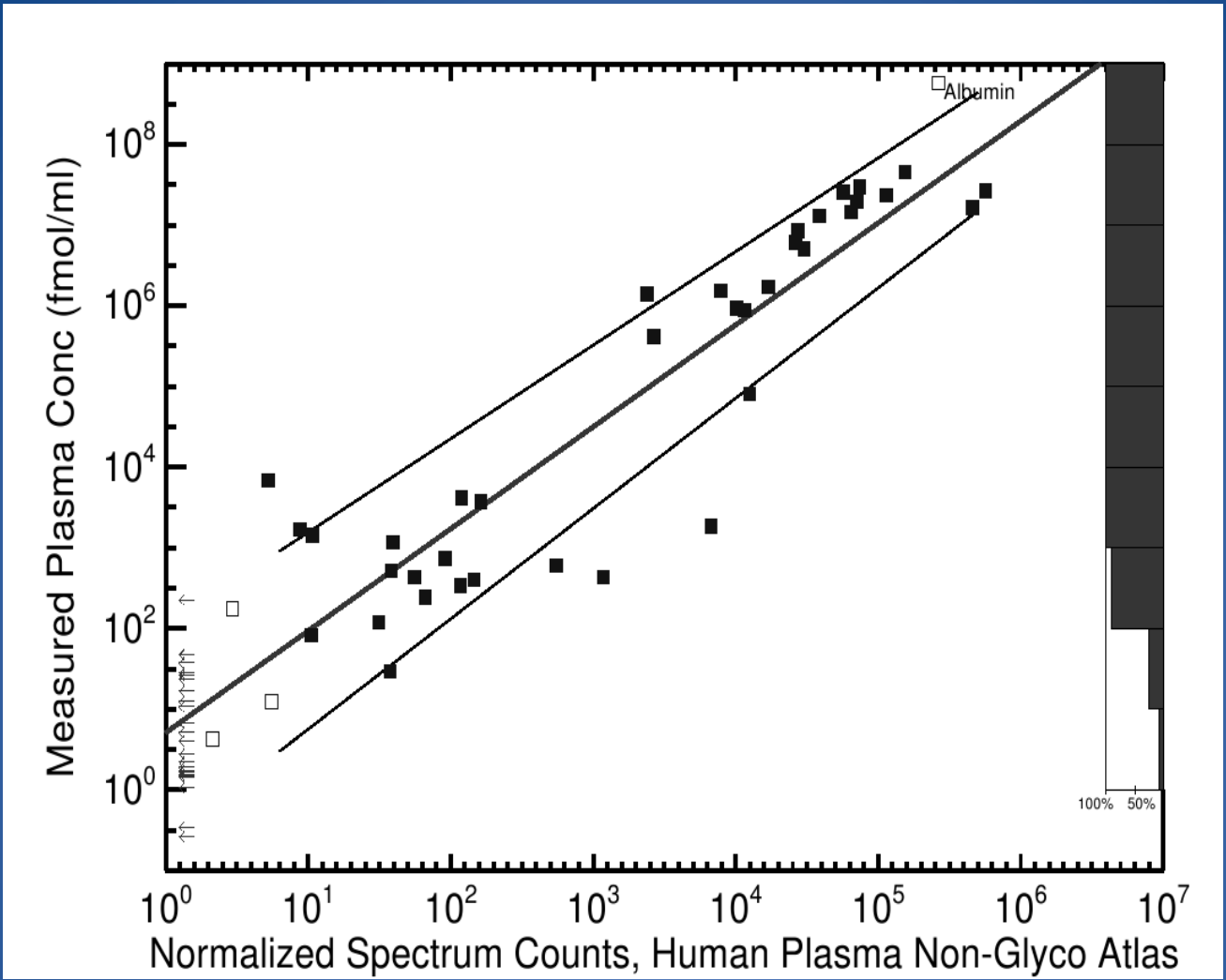
- The PeptideAtlas' 1929 canonicals include all of Hortin's 126 clinical plasma proteins, Kuzyk's 45 cardiovascular plasma biomarker candidates, and Carrascal's 44 phosphoproteins (except 1 manually excluded).
- The Schenk et al (2008) plasma reference set has 697 non-redundant, non-IG proteins; we have 503 of the 554 (91%) in the sequence databases; the remaining 51 are not in our PeptideAtlas exhaustive set. Probably, most would be added to our canonical or larger lists, if we had the spectra.

Splice Variant Protein Isoforms

- Splice variants are a special interest of mine, including those differentially expressed in tumors.
- Swiss-Prot has only one entry for which two splice isoforms exist in the canonical set: mannan-binding lectin serine protease I, both of which we detect in human plasma. Twelve additional Swiss-Prot ASVs are noted as possibly-distinguished.
- 131 canonical protein sequences came from IPI or Ensembl, not in SwissProt; some may be ASVs.

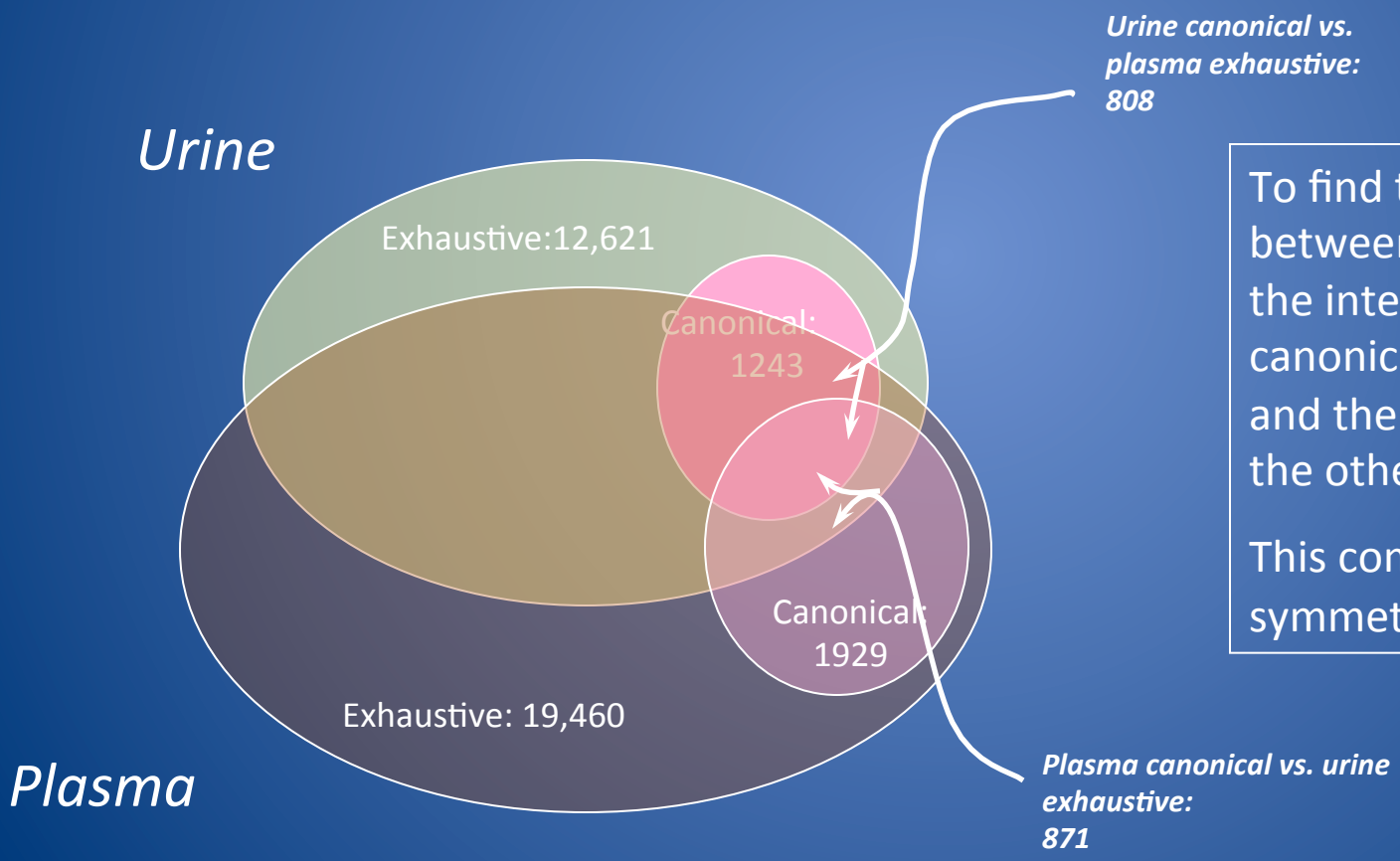
Application of PeptideAtlas to SRM Experiments

- Peptide Atlas shows which peptides have been detected; we have created an “empirical observability score” (EOS).
- PeptideAtlas links to the spectral library to guide SRM peptide production and standards
- Protein concentrations estimated from correlations of spectral counts with immunoassay results guide the spike-in design



Plasma protein concentrations determined using immunoassay and antibody microarray analysis (HPPP, Haab 2005) versus normalized spectral counts from the Human Plasma Non-glyco PeptideAtlas, plotted on a log scale. Values in ng/ml, with uncertainty factors, are now in PeptideAtlas.

Application: Comparing Protein Lists for Human Plasma and Urine

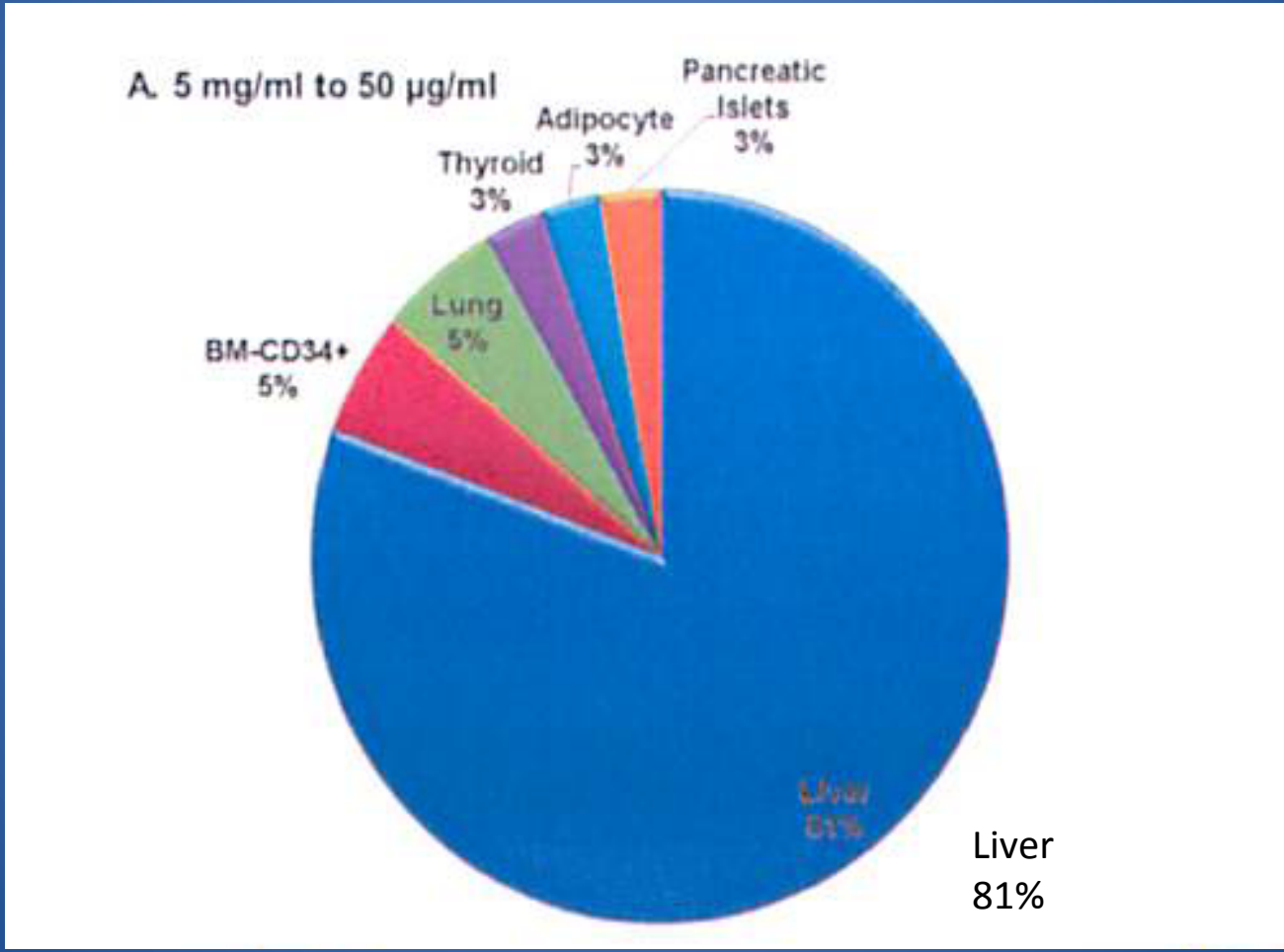


To find the overlap between two atlases, take the intersection of the canonical list for one atlas and the exhaustive list for the other.

This comparison is not symmetric.

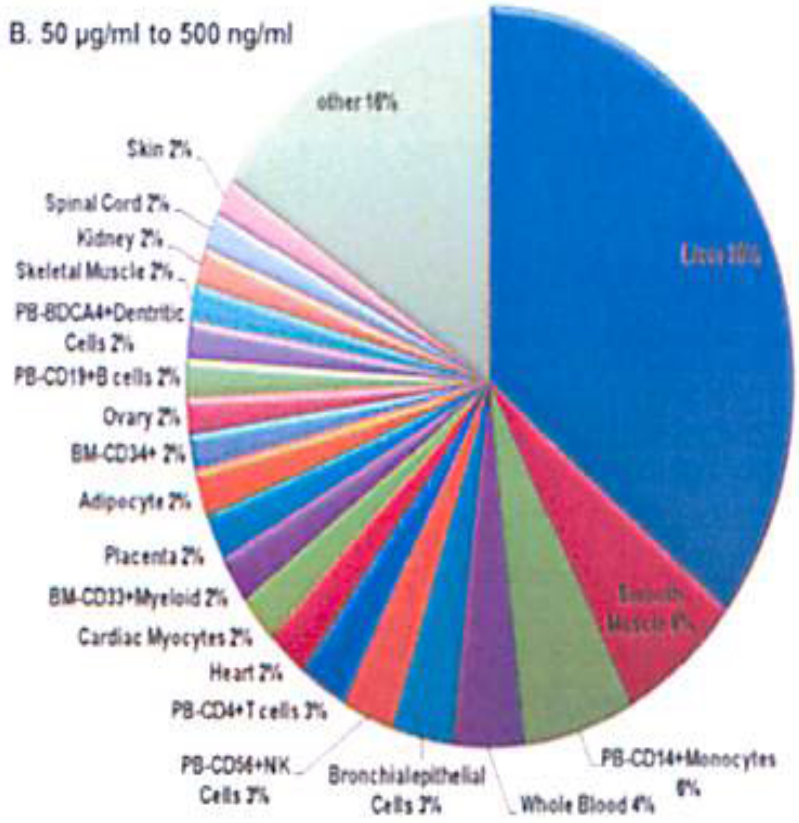
Organ/Cellular Origins of Abundant Human Plasma Proteins

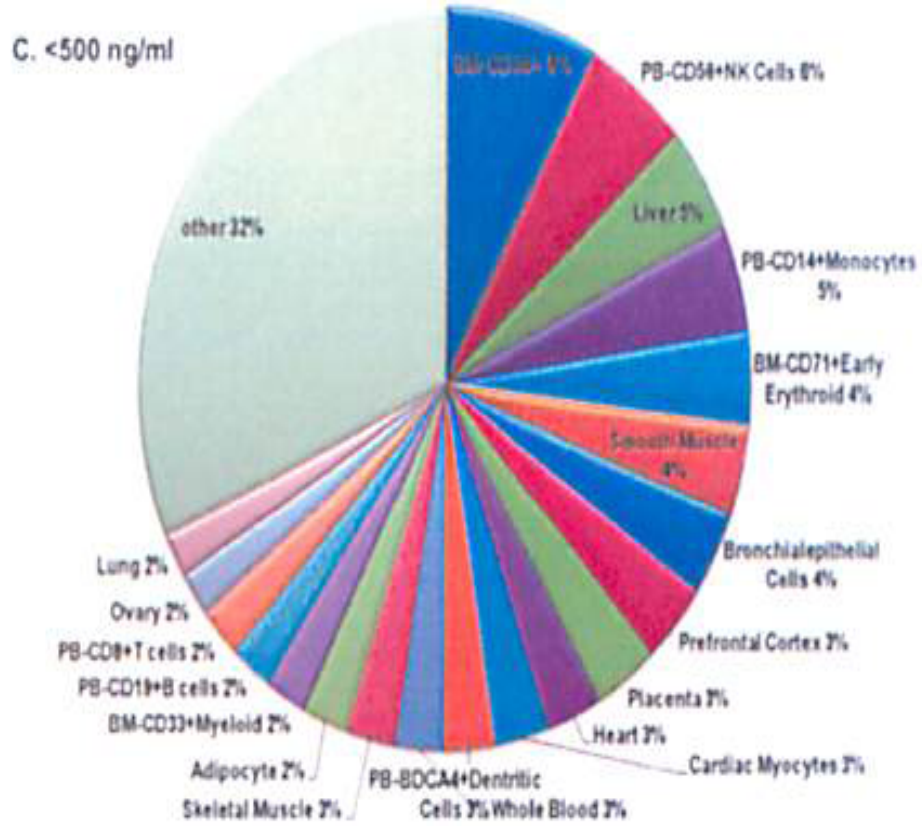
(Zhang et al, JPR 2010)



Liver
81%

Organ/Cellular Origins of Lower Abundance Plasma Proteins





Further Uses of this Scheme

- We are going to add this Cedar tiered protein identification scheme to the freely-available, widely-used TransProteomicPipeline (TPP).
- Thus, it will be applicable for any dataset convertible to mzML or mzXML.
- All steps except the manual validation of single-PSM identifications are clearly defined and reproducible.
- We propose this scheme as a standard for the community, including the Human Proteome Project.

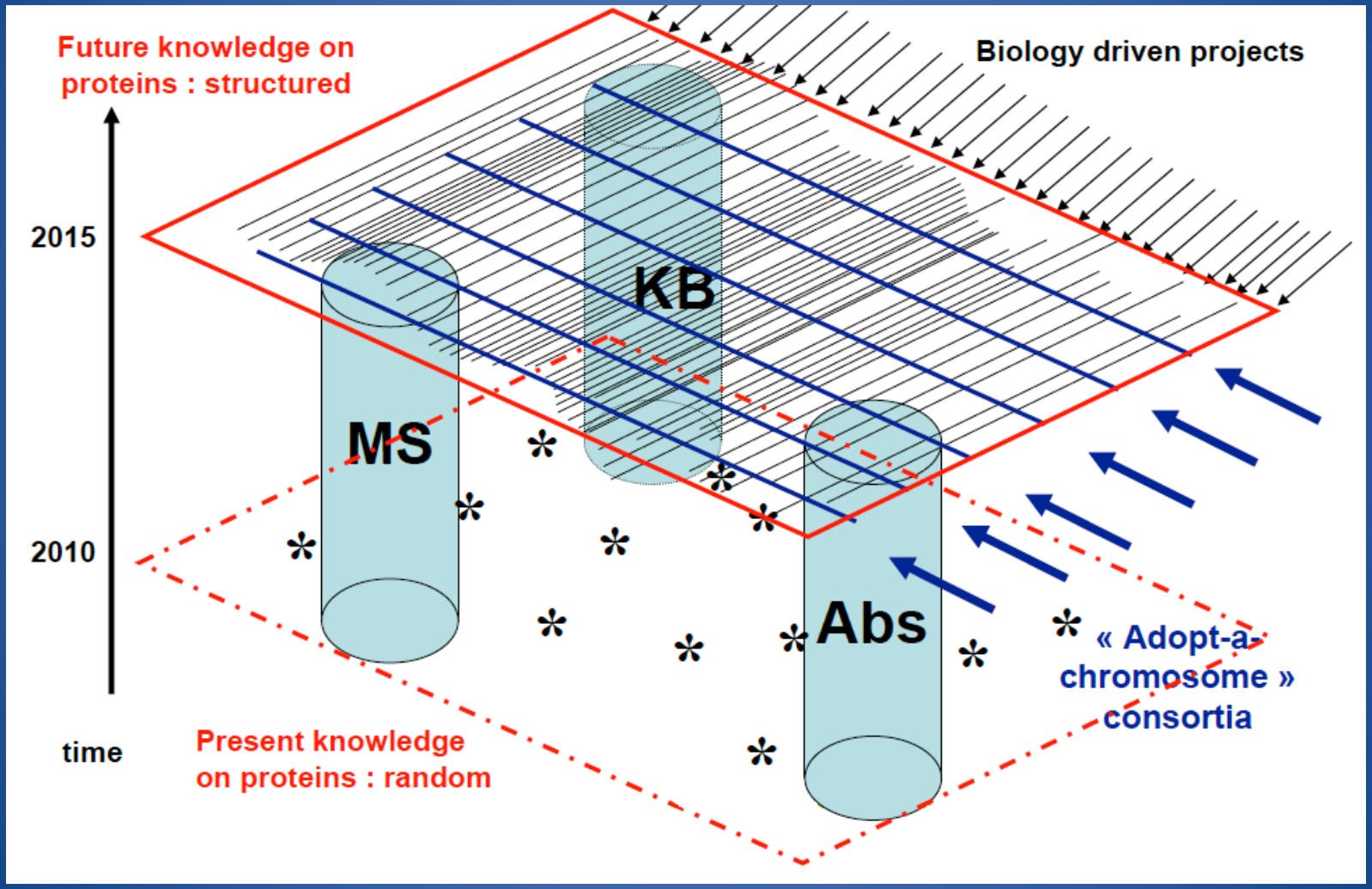
The Mission of the HPP

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 to use proteomics for diagnosis, prevention, therapy, and cure of diseases and improved health worldwide.

The Vision of the HPP



Acknowledgements and Citation

A High-Confidence Human Plasma Proteome Reference Set with Estimated Concentrations in PeptideAtlas.

Mol Cell Proteomics 1 June 2011, M110.006353

(Presented at HUPO Congress 2010; US HUPO Mtg 2011)

Terry Farrah, Eric W. Deutsch, Gilbert S. Omenn, David S. Campbell, Zhi Sun, Julie A. Bletz, Parag Mallick, Jonathan E. Katz, Johan Malmström, Reto Ossola, Julian D. Watts, Biaoyang Lin, Hui Zhang, Robert L. Moritz, Ruedi Aebersold

Institute for Systems Biology, Seattle, WA

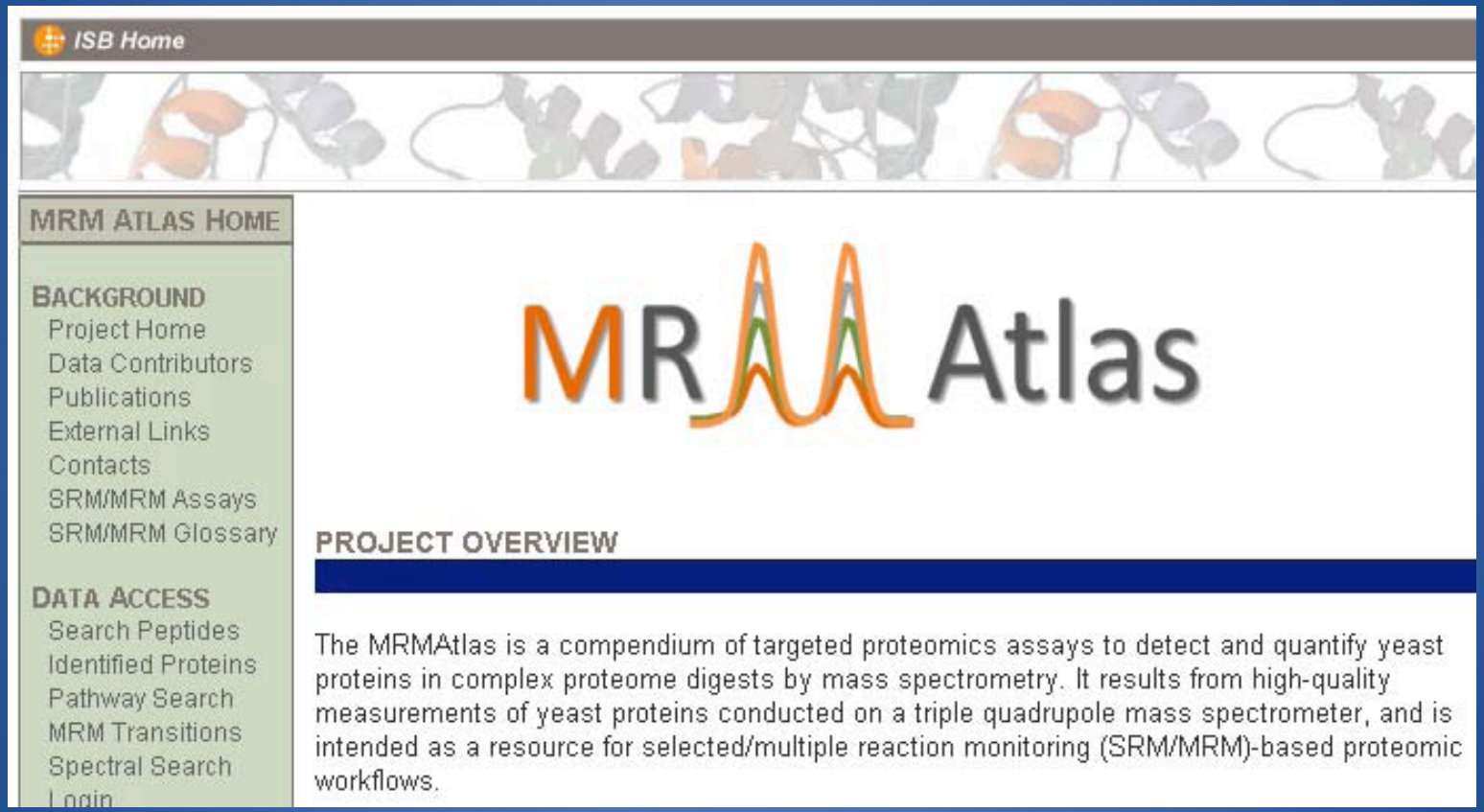
Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

Department of Biology, Institute of Molecular Systems Biology, ETH (Swiss Federal Institute of Technology), Zurich, Switzerland

HUPO Human Plasma Proteome Project: co-chairs Young-Ki Paik, Ruedi Aebersold, Mark Baker, Gil Omenn; founder of HUPO Initiatives, Sam Hanash; and all participating investigators.

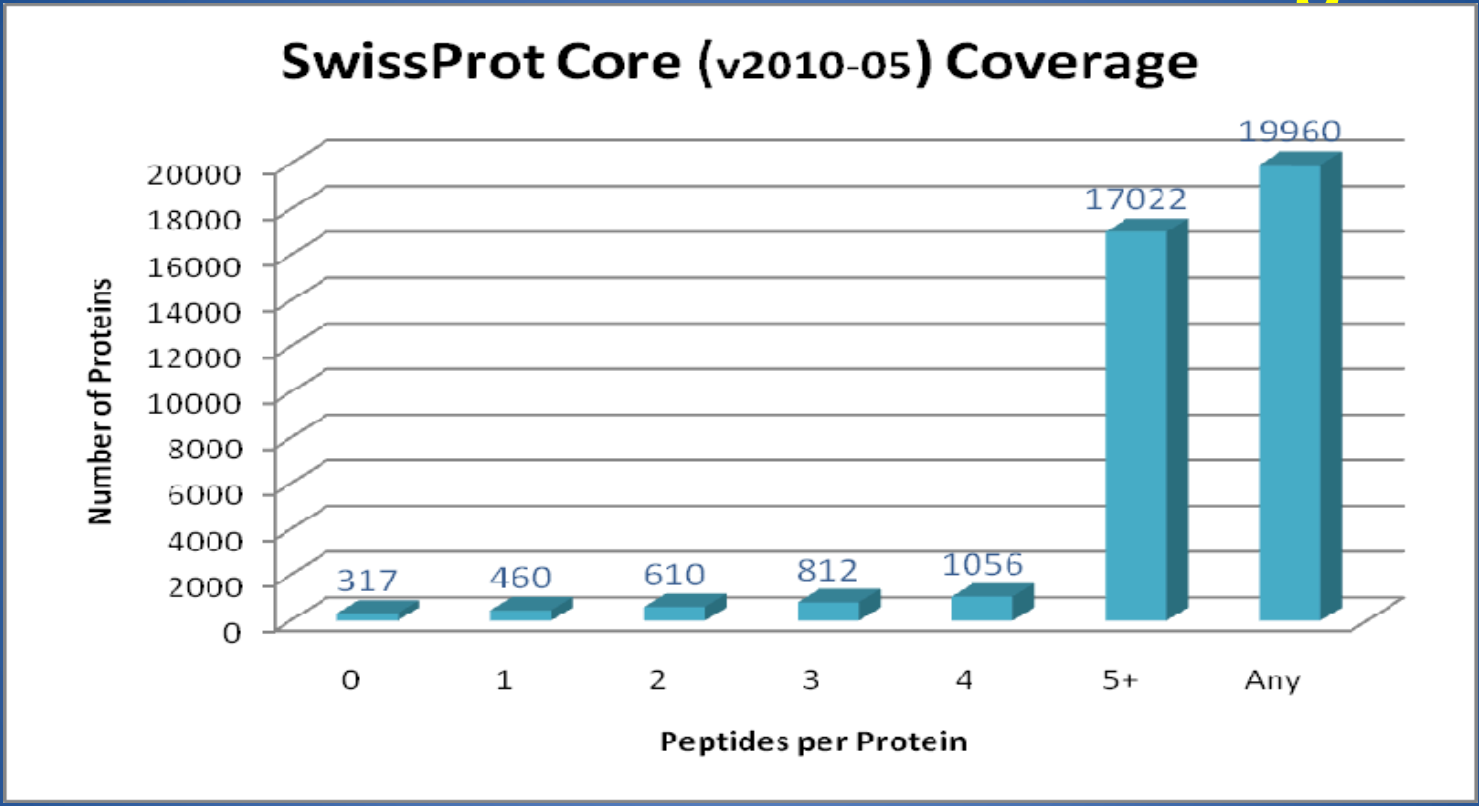
Extra Slides

www.mrmatlas.org



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

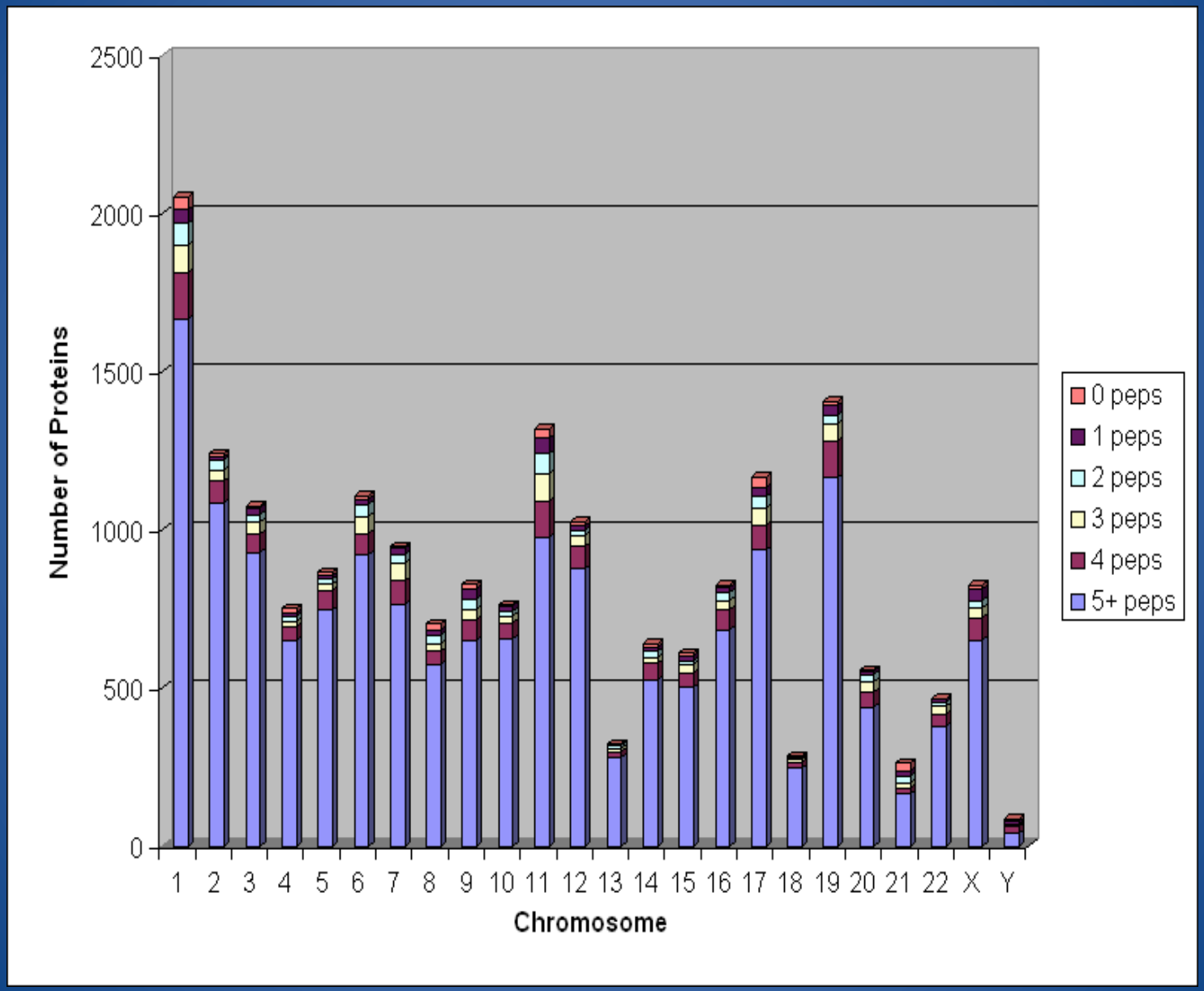
Human Proteome SRM Coverage



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

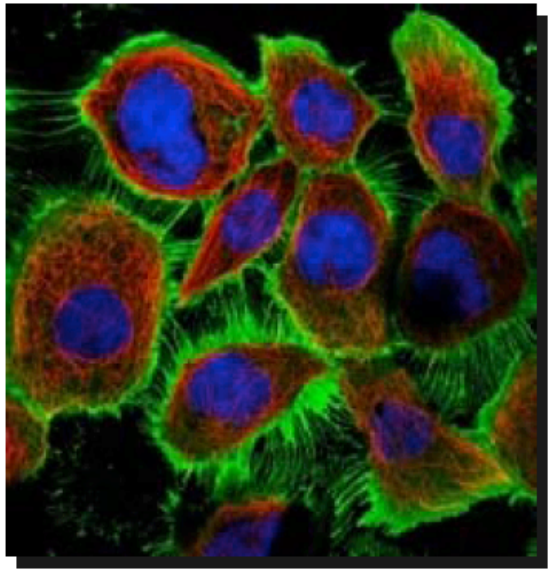
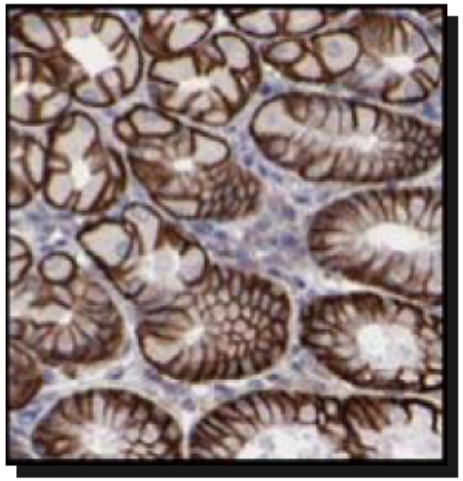
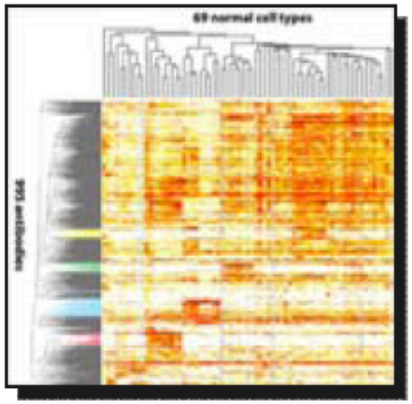
Distribution of Peptide/Protein SRM Coverage by Chromosome



Next Phase of MS Pillar to 2016

- Make data from first phase publicly accessible
- Extend SRM library approach
  - PTMs
  - SNPs
  - Splice variants
- 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—compare protein IDs and tissue expression from MS and from Ab studies

Contribution to the HPP by Antibody-based Proteomics

Mathias Uhlén, HumanProtein Atlas

Knowledge-Based Resource
Normal and Disease Profiles of 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 Resources

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**

Status of Human Protein Atlas v7.0
www.hpa.org (15 November 2010)

- 13,000 Antibodies against 10,000 proteins
- Tissue expression profiles for 46 normal tissues and 47 cell lines
- Subcellular localization by immunohistochemistry
- Expression profiles for 20 cancers (12 individuals)
- Antibody and protein meta-data available
- Needed: comparison with protein IDs and localization with Mass Spectrometry
- Next version expected at HUPO-2011 Geneva Congress 4-7 September

Five-Year Antibody-Based Objectives of HPP (2016)

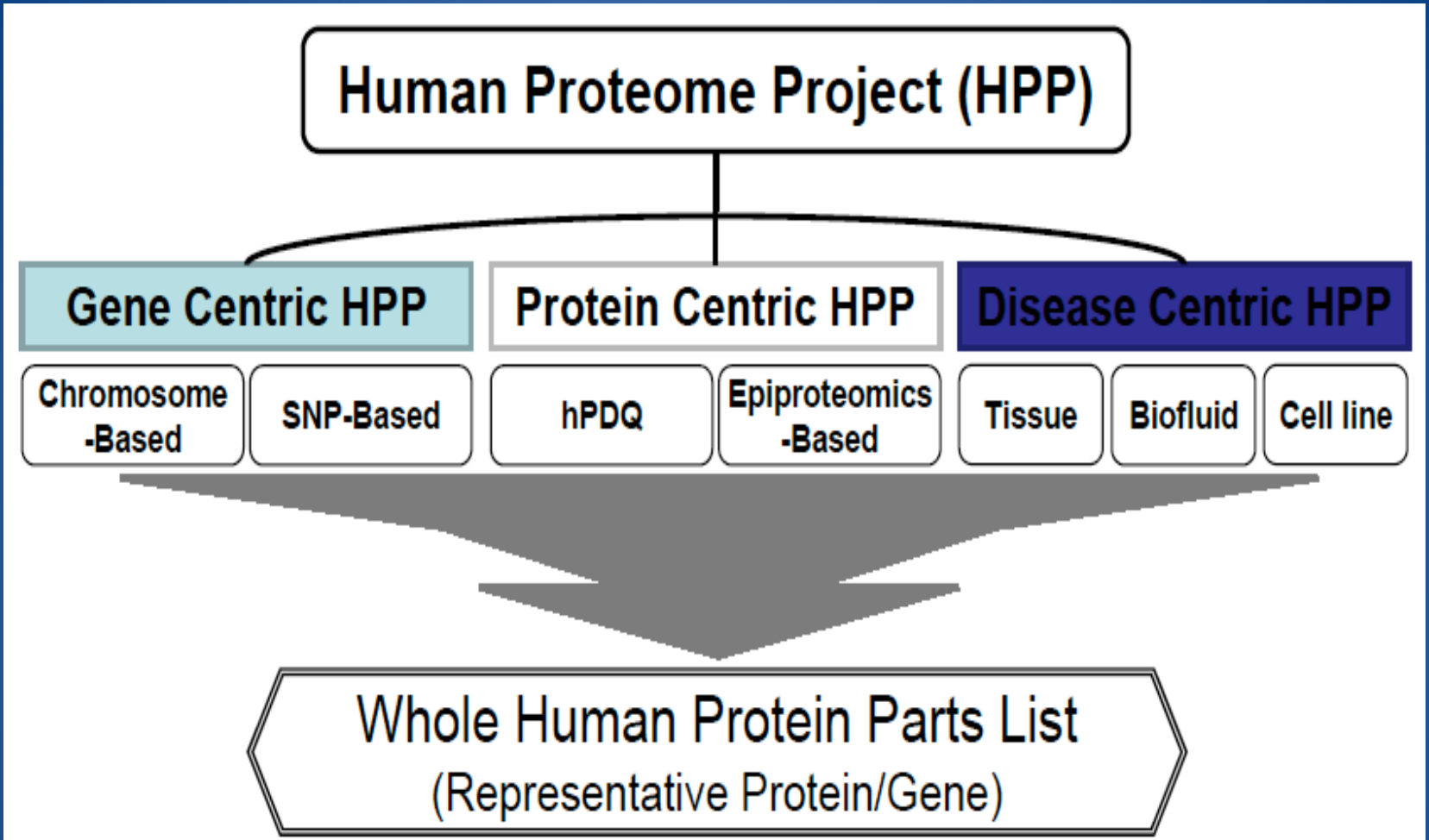
- At least one antibody to a representative protein from every gene
- Contributions from both academic and commercial providers
- Technology to generate recombinant binders in a systematic manner
- Validation of all "HPP antibodies" with new technologies (such as siRNA)
- Draft version of subcellular localization covering all human genes/proteins
- 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)

Database and Knowledge Resources

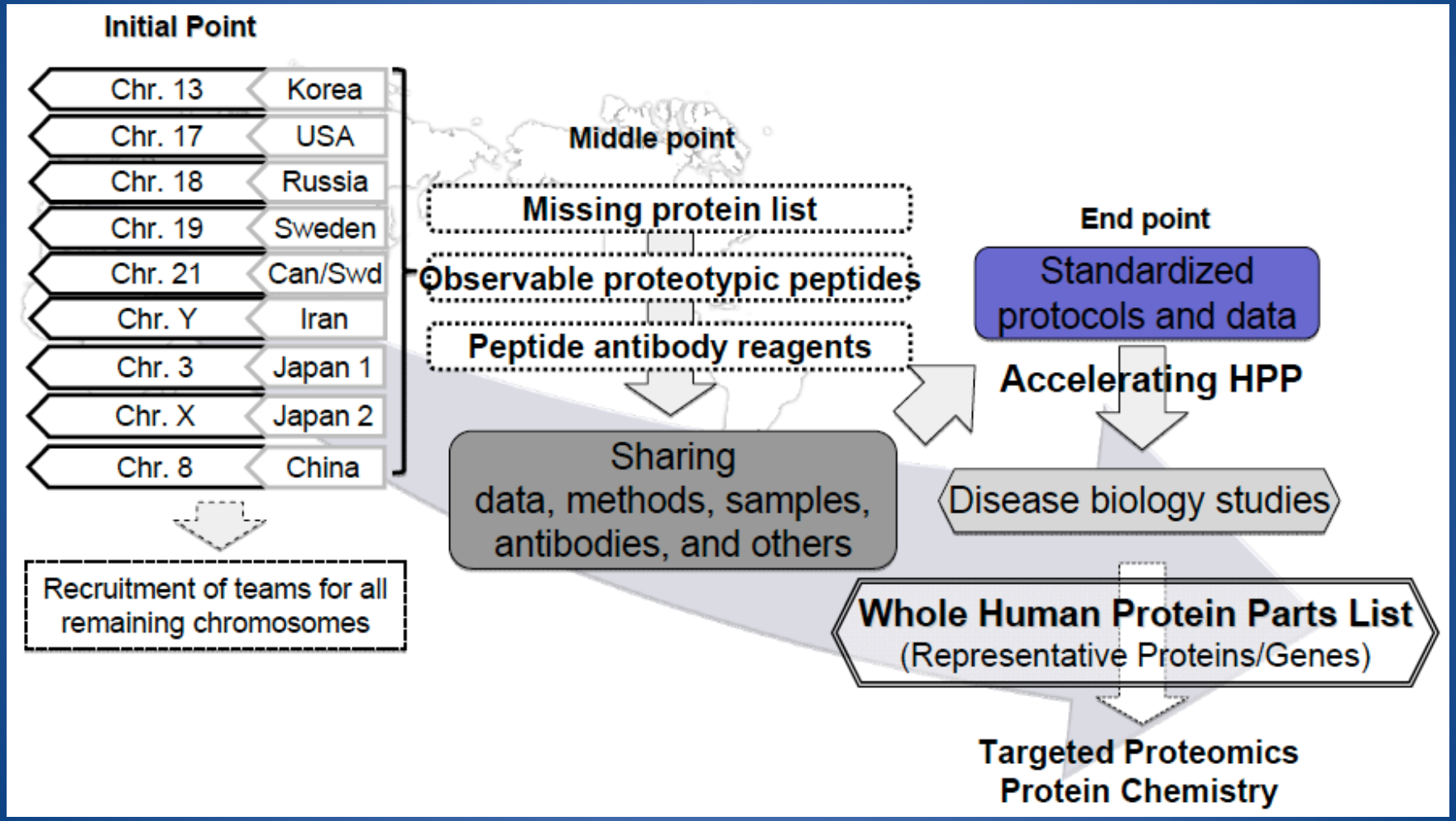
- **Sequences (including PTMs, SNPs, splice isoforms):** ***UniProtKB/Swiss-Prot human section; NeXt-Prot***

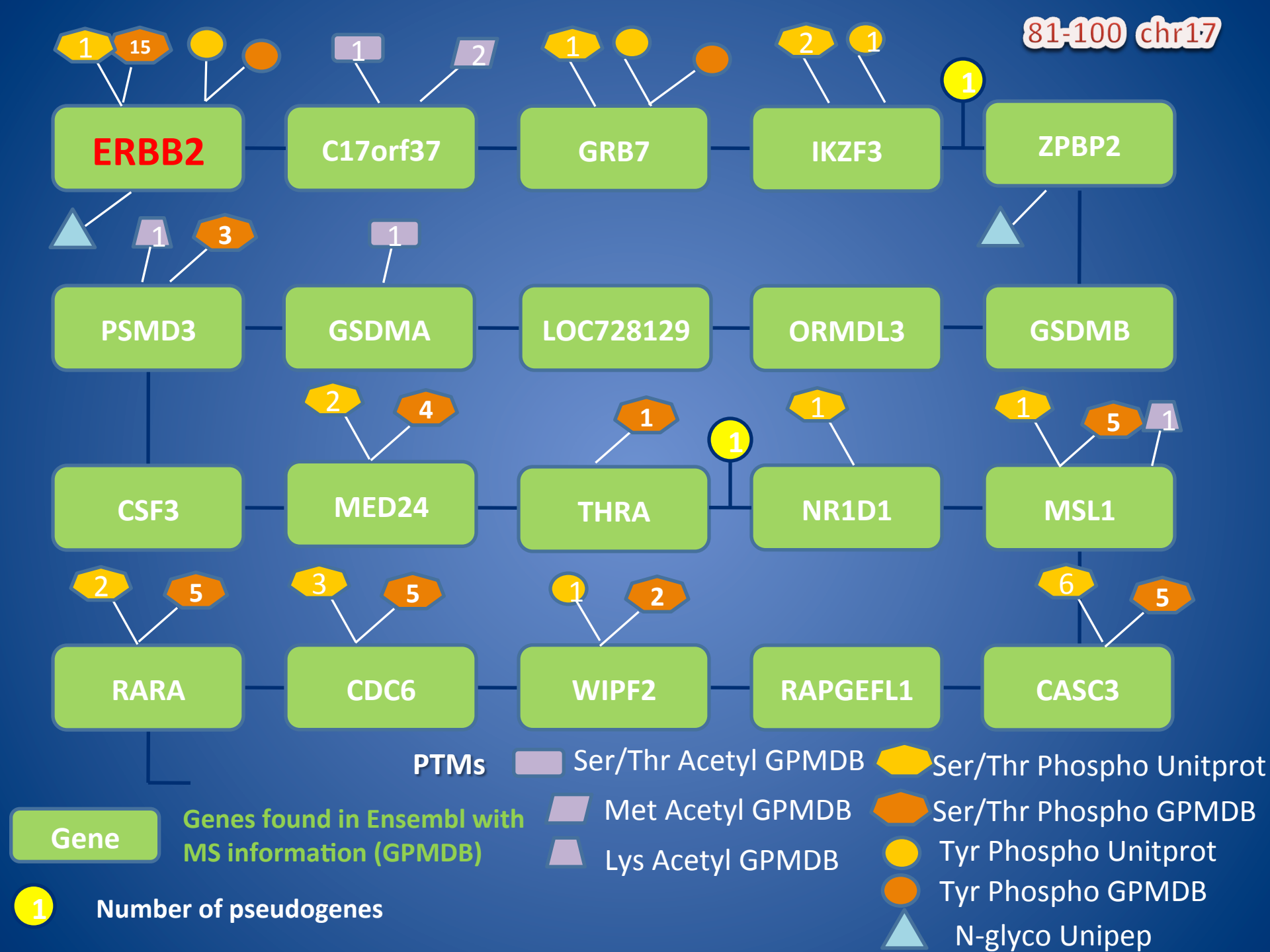
- **For protein IDs, participating labs should use HUPO PSI Extended FASTA Format (*PEFF), TPP, and Cedar*.**

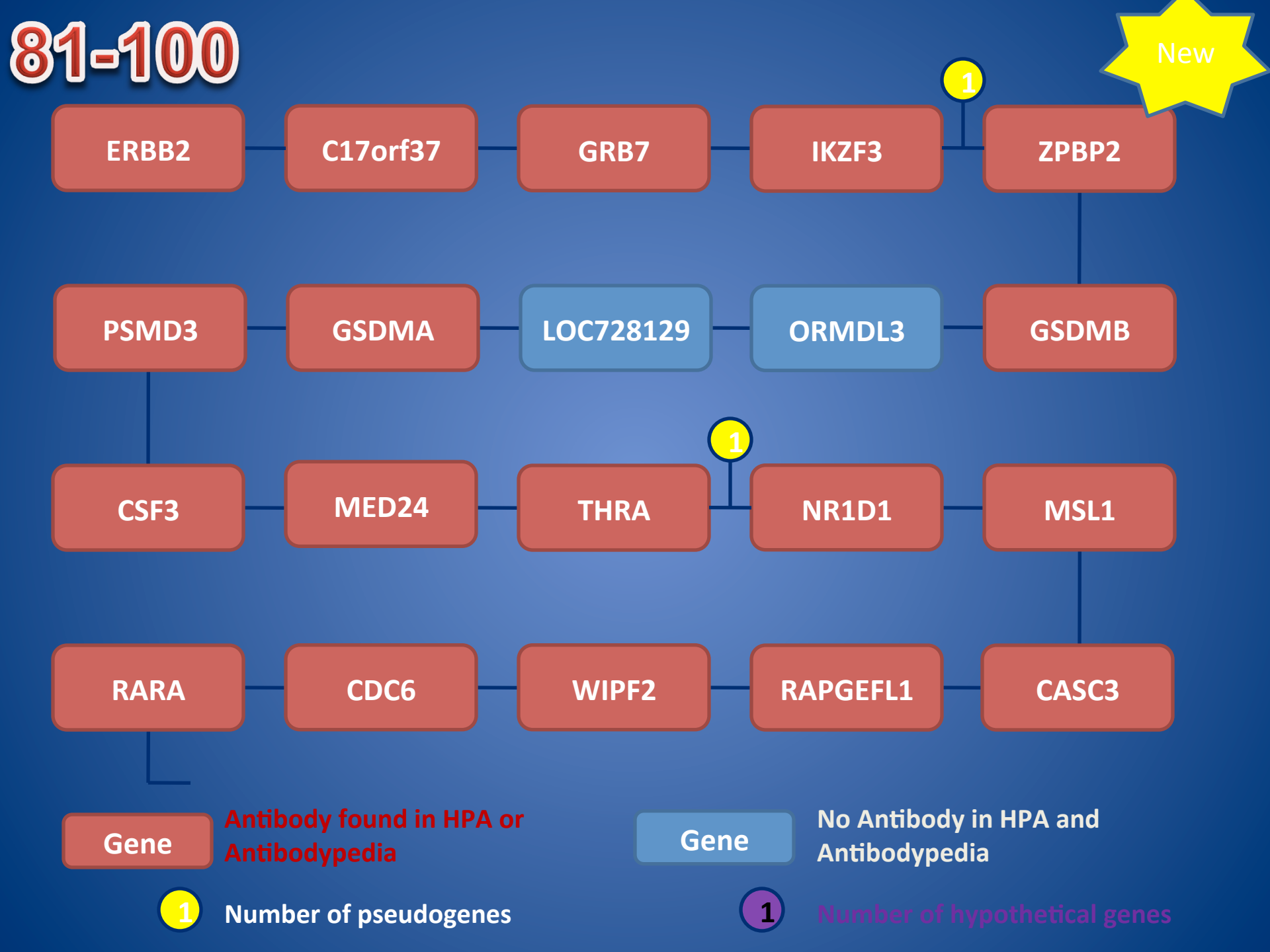
- **Identification data should be deposited in the proteomics repositories (*PRIDE, Tranche)* and be subjected to uniform re-analysis *(PeptideAtlas, GPMDB)* complying with the standards that are being developed in the framework of the EU-funded *ProteomeXchange* program.**


Human Proteome Project (HPP)
Gene Centric HPP
Protein Centric HPP
Disease Centric HPP
Chromosome -Based
SNP-Based
hPDQ
Epiproteomics -Based
Tissue
Biofluid
Cell line
Whole Human Protein Parts List
(Representative Protein/Gene)

Chromosome-Based HPP (as of 9/2010)



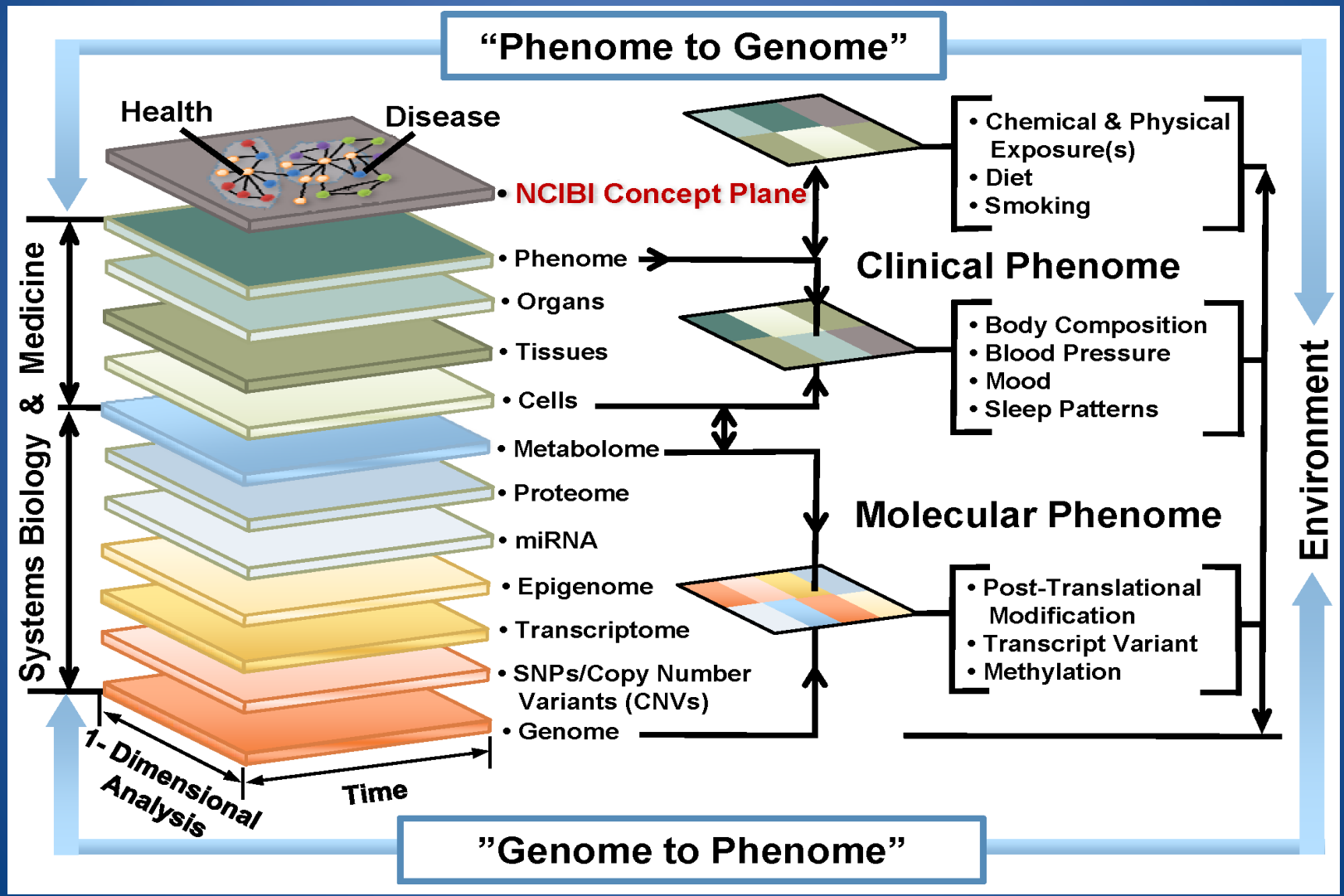

81-100 chr17
ERBB2
C17orf37
GRB7
IKZF3
ZPBP2
PSMD3
GSDMA
LOC728129
ORMDL3
GSDMB
CSF3
MED24
THRA
NR1D1
MSL1
RARA
CDC6
WIPF2
RAPGEFL1
CASC3
PTMs
Ser/Thr Acetyl GPMDB
Ser/Thr Phospho Unitprot
Met Acetyl GPMDB
Ser/Thr Phospho GPMDB
Lys Acetyl GPMDB
Tyr Phospho Unitprot
Tyr Phospho GPMDB
N-glyco Unipep
Gene
Genes found in Ensembl with MS information (GPMDB)
Number of pseudogenes


81-100
New
1
ERBB2
C17orf37
GRB7
IKZF3
ZPBP2
PSMD3
GSDMA
LOC728129
ORMDL3
GSDMB
1
CSF3
MED24
THRA
NR1D1
MSL1
RARA
CDC6
WIPF2
RAPGEFL1
CASC3
Gene
Antibody found in HPA or Antibodypedia
Gene
No Antibody in HPA and Antibodypedia
1
Number of pseudogenes
1
Number of hypothetical genes

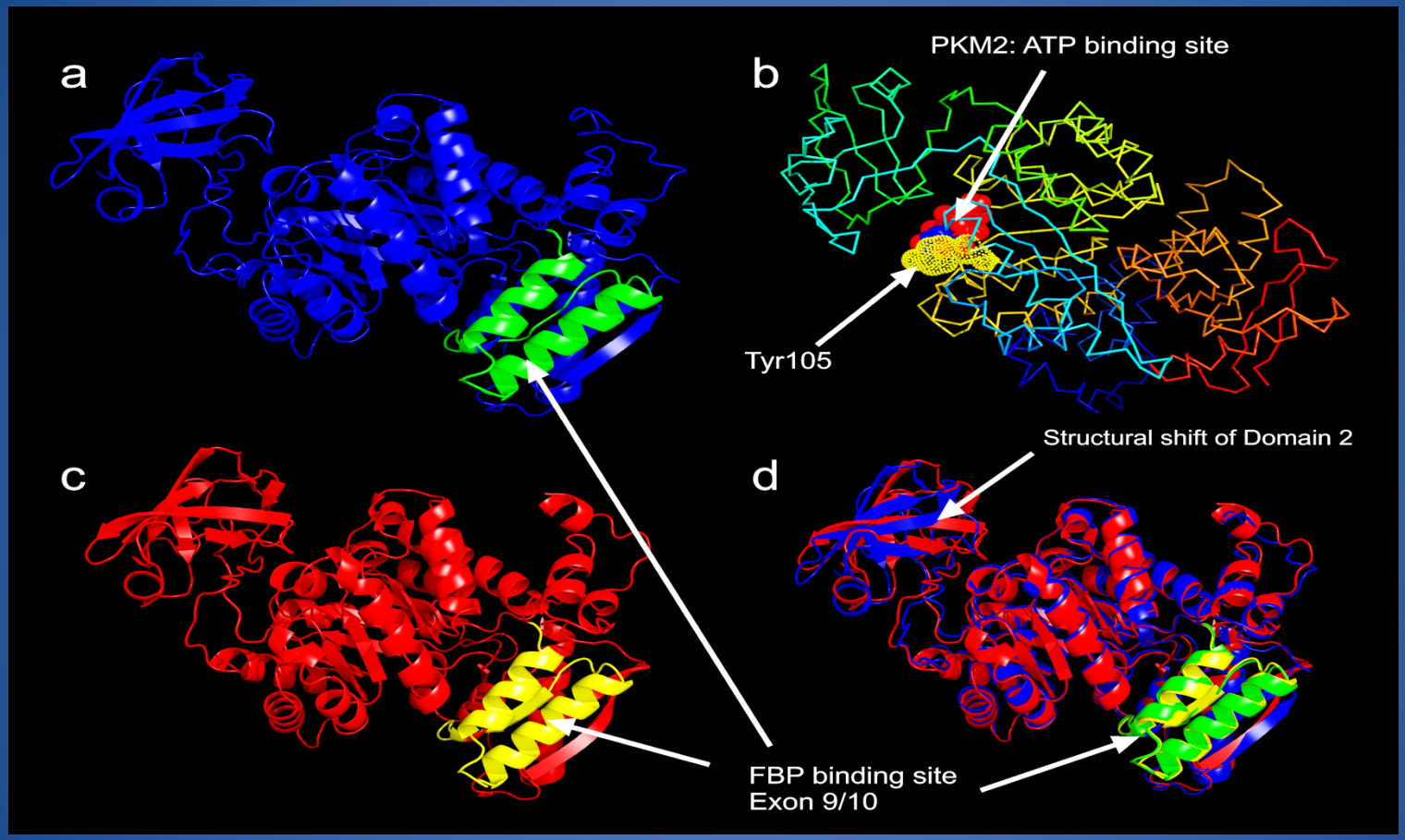
Vision of Biology and Pathology as Information Sciences: Key Components

- An avalanche of genomic information: validated SNPs, haplotype blocks, candidate genes/alleles, sequences, proteins, & metabolites—to be associated with disease risks
- Powerful computational methods
- Effective linkages with better environmental, dietary, and behavioral datasets for eco-genetic analyses
- Credible privacy and confidentiality protections in research and clinical care
- Breakthrough tests, vaccines, drugs, behaviors, and regulatory actions to reduce health risks and cost-effectively treat patients globally.

Integrating High-Throughput Measurements with the Phenotype: from Science to Medicine



PKM2 Structures Predicted with I-TASSER Threading Assembly (Y.Zhang)



PKM2 normal and cancer isoforms with exons 9 and 10 superimposed