Top Down Proteomics: Has It’s Time Now Come?

Neil L. Kelleher
Northwestern University
The Chicago Biomedical Consortium

World HUPO
Boston, MA - September 9th, 2012
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
The pace of development in genomics is breathtaking

Richard Resnick: Welcome to the genomic revolution
The pace of development in genomics is breathtaking

TEDxBoston 2011, Filmed Jul 2011; Posted Sep 2011

From TedX Boston- Richard Resnick
# The Human Genome Project: Rewind

<table>
<thead>
<tr>
<th></th>
<th>Aspect of Project</th>
<th>Human Genome Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Required Major Leap in Technology?</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Required Mapping Phase?</td>
<td>Yes (Genetic + Physical)</td>
</tr>
<tr>
<td>3</td>
<td>Body/Cell Context Matters?</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Target Size</td>
<td>$3 \times 10^9$ base pairs</td>
</tr>
<tr>
<td>5</td>
<td>Number of Donors</td>
<td>5 – 22 People</td>
</tr>
<tr>
<td>6</td>
<td>Model Systems?</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>Time</td>
<td>17 years</td>
</tr>
<tr>
<td>8</td>
<td>$$(Pilot Projects)$$</td>
<td>$15 B \ (\sim 10-100 \ M)$$</td>
</tr>
</tbody>
</table>
The Human Genome Project: Rewind

• Initial Phases of Project:
  – Genome Mapping
  – Technology Development

• Project Tenets: Fundamental Values
  – Normal, and NOT Disease Biology
  – Limited Population Sampling
  – Definition of Depth → cost vs. progress knowable
  – Was a Structural Project, not Functional (comes later)
The Human Genome

Humans are Diploid: each chromosome has one homologous partner

Giemsa staining
Initial Stages of the HGP

Genome Mapping and Development of Sequencing Technology
From Chromosome to Sequence
Sequencing at last...
The ABI 3700
The Engineering Ethos took over the Human Genome Project
Progress of the Public HGP

Figure 4 Total amount of human sequence in the High Throughput Genome Sequence (HTGS) division of GenBank. The total is the sum of finished sequence (red) and unfinished (draft plus predraft) sequence (yellow).
Concept of the Long Ball
The Human Genome Project

Public Project
Dr. Francis Collins, Director (2001)

Drs. Craig Venter and Claire Fraser of Celera (2001)
The pace of development in genomics is breathtaking

Cancer genome maps

The pace of development in genomics is breathtaking.
Executive Summary

- The Genomics Revolution: A Retrospective
  - Proteins as Measurement Targets

- Versions of the HPP (B/D- and C- HPP)

- Top Down Proteomics for Cataloging Protein Molecules Precisely
  - An Early Example → Human Histones

- Levels of Organization in the Human Body

- The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
Abundance range of protein molecules spans >1 million fold, and there is no way to amplify them.

- 10^7 copies/cell
- 50 copies/cell

~10,000 proteins in a single cell type (~1/2 that encoded by the genome)
From One Gene, Many Protein Forms: A Major Theme in Human Biology

DNA → mRNA → Protein

mutation → Alternative splicing → Covalent Modification

20,300 human genes

RNA messages

distinct forms of protein molecules
Origins of Complexity in the Human Proteome: The Age of Protein Isoforms

Key Concept: sources of protein variability result in a large, but finite number of protein forms, resulting in a vast measurement challenge.
The “Protein Inference” Problem (or the “Protein Isoform” Problem)

Proteoform: A Single Term to Capture Protein Complexity

With Lloyd Smith, University of Wisconsin
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus \( D_x \) and \( R_x \) Payoff
# Human Proteome Project(s)

## Abbreviations of Two Articulations of the HPP

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Project</th>
<th>Year Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>B/D-HPP</td>
<td>Biology/Disease-Based Human Proteome Project*</td>
<td>2002</td>
</tr>
<tr>
<td>C-HPP</td>
<td>Chromosome-Centric Human Proteome Project**</td>
<td>2010</td>
</tr>
</tbody>
</table>

(*also known as the Organ/Tissue-Based HPP)  
(**also known as the Gene-Centric HPP)
The human proteome project: Current state and future direction.


To the Editor:
The Chromosome-Centric Human Proteome Project (C-HPP) aims to define the full set of proteins encoded in each chromosome through development of a standardized approach for analyzing the massive proteomic data sets currently being generated from dedicated efforts of national and international teams. The initial goal of the C-HPP is to identify at least one representative protein encoded by each of the approximately 20,300 human genes\textsuperscript{1,2}. The proteins will be characterized for tissue localization and major isoforms, including post-translational modifications (PTMs), using quantitative mass spectrometry and antibody reagents. Our rationale is that effective integration of proteomics data into a genomic framework will lead to improved knowledge of complex biological systems and facilitate access to protein level data. Although the intent to engage in a C-HPP program has been noted\textsuperscript{1-3}, our objective here is to define the goals and process for its development as a multinational program.

Over the past three years, the Human utility for biological and disease studies. With development of new tools for in-depth characterization of the transcriptome and proteome, the HPP is well positioned to have a strategic role in addressing the complexity of human phenotypes. With this in mind, the HUPO has organized national chromosome teams that will collaborate with well-established laboratories building complementary proteotypic peptides, antibodies and informatics resources.

An important C-HPP goal is to encourage capture and open sharing of proteomic data sets from diverse samples to enhance a gene- and chromosome-centric display. This will display several layers of biological information on a common reference platform comparable to a genome browser. Such context will effectively integrate transcriptomics data such as RNA-Seq with proteomic data sets (Fig. 1).

Although the C-HPP program has some similarities to the Human Genome Project (HGP)\textsuperscript{4} in its quest for complete coverage across the genome, the C-HPP has the added challenge of characterizing machine database (GPMDB), UniProt and neXtProt (Supplementary Fig. 3).

The C-HPP does not propose any alteration in the work flow of a typical proteomics laboratory; instead, it seeks more effective use of data encompassed in existing bioinformatics resources, which will be combined with targeted studies to generate a robust list of observed protein

C-HPP

Year ~2022
(10 year time horizon)

The project will meet its aims when the comprehensive C-HPP database is 100% matched with the 20,300 protein-coding genes annotated on the human genome sequence, including at least one representative AST and nsSNP, tissue localization and three classes of PTMs in whole-chromosome sets (22 autosomal, X and Y; Supplementary Table 2).

Characterizing Proteins Precisely
(gene specific ID, splicing, modifications)

Splice Variants
and Modifications

Integrated Informatics

RNA-seq

Bottom Up Proteotypic Peptides

Quantitative Multitarget MS
Isobaric species of histone H3 K27-K36 peptide

xKxKxSA.PATGGVKKx3Kx2Kx1Kx1Kx1Sx2Kx2Kx2PHR
Kx2KxKx1Kx1Kx1Sx2Kx2Kx2PHR
Kx2KxKx2Sx2Kx2Kx2Kx2PHR
Characterizing Proteins Precisely (gene specific ID, splicing, modifications)

Top Down Proteomics

Proteoforms (Splice Variants and Modifications)

Integrated Informatics

Bottom Up Proteotypic Peptides

RNA-seq

Quantitative Multitarget MS
Isobaric species of histone H3 K27-K36 peptide
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
Top Down MS Solves the Protein Inference Problem

Intact mass determination and N- and C-terminal fragmentation differentiates highly similar protein forms

Abundance range of protein molecules spans ~1 million fold, and there is no way to amplify them.
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)
  – Top Down Proteomics for Cataloging Protein Molecules Precisely
  – Levels of Organization in the Human Body

• Early Example → Human Histones

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
Packaging of DNA into Chromatin

Top Down Mass Spectrometry of Human Histones

Nucleosome

6 Modifications Automatically Detected and Localized

Human Histone Analysis by
Top Down Mass Spectrometry (Q-FTMS)

1) Electron Capture Dissociation
2) Custom Database Search

LRDNIQGITKPAIRRLARRGGVK
RISGLIYEETRGVLKVFLENVIRD
AVTYTEHAKRKTVTAMDVVYAL
KRQGRTLYGF GG

+70  +112  +154  +196  +238

11264  11319  11429  11484

mass
For Histone H4

10^7 copies/cell

10^3 copies/cell

N-Acetyl and Lys20 dimethyl

75 Proteoforms

N-Acetyl, Arg3 dimethyl, and Lys20 dimethyl
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
The Levels of Organization in the Human Body

- **Organ/Tissue**
- **Cells**
- **Organelles**
- **Protein Complexes**
- **Protein Molecules**

**Key Concept:** Analysis of protein molecules can be done at selected levels in this hierarchy.
The Levels of Organization in the Human Body

Organ/Tissue → Cells → Organelles → Protein Complexes → Proteoforms

Key Concept: Analysis of protein molecules can be done at selected levels in this hierarchy.
The Levels of Organization in the Human Body

Organ/Tissue → Cells → Organelles → Protein Complexes → Proteoforms

Key Concept: Analysis of protein molecules can be done at selected levels in this hierarchy.
CELLPEDIA: a taxonomy and repository for human cell types (information on morphologies, gene expression, etc.)

Classification scheme
(1) physical locations + conventional taxonomy  
(2) cell differentiation pathways compiled from biomedical textbooks and journal papers

human differentiated cells
2718 taxonomy keys

stem cells
66 cell taxonomy keys

934 parent–child relationships reported in cell differentiation or transdifferentiation pathways are retrievable

http://cellpedia.cbrc.jp/cgi-bin/index.cgi
The Levels of Organization in the Human Body

Key Concept: Analysis of protein molecules can be done at selected levels in this hierarchy.

- Proteoforms: N ~ 250,000
- Protein Complexes: N ~ 4000
- Organelles
- Cells
- Organ/Tissue

Key Concepts:
The Cell-Based Human Proteome Project (CB-HPP)

1 Cell Type × Proteoforms = A Cellular Proteome

~4,000 Cell Types × 250,000 Proteoforms/Type = A Cell-Based Proteome Project (1,000,000,000,000 Proteoforms)
## Comparing the Genome Project and the CB-HPP

<table>
<thead>
<tr>
<th></th>
<th>Aspect of Project</th>
<th>Human Genome Project</th>
<th>Cell Based - Human Proteome Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Required (s) Major Leap in Technology?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Required (s) Mapping Phase?</td>
<td>Yes (Genetic + Physical)</td>
<td>Yes (Cell-based)</td>
</tr>
<tr>
<td>3</td>
<td>Body/Cell Context Matters?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Target Size</td>
<td>$3 \times 10^9$ base pairs</td>
<td>$1 \times 10^9$ proteoforms</td>
</tr>
<tr>
<td>5</td>
<td>Model Systems?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>microorganisms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Number of Donors</td>
<td>4 – 22 People</td>
<td>thousands</td>
</tr>
<tr>
<td>7</td>
<td>Time</td>
<td>17 years</td>
<td>15-20 years</td>
</tr>
<tr>
<td>8</td>
<td>$ (Pilot Projects)</td>
<td>15 B ($\sim$10-100 M)</td>
<td>? B ($\sim$10-100 M)</td>
</tr>
</tbody>
</table>
Questions: The Big Three

- How? Methods and implementation?
- How much?
- Why? Value of the CB-HPP transformative?
Questions: The Big Three

- How? Methods and implementation?
- How much?
- Why? Value of the CB-HPP transformative?
Cell-Specific Proteomics
General Experiment Schematic

Highly Sensitive Proteome Analysis of FACS-Sorted Adult Colon Stem Cells
Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum

CyTOF® Mass Cytometer: Single Cell Analysis

replace fluorophores and fluorescence ...

with metals and atomic mass spectrometry
Rediscovery of canonical signaling pathways validates method
Questions: The Big Three

• How? Methods and implementation?
• How much?
• Why? Value of the CB-HPP transformative?

...per proteoform

= $1 Billion
Executive Summary

• The Genomics Revolution: A Retrospective
  – Proteins as Measurement Targets

• Versions of the HPP (B/D- and C- HPP)

• Top Down Proteomics for Cataloging Protein Molecules Precisely
  – An Early Example → Human Histones

• Levels of Organization in the Human Body

• The Need for Disruption in Proteomics, Plus $D_x$ and $R_x$ Payoff
The pace of development in genomics is breathtaking

Richard Resnick: Welcome to the genomic revolution
Historical Progression of Mass Spectrometry

1901 → present
Organ/Tissue

Cells

Organelles

Protein Complexes

Proteoforms

Mapping intact protein isoforms in discovery mode using top-down proteomics

John C. Tran1,2, Leonid Zamdborg1, Dorothy R. Ahlf1,2, Ji Eun Lee1,3, Adam D. Catherman1,2, Kenneth R. Durbin1,2, Jeremiah D. Tipton1, Adaikalam Vellaichamy1, John F. Keil1,2, Mingxi Li1,2, Cong Wu1, Steve M. M. Sweet1,2, Bryan P. Early1,2, Nertila Siuti1, Richard D. LeDuc1, Philip D. Compton1, Paul M. Thomas1,2 & Neill L. Kelleher1,2

Mapping intact protein isoforms in discovery mode using top-down proteomics
Top Down Proteomics of >1000 Proteins

- Organ/Tissue
- Cells
- Organelles
- Protein Complexes
- Proteoforms

Published Oct. 30, 2011
Top Down Proteomics of >1000 Proteins and >3000 Proteoforms

Published Oct. 30, 2011

Organ/Tissue

Cells

Organelles

Protein Complexes

Proteoforms

Mono- / Di-Phosphorylation
Trimethylation / Acetylation
Mono- / Di-Methylation

Retention Time (min)

Isoelectric Point (pI), from sIEF

LC Retention Time (each box): 15 min → 55

Protein Mass (kDa)

MW Range (kDa), from GEL/GE

1) GRP 78: 70650 Da E-value = 10^-3
2) Mitaxin: 70905.7 Da E-value = 10^-4
3) PCK 1: 13818.0 Da E-value = 10^-7

Phosphorylations

Methylation

Da

Δ 80 Da

Δ 80 Da

Δ 14 Da
Progression of “Top Down” and FT Mass Spectrometry

1999 — 2012
Transformation Requires Innovation

Sanger Sequencing
1977 \(\rightarrow\) 2003

Next Generation Sequencing
1996 \(\rightarrow\) Today

10^5 – 10^6 bases / day

10^9 – 10^{10} bases / day
Transformation Requires Innovation

Mass Spectrometry
2012

2020
Transformation Requires Innovation

Mass Spectrometry
2012

?
Small Steps: Easy to use, high performance nanoLC-MS

Complex
Specialized expertise

Simple
Universal productivity

PicoChip™ and Stage on a Q Exactive
Top Down Proteomics: Faster and Cheaper

In House PLRPS

87 Unique Accession Numbers (p<1E10)

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P14927</td>
<td>Cytochrome b-c1 complex subunit 7</td>
</tr>
<tr>
<td>P14406</td>
<td>Cytochrome c oxidase subunit 7A2</td>
</tr>
<tr>
<td>O43677</td>
<td>NADH dehydrogenase 1 subunit C1</td>
</tr>
<tr>
<td>P56134</td>
<td>ATP synthase subunit f</td>
</tr>
<tr>
<td>Q9P0S9</td>
<td>Transmembrane protein 14C</td>
</tr>
<tr>
<td>Q9P0U1</td>
<td>Mitochondrial import receptor subunit TOM7</td>
</tr>
</tbody>
</table>

PicoCHIP PLRPS

with Gary Valaskovic
Questions: The Big Three

• How? Methods and implementation?
• How much?
• Why? Value of the CB-HPP transformative?
Primary Outcomes of the CB-HPP

- A clear taxonomy of human cell types and their natural variation
- Technologies and reagents to define, sort, and in-situ image cell types
- Technologies for “next-generation” proteomics
- A reference list of proteoforms within all cell types
...Only about 25 percent of men who have prostate biopsy due to an elevated PSA level actually have prostate cancer ~National Cancer Institute (using older PSA testing)
Many Proteoforms Confuse PSA Testing

Over 80 proteoforms possible with known modifications alone
Top Down alone can link these together!

“We have to do the best we can..., and keep working to learn more.”
~Dr. Catalona, Northwestern University

Complement New Testing: free-PSA, PSA velocity, PSA density, pro-PSA-based phi Test, PCA3 urine testing

Data Courtesy of Rosa Viner and Colleagues, Thermo Fisher Scientific
A Cell-Based Approach to the Human Proteome Project

1 Cell Type

Proteoforms

A Cellular Proteome

\(~4,000\) Cell Types

250,000 Proteoforms/Type

A Cell-Based Proteome Project

\((1,000,000,000\) Proteoforms)
Acknowledgements

• Kelleher Laboratory

• Funding: Northwestern University, NIH GM 067193, and the Chicago Biomedical Consortium
Sponsors and Supporters of the Kelleher Group at Northwestern

NORTHWESTERN UNIVERSITY

FEINBERG SCHOOL OF MEDICINE

CHICAGO BIOMEDICAL CONSORTIUM
CBC is funded by the Searle Funds at The Chicago Community Trust

NORTHWESTERN UNIVERSITY

WEINBERG COLLEGE OF ARTS & SCIENCES

Robert H. Lurie Comprehensive Cancer Center
OF NORTHWESTERN UNIVERSITY

Chemistry of Life Processes Institute

Thermo SCIENTIFIC
Thank You.
Consortium for Top Down Proteomics (CTDP)

Mission Statement

To promote innovative research, collaboration and education accelerating the comprehensive analysis of intact proteins in complex systems.

http://www.topdownproteomics.org/
# Proteoform Repository

on Wednesday, 14 March 2012.

## Protein Proteoforms, 075531 (075531)

<table>
<thead>
<tr>
<th>Proteoform Id</th>
<th>Sequence</th>
<th>Description</th>
<th>Monoisotopic Mass (Da)</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID_00003D</td>
<td><strong>TSQKHDVFVAEMGKEKPVGSLAGIGEVLGKKEERGFDKAYVVLQFYLVKDEDFLREWLDKTCGANAKQSRDCFGCLESWCDAL</strong></td>
<td>Barrier-to-autointegration factor</td>
<td>10,042.96</td>
<td></td>
</tr>
<tr>
<td>PID_00003E</td>
<td><strong>TSQKHDVFVAEMGKEKPVGSLAGIGEVLGKKEERGFDKAYVVLQFYLVKDEDFLREWLDKTCGANAKQSRDCFGCLESWCDAL</strong></td>
<td>Barrier-to-autointegration factor</td>
<td>9,963.00</td>
<td></td>
</tr>
</tbody>
</table>

From Gene Sequence to Traits and Treatment of Complex Disease

Human Genome Sequences

Complex Human Disease

Phenotypic Variation

Drugs & Diagnostics

Kidney Cancer

Ac Me P1 X C

83
# Abbreviations for Versions of the Human Proteome Project

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Project</th>
<th>Year Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>B/D-HPP</td>
<td>Biology/Disease-Based Human Proteome Project*</td>
<td>2002</td>
</tr>
<tr>
<td>C-HPP</td>
<td>Chromosome-Centric Human Proteome Project**</td>
<td>2010</td>
</tr>
<tr>
<td>CB-HPP</td>
<td>Cell-Based Human Proteome Project</td>
<td>2012</td>
</tr>
</tbody>
</table>

(*also known as the Organ/Tissue-Based HPP)  
(**also known as the Gene-Centric HPP)