Les applications de la protéomique à haut-débit pour l'étude des interactions hôte-microbiôme chez l’humain

A E. Starr¹, X. Zhang¹, R Chen¹, S. A. Deeke¹, H. Zhou², D. Mack³, A. Stinzi¹ and D. Figeys¹

¹ Ottawa Institute of Systems Biology, University of Ottawa
² Shanghai Institute of Materia Medica, China
³ Children Hospital of Eastern Ontario
Technology Development
Proteomic Reactor
Sample processing
Lipidomic approach
Post translational modifications
Microfluidic devices

Biological applications
Lysine methyl transferases
Lipid metabolism
Brain and neurodegeneration
Inflammatory Bowel Disease
Can “big data” based on “omics” deliver for clinical discovery?

Are tools in place to mine multi-omics “big data”?
Inflammatory Bowel Disease: Crohn’s Disease and Ulcerative Colitis

Crohn’s Disease
Ulcerative colitis

200000 Canadians with IBD
1.4 million Americans with IBD
2.2 million Europeans with IBD

Ng et al, Gut 2013
Pediatric patients: a naive system with different challenges

Environmental risk factors of IBD

CD
UC

A multi-omics approach to IBD research

Biopsy → Lavage

Lavage

- Supernatant
  - Quantitative Proteomics
  - Glycoproteomics
  - N-terminomics
  - Metabolomics

- Bacterial Pellet
  - Metaproteomics
  - Metagenomics
  - Metatranscriptomics
  - 16S sequencing

Biopsy

- Quantitative Proteomics
- Epigenetics

Extracellular vesicles

- Proteomics
- Transcriptomics
Gut microbiota as a global research interest

- **Second genome of human**
  - 9.9 million bacterial genes
    (~25,000 human genes)
  - *Diverse metabolic activity*
    - Metabolic/Nutrition/Immunity
  - *Close relation with health*
    - *Intestinal diseases*
    - *Non-intestinal diseases*
      - metabolic syndrome
      - Cardiovascular diseases
Meta-proteomics: a complex problem

http://unipept.ugent.be/

https://en.wikipedia.org/wiki/Phylogenetic_tree
Meta-proteomics: large matrix of proteins and species
Bottom-up proteomics

Complex protein mixture → Even more complex peptide mixture → Peptide identity and intensity

Which proteins did the peptides come from??

http://www.strgen.org/proteome/
Meta-proteomics: expended space when dealing with peptides

http://unipept.ugent.be/
Proteomics, 2015, 15 (8), pp 1437–1442
Meta-proteomics:
*expanded space when dealing with peptides*
Metaproteomics can identify proteins all the way to the species level.
Host-Microbiota interactions
A multidimensional data space
Host-Microbiota interactions in IBD
A multi-omics approach to IBD research

Biopsy

Lavage

Supernatant

Bacterial Pellet

Quantitative Proteomics
Glycoproteomics
N-terminomics
Metabolomics

Quantitative metaproteomics
Metagenomics
Metatranscriptomics
16S sequencing

Extracellular vesicles

Proteomics
Transcriptomics
Samples obtained at diagnosis: IBD and non-IBD control patients

Biopsies from macroscopically inflamed and non-inflamed areas

Number of Patients

Control | CD | UC

N=31 | N=40 | N=28

Age (years)

Control | CD | UC
A proteomics approach to identify subtype biomarkers

FASP Tryptic Digest
SCX Fractionation; C18-desalting
LC MS/MS
MaxQuant, Perseus
Matlab, Rocct, R

Validation of Biomarkers

Quantitative Proteomics of biopsies

> 3 year recruitment

> 400 hours processing

> 2000 hours MS

> 240 hours search of 48 MB database

Ongoing bioinformatics

138 patients

166 biopsies

741 fractions

2,123,027 MS

25,883,537 MS/MS

248,000,000 Peaks

> 300 GB data

48735 Peptide sequences

4767 Protein groups

> 30 GB biomarker-associated data
Candidate biomarker identification methodology

4226 Proteins Quantified

Principle Component Analysis

Explore:
Partial Least Squares Discriminant Analysis

Control vs IBD
Identify Set 1 of Potential Biomarkers
Test “Unknowns” → Validation

CD vs UC
Identify Set 2 of Potential Biomarkers
Test “Unknowns” → Validation
We can differentiate Control from IBD by PCA

PCA of 1353 proteins

Trend towards increase severity
90% of “Unknown” cases accurately separated between control and IBD

98% accurately diagnosed with 15 proteins

98% accurate by cross-validation

90% accurate classification of “Unknowns”
Proteins elevated in IBD patients vs non-IBD control patients

Inflammatory proteins

- Superoxide Dismutase
- IL-25

Novel IBD biomarkers

- Calumenin
- Leucine Aminopeptidase 3

Adult IBD biomarkers

- S100A8
- S100A9

GAPDH
Proteomic identified biomarkers validated for pediatric IBD diagnosis

Calumenin

Leucine Aminopeptidase 3 (LAP3)

Creatine Kinase B-chain (B-CK)

GAPDH
85% of IBD cases were accurately predicted as CD or UC

91% accurately classified with 15 proteins

81% accurate by cross-validation
85% accurate classification of “Unknowns”
Severity of the diseases:
Proteins that differentiate IBD from controls

\[ r^2 = 0.3532 \]
\[ p = 0.0195 \]

\[ r^2 = 0.5186 \]
\[ p = 0.0025 \]
Severity of the disease protein that differentiate Crohn’s Disease and Ulcerative Colitis

- $r^2 = 0.5975$, $p = 0.0007$
- $r^2 = 0.2791$, $p = 0.0354$

ELISA

- $r^2 = 0.705$, $p = 0.018$
Integrating proteomics and metagenomics

~100 CD patients (mild, moderate, severe)

Metagenomics (microbes)
>1 million 16S rDNA reads per sample
Total of 4961 OTUs (operational taxonomic unit)

Proteomics (host)
3323 proteins identified
Step 1: Differentially abundant OTUs across CD severities

4961 OTUs

Linear mixed models

161 OTUs

Phylogenetic tree of the differentially abundant OTUs (fold change ≥2 and P<0.05); an increasing red intensity indicates OTUs whose relative abundance increased, whereas an increasing blue intensity indicates OTUs whose relative abundance decreased in CD patients with severe inflammation as compared to mild (outer circle) or severe as compared to moderate (inner circle).
Step 2: Identification of differentially abundant proteins

- 526 proteins were found to be differentially expressed (CD vs control)
- Mitochondrial proteins were identified as the major discriminant feature (129 differentially expressed mitochondrial proteins)
Step 3: “transkingdom” correlation analysis

Kendall correlation

35 OTUs
Step 4: Hypothesis generation

Data reduction:

4661 OTUs → 161 OTUs → 35 OTUs → 4 OTUs
Step 5: Hypothesis testing

+/- A. parvulun Colitis?

Germ-free Il10^-/- mice

SPF + A. parvulum

SPF + A. parvulum

SPF A. parvulum
## Expanding raw data: Current estimation of the IBD project

<table>
<thead>
<tr>
<th>Biopsy</th>
<th>Estimated Raw Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative Proteomics</td>
<td>1 TB</td>
</tr>
<tr>
<td>50 MB database</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supernatant</th>
<th>Estimated Raw Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative Proteomics</td>
<td>200 GB</td>
</tr>
<tr>
<td>Glycoproteomics</td>
<td>200 GB</td>
</tr>
<tr>
<td>N-terminomics</td>
<td>200 GB</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>? GB</td>
</tr>
<tr>
<td>50 MB database</td>
<td></td>
</tr>
<tr>
<td>PTM increase search times</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bacterial Pellet</th>
<th>Estimated Raw Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metaproteomics</td>
<td>300 GB</td>
</tr>
<tr>
<td>Metatranscriptomics</td>
<td>500 GB</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>500 GB</td>
</tr>
<tr>
<td>16S sequencing</td>
<td>40 GB</td>
</tr>
<tr>
<td>1GB database</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extracellular vesicles</th>
<th>Estimated Raw Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteomics</td>
<td>1 GB</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>10 GB</td>
</tr>
</tbody>
</table>

- **Quantitative Proteomics**: The estimated raw data is 1 TB. The database size is 50 MB. **Glycoproteomics**: The estimated raw data is 200 GB. **N-terminomics**: The estimated raw data is 200 GB. **Metabolomics**: The estimated raw data is unknown. **Metaproteomics**: The estimated raw data is 300 GB. **Metatranscriptomics**: The estimated raw data is 500 GB. **Metagenomics**: The estimated raw data is 500 GB. **16S sequencing**: The estimated raw data is 40 GB. **Proteomics**: The estimated raw data is 1 GB. **Transcriptomics**: The estimated raw data is 10 GB. **PTM increase search times**. **uOttawa**
Can “big data” based on “omics” deliver for clinical discovery? ✔

Are tools in place to mine multi-omics “big data”? 😐
Acknowledgements

People:
- P. Adler
- A. Blanchard
- R. Chen
- K. Cheng
- C.K. Chiang
- K. Chu
- S. Deeke
- J. Mayne
- Z. Ning
- J. Moore
- D. Seebun
- A. Star
- A. Starr
- M. Weng
- B. Xu
- H. Xu
- X. Zhang