A meta-omics workflow to generate a novel host and microbial biomarker panel for early detection of ovarian cancer

Subina Mehta

Galaxy for Proteomics (Galaxy-P) team

University of Minnesota

Kristin Boylan, Ashley Petersen, Amy Skubitz, Pratik Jagtap, Timothy Griffin
Introduction

OCRA
ovarian cancer research alliance

woman's lifetime risk of developing OC is 1 in 78

5-year survival rate
Stage 1- 92.4%
Stage 2 or 3 - 71%

Diagnostic Biomarkers
CA-125
HE4

Dysbiosis in female reproductive tract linked to Ovarian cancer development

MS-based Bioinformatic analysis indicates pap test fluid (PTF) samples are a rich source of microbial and human proteins

PAP-TEST FLUID AS SOURCE?

Microbial attribute?
Hypothesis

Microbiome-expressed biomarkers alongside host for early detection?

Analysis

Aim to use Galaxy platform to investigate and generate host-microbial biomarker panel for early detection of ovarian cancer

Source

PAP-TEST FLUID AS SOURCE?
Database Generation

137 taxa

~6.75M Protein Sequences

Metanovo

~38K Protein Sequences

Microbial + Human SwissProt + Contaminants

~57K Protein Sequences

Matthys G Potgieter et al. (2022) - https://www.biorxiv.org/content/10.1101/605550v7
Discovery Workflow

**DATABASE**
- ~57K protein sequences

**96 MSMS datasets**
- 4Plex-24 fraction

**Database Searching**

- **SearchGUI**
- **PeptideShaker**
- **MaxQuant**
- **Fragpipe**

**5599 microbial peptides**
5599 MICROBIAL PEPTIDES


verified microbial peptides

HUMAN + contaminants

PepQuery*

5599 microbial peptides

DeepVenn (C) 2020 Tim Hulsen

FragPipe

2220

216

384

1621

26

538

596

Search GUI

Peptide Shaker

MaxQuant

FragPipe

Verification Workflow

5,599 Peptides → PepQuery → 3,887 Peptides → Extract Protein sequences → 2,044 Protein sequences → Human + Microbial database+ contaminants ~83k protein sequences → CONTAMINANTS → Human UniProt → Quantitation workflow
Quantitation workflow

DATABASE ~83K protein sequences

96 MSMS datasets 4Plex-24 fraction

MaxQuant

Quantified Peptides and Proteins

4779 Human Proteins

731 Microbial Proteins
Statistical Analysis

01 Normalization
- Limma based
- Log ratios - median normalized

02 Data filtering criteria
- Observed in \( \frac{3}{4} \) plex’s
- 10% non plex missingness

03 Performance Measurement
- ROC curve

04 Data Quality
- Missingness observed
- OVCA-Benign

[Diagram showing box plots and ROC curve]

Ashley Petersen, PhD
## Microbial Protein Abundance Results

### Differentially abundant

<table>
<thead>
<tr>
<th></th>
<th>OVCA–H</th>
<th>OVCA–B</th>
<th>Benign–H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microbial Peptides</td>
<td>104</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Microbial Proteins Groups</td>
<td>26 (16↓ and 8↑)</td>
<td>6 (4↓ and 2↑)</td>
<td>3 (2↓ and 1↑)</td>
</tr>
</tbody>
</table>

### Graphs

- **OVCA vs Healthy**
- **OVCA vs Benign**
- **Benign vs Healthy**
Microbial Protein Results

**OVCA-Healthy**
- Surface layer Protein A domain containing protein
  - *Lactobacillus crispatus*
  - HTH cro/C1-type domain containing protein
  - *Clostridiales bacterium*

**OVCA-Benign**
- DUF4879 domain containing protein
  - *Bacteroides thetaiotaomicron*
- Thiol-disulfide interchange protein Dsb-D
  - *Escherichia coli*

**Benign-Healthy**
- N-acetyleneuraminate lyase
  - *Lactobacillus crispatus*
- Phage tail tape measure protein
  - *Enterococcus faecalis*
Human proteins

Differentially abundant Proteins Groups

<table>
<thead>
<tr>
<th></th>
<th>OVCA-H</th>
<th>OVCA-B</th>
<th>Benign-H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins Groups</td>
<td>167</td>
<td>107</td>
<td>39</td>
</tr>
</tbody>
</table>
Conclusion

01 Developed MS-based Galaxy-driven bioinformatics workflow for processing of microbial and host proteins, generating verified microbial peptide candidates suitable for targeted analysis within individual patient samples.

02 Preliminary results show that microbial proteins along with host proteins could improve early detection.

03 Predictive analysis looks promising and plan to follow up with more studies.
### Ongoing Plans

<table>
<thead>
<tr>
<th></th>
<th>Literature survey database</th>
<th>16S rRNA based database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxa identified (genera and species incl)</td>
<td>137</td>
<td>192</td>
</tr>
<tr>
<td>Comprehensive protein sequence database</td>
<td>~6.75 million protein sequences</td>
<td>~23.2 million protein sequences</td>
</tr>
<tr>
<td>After Metanovo (incl Human SwissProt and contaminants)</td>
<td>~57 thousand protein sequences</td>
<td>~142 thousand protein sequences</td>
</tr>
<tr>
<td>Database search microbial Peptides</td>
<td>5599 microbial peptides</td>
<td></td>
</tr>
</tbody>
</table>

Verification, Quantification and statistical analysis to follow.....
ACKNOWLEDGMENTS

University of Minnesota
Timothy Griffin
Pratik Jagtap
Monica Kruk
Andrew Rajczewski
Danielle Weise
Katherine Do

Minnesota Supercomputing Institute
James Johnson
Reid Wagner
Alexey Nesvizhskii
Fengchao Yu
University of Michigan

Valerie Schiml
Magnus Øverlie Arntzen
NMBU, Oslo, Norway

Biologists / collaborators
Chris Wendt
Teresa Laguna
Maneesh Bhargava
Ryan Hunter
David Largaespada

University of Minnesota
Amy Skubitz
Kristin Boylan
Ashley Petersen

University of Freiburg,
Freiburg, Germany

Matt Chambers
Nashville, TN

Netherlands

Alessandro Tanca
Italy

Magnum

University of Freiburg,
Freiburg, Germany

Anshu Bharadwaj
Chandigarh
India

Surbhi Bihani
Sanjeeva Srivastava
IIT, Mumbai
India

Saskia Hiltemann

Minneapolis

Valdemir Carvalho
Fleury Group, Brazil

Thilo Muth
Germany

Valerie Schiml

Minnesota Ovarian Cancer Alliance

Jean Armengaud
France

http://galaxyp.org/contact/
galaxyp.org
twitter.com/usegalaxyp

Biologists / collaborators
Bing Zhang
Bo Wen
Baylor College of Medicine

Michael Shortreed
UW-Madison

Björn Gruening
Melanie Foell
University of Freiburg,
Freiburg, Germany

Brook Nunn
U of Washington

Surbhi

University of Michigan

Jagroop Pandhal

Maria Doyle
Melbourne, Australia

Alexey Nesvizhskii
Fengchao Yu

Thilo Muth
Germany
Unipept Analysis