Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease

Brandi Cantarel
IHMC
March 11, 2011
• **Background**
  - What is Crohn’s Disease?
  - What is our study?
  - What is our experimental design?

• **Findings**
  - Taxonomic Community (WGS vs 16S)
  - Functional Profile (WGS)
  - Highly Abundant Proteins (Proteomics)
Crohn’s Disease Causes
Intestinal Inflammation

- 0.1% of western population
- About 1/2 million Americans
- Ileum and colon
- Imbalanced immune response
- No cure
- Varying treatment options
- No good monitoring tools
- Dietary restrictions
- Problems with nutrient absorption can lead to malnutrition
- Causes are genetic and environmental factors
This Study Aims to Identify Associated Genes

• **Hypothesis:**
  - Microbes in the intestinal tract play a role in Crohn’s disease incidence and/or development.
    - Dysbiosis of the gut microbiota (Seksik et al., 2005; Tamboli et al., 2004).

• **Aim:**
  - To identify microbial signatures associated with Crohn’s disease
    - Difficult to assess due to individual variation of gut flora
Twins were studied to control for Human Genetic Background

- Identical genetic makeup
- Presumably similar environmental exposures and dietary habits in childhood
- Similar gut microbiota
CSI: Human Gut
“Omics pipeline” provides different levels of information

Existing data for the Swedish twin Cohort

16S rRNA gene pyrosequencing
Willing et al., 2010. Gastroenterology.

Metabolite profiling

16S Metagenome
WGS Metagenome
Meta-Transcriptome
Meta-Proteome
Metabolome
Metagenomics vs Metaproteomics

- Genomics/Metagenomics
  - Information on gene content
  - Assessment of functional capability
- Proteomics/Metaproteomics
  - Evidence of gene expression under a given condition
  - Assessment of active functional processes
“Omics” Pipeline

Human Fecal Samples → Density Centrifugation to Extract Bacterial Cells → Protein Extraction for Mass Spectrometry → Protein Digestion → 2D LC-MSMS → Filter union → SEQUEST Search → Details See Poster #20

DNA Extraction → Genomic DNA → 454 Sequencer

Metagenomic Annotation Pipeline → Protein Database

Human Fecal Samples → DNA Extraction → Genomic DNA → 454 Sequencer → Metagenomic Annotation Pipeline → Protein Database
• **Background**
  - What is Crohn’s Disease?
  - What is our study?
  - What is our experimental design?

• **Findings**
  - Taxonomic Community (WGS vs 16S)
  - Functional Profile (WGS)
  - Highly Abundant Proteins (Proteomics)
### Metagenomic Sequencing Data

- **1** healthy twin pair
- **1** concordant colonic CD
- **2** discordant ileal CD
- **2** concordant ileal CD

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of BP (M)</th>
<th>Number of Reads (K)</th>
<th>Number of Singletons (K)</th>
<th>Number of Contigs</th>
<th>BP in Contigs (M)</th>
<th>Average Contig Length</th>
<th>Number of Contigs &gt; 500BP</th>
<th>Fraction Singleton</th>
<th>Fraction Contigs &gt; 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>10a</td>
<td>502</td>
<td>1,211</td>
<td>82</td>
<td>36,328</td>
<td>49</td>
<td>1337</td>
<td>24,726</td>
<td>0.0681</td>
<td>0.6806</td>
</tr>
<tr>
<td>10b</td>
<td>541</td>
<td>1,327</td>
<td>140</td>
<td>30,642</td>
<td>39</td>
<td>1259</td>
<td>20,219</td>
<td>0.1053</td>
<td>0.6598</td>
</tr>
<tr>
<td>15a</td>
<td>283</td>
<td>1,358</td>
<td>256</td>
<td>38,046</td>
<td>42</td>
<td>1109</td>
<td>25,909</td>
<td>0.1887</td>
<td>0.6810</td>
</tr>
<tr>
<td>15b</td>
<td>279</td>
<td>1,270</td>
<td>272</td>
<td>36,871</td>
<td>36</td>
<td>981</td>
<td>21,555</td>
<td>0.2146</td>
<td>0.5846</td>
</tr>
<tr>
<td>16b</td>
<td>148</td>
<td>1,126</td>
<td>511</td>
<td>46,845</td>
<td>28</td>
<td>591</td>
<td>23,857</td>
<td>0.4541</td>
<td>0.5093</td>
</tr>
<tr>
<td>18a</td>
<td>221</td>
<td>1,341</td>
<td>215</td>
<td>26,803</td>
<td>27</td>
<td>1014</td>
<td>17,971</td>
<td>0.1604</td>
<td>0.6705</td>
</tr>
<tr>
<td>18b</td>
<td>143</td>
<td>955</td>
<td>372</td>
<td>28,631</td>
<td>19</td>
<td>671</td>
<td>13,862</td>
<td>0.3898</td>
<td>0.4842</td>
</tr>
<tr>
<td>6a</td>
<td>280</td>
<td>1,901</td>
<td>299</td>
<td>55,650</td>
<td>66</td>
<td>1195</td>
<td>39,318</td>
<td>0.1570</td>
<td>0.7065</td>
</tr>
<tr>
<td>9a</td>
<td>245</td>
<td>782</td>
<td>155</td>
<td>19,893</td>
<td>21</td>
<td>1053</td>
<td>13,439</td>
<td>0.1988</td>
<td>0.6756</td>
</tr>
<tr>
<td>9b</td>
<td>290</td>
<td>845</td>
<td>197</td>
<td>35,571</td>
<td>27</td>
<td>754</td>
<td>20,011</td>
<td>0.2328</td>
<td>0.5626</td>
</tr>
</tbody>
</table>

- Patients in Remission
## MetaProteomics Metrics

### Matched MetaGenomes, Human Reference DB

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Sample</th>
<th>MS Run</th>
<th>Proteins</th>
<th>Peptides</th>
<th>Total Identified MS/MS</th>
<th>Total Collected MS/MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>healthy</td>
<td>6a</td>
<td>run 2</td>
<td>2,315</td>
<td>9,110</td>
<td>15,724</td>
<td>78,381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 3</td>
<td>2,009</td>
<td>8,005</td>
<td>15,452</td>
<td>70,492</td>
</tr>
<tr>
<td></td>
<td>6b</td>
<td>run 2</td>
<td>1,385</td>
<td>5,548</td>
<td>11,300</td>
<td>80,797</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 1</td>
<td>1,413</td>
<td>5,796</td>
<td>11,258</td>
<td>81,202</td>
</tr>
<tr>
<td>colonic CD</td>
<td>9a</td>
<td>run 1</td>
<td>871</td>
<td>3,738</td>
<td>7,628</td>
<td>92,865</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 2</td>
<td>829</td>
<td>3,381</td>
<td>7,879</td>
<td>91,745</td>
</tr>
<tr>
<td></td>
<td>9b</td>
<td>run 1</td>
<td>723</td>
<td>2,551</td>
<td>5,153</td>
<td>82,546</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 2</td>
<td>728</td>
<td>2,532</td>
<td>5,074</td>
<td>84,089</td>
</tr>
<tr>
<td>ileal CD</td>
<td>10a</td>
<td>run 3</td>
<td>1,089</td>
<td>4,474</td>
<td>8,891</td>
<td>90,775</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 1</td>
<td>1,049</td>
<td>4,286</td>
<td>8,244</td>
<td>91,145</td>
</tr>
<tr>
<td></td>
<td>10b</td>
<td>run 1</td>
<td>1,118</td>
<td>4,078</td>
<td>8,057</td>
<td>75,873</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 2</td>
<td>1,141</td>
<td>3,984</td>
<td>8,546</td>
<td>74,574</td>
</tr>
<tr>
<td>ileal CD</td>
<td>15a</td>
<td>run 1</td>
<td>946</td>
<td>3,913</td>
<td>9,276</td>
<td>83,254</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 2</td>
<td>1,183</td>
<td>4,862</td>
<td>10,616</td>
<td>77,906</td>
</tr>
<tr>
<td></td>
<td>15b</td>
<td>run 2</td>
<td>787</td>
<td>3,890</td>
<td>8,647</td>
<td>81,970</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 1</td>
<td>769</td>
<td>3,421</td>
<td>8,619</td>
<td>80,718</td>
</tr>
<tr>
<td>ileal CD</td>
<td>16a</td>
<td>run 2</td>
<td>369</td>
<td>1,716</td>
<td>4,079</td>
<td>78,811</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 1</td>
<td>407</td>
<td>1,844</td>
<td>4,012</td>
<td>73,878</td>
</tr>
<tr>
<td>healthy</td>
<td>16b</td>
<td>run 2</td>
<td>1,248</td>
<td>4,581</td>
<td>11,659</td>
<td>92,460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 3</td>
<td>1,256</td>
<td>4,416</td>
<td>11,914</td>
<td>91,865</td>
</tr>
<tr>
<td>ileal CD</td>
<td>18a</td>
<td>run 2</td>
<td>687</td>
<td>3,233</td>
<td>7,951</td>
<td>76,489</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 1</td>
<td>654</td>
<td>2,733</td>
<td>7,281</td>
<td>78,661</td>
</tr>
<tr>
<td>healthy</td>
<td>18b</td>
<td>run 1</td>
<td>794</td>
<td>3,018</td>
<td>7,640</td>
<td>92,795</td>
</tr>
<tr>
<td></td>
<td></td>
<td>run 2</td>
<td>829</td>
<td>3,245</td>
<td>7,442</td>
<td>93,485</td>
</tr>
</tbody>
</table>

Protein counts (not clusters); Non-Redundant peptide and spectra counts
Changes in Microbiome Community in Crohn’s Subjects (WGS)

Willing et al. 2010. Gastroenterology
Predicted ORFs are Clustered to Identify Differences

Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Intracellular trafficking, secretion, and vesicular transport
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
Broad Functional Categories of Metagenomic Data

Mean Relative Count

- RNA processing and modification
- Chromatin structure and dynamics
- Cell wall/membrane/envelope biogenesis
- Cell motility
- Cytoskeleton
- Extracellular structures
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme transport and metabolism
- Lipid transport and metabolism
- Inorganic ion transport and metabolism
- Secondary metabolites biosynthesis, transport and catabolism
COG classification of proteins that differ in relative amounts in ICD compared to healthy

- Cytoskeleton
- Defense
- Intracellular trafficking
- Signal transduction
- Unknown
- General Function
- Secondary Metabolites
- Inorganic Metab.
- PTM, Turnover
- Cell motility
- Cell wall/membrane
- Replication
- Transcription
- Translation
- Lipid metab.
- Coenzyme metab.
- Carbohydrate metab.
- Neucleotide Metab.
- Amino Acid Metab.
- Cell cycle
- Energy production
- Chromatin structure
- RNA processing

Healthy
ICD
Genes vs Proteins

Metabolism
Cell Cycle

<table>
<thead>
<tr>
<th>Proteins</th>
<th>ileal CD</th>
<th>18a</th>
<th>run 2</th>
<th>687</th>
<th>3,233</th>
<th>7,951</th>
<th>76,489</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>run 1</td>
<td>654</td>
<td>2,733</td>
<td>7,281</td>
<td>78,661</td>
<td></td>
<td></td>
</tr>
<tr>
<td>healthy</td>
<td>18b</td>
<td>run 1</td>
<td>794</td>
<td>3,018</td>
<td>7,640</td>
<td>92,795</td>
<td></td>
</tr>
<tr>
<td></td>
<td>run 2</td>
<td>829</td>
<td>3,245</td>
<td>7,442</td>
<td>93,485</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Microbial Metagenome

EC2.8.1.1 Thiosulfate sulfurtransferase
EC5.99.1.2 DNA topoisomerase

sulfate-reducing bacteria?
Microbial Metaproteome

- Alcohol dehydrogenase, class IV
- GTPases - translation elongation factors, Chaperonin GroEL (HSP60 family), ABC-type sugar transport systems-ATPase components
- Bacterial nucleoid DNA-binding protein, Ribosomal proteins, Carbon dioxide concentrating mechanism, transaldolase and transketolase and enolase
- GTPases - translation elongation factors, Chaperonin GroEL (HSP60 family), Branched-chain amino acid aminotransferase, Phosphomannomutase, Adenylosuccinate lyase, Membrane protease subunits, stomatin/prohibitin homologs

Crohns Disease  Healthy
Differentially abundant proteins reflect differences in taxonomic abundances
GO Terms for the Statistically Differentially Abundant Human Proteins

- monosaccharide catabolic process
- hexose catabolic process
- glucose catabolic process
- glycolysis
- regulation of heart rate
- heterocycle catabolic process
- translational elongation
- regulation of blood pressure
- response to reactive oxygen species
- response to oxidative stress
- cardiac myofibril assembly
- killing of cells of another organism
- cardiac muscle contraction
- ventricular cardiac muscle morphogenesis
- cell killing
- multicellular organismal catabolic process
- regulation of ATPase activity
- heart process
- heart contraction
- cardiac muscle tissue development
- heart morphogenesis
- cardiac muscle tissue morphogenesis
- muscle tissue morphogenesis
- ATP metabolic process
- cellular component assembly involved in morphogenesis
- actomyosin structure organization
- myofibril assembly
- muscle cell development
- striated muscle cell development
- epithelial cell differentiation
- regulation of response to external stimulus
- striated muscle cell differentiation
- response to metal ion
- muscle cell differentiation
- muscle tissue development
- striated muscle tissue development
- circulatory system process
- blood circulation
- tissue morphogenesis
- anti-apoptosis
- striated muscle contraction
- regulation of system process
- defense response to bacterium
- response to bacterium
- muscle organ development
- response to endogenous stimulus
- actin cytoskeleton organization
- response to inorganic substance
- cellular component morphogenesis

Almost 30% of proteins were human secreted proteins
The Victim

16S Metagenomics

WGS Metagenomics

Willing et al. 2010. Gastroenterology

Phosphotransferase system (PTS)
Membrane Transport

Faecalibacterium *

Metaproteomics
An Integrative Approach Provides Robustness

- Integration of 16S rRNA and WGS Metagenomics can give us 2 DNA centric views of the gut community
  - We can confirm differences by agreement of both methods.
  - Integration of WGS Metagenomics and Proteomics can give us 2 views of function -- capability vs activity
  - Integration can also highlight biases introduced by various methods.

Methods Questions? See Poster #20
Twinness Important?

- Examination of 16S metagenomics, WGS metagenomics and proteomics shows clustering by disease status, rather than twinness.
- While host genetics are important, we see that differences in microbiomes correlate better with disease status rather than host genetics.
• Who is there?
  • 16S Metagenomics
  • WGS Metagenomics

• What are the important functional players
  • WGS Metagenomics
  • Metaproteomics

• Taxa that are important for function?
  • Integrated Datasets
Future Plans

- For this study
  - Patient Cohort, Remission vs Active Disease
  - Deeper Coverage
    - Illumina WGS sequencing
    - Orbitrap Velos MS
    - Incorporation of New Reference Genomes
- For the field
  - High Throughput Biochemistry?
  - Diagnostic Tools
    - Assaying Protein Markers
    - Community profiling using genomic specific markers
  - Host-Microbial Interaction and GWAS studies
  - Knock-Out Microbiomes?
International Crohn’s Omics Team

UMB SOM, Baltimore
Claire Fraser-Liggett
Emmanuel Mongodin
Zhenqiu Liu
LBL, Berkley
Regina Lamendella
Janet Jansson
ORNL, Knoxville
Robert Hettich
Nathan Verberkmoes
Alison Erickson
Chongle Pan

VIB, Brussel
Jeroen Raes
Youssef Darzi
Karolinska Institute, Sweden
Lars Engstrand
Örebro University Hospital, Sweden
Curt Tysk
Jonas Halfvarson
CNRS, Marseille
Bernard Henrissat

Funding: NIH, HMP