Application of ‘Omics Technology to Infectious Diseases and the Human Microbiome

Karen Nelson
Marcus Jones
Historical...

- Genome of *Haemophilus influenzae* - 1995
- Reverse vaccinology - 2000
- Sargasso Sea Study - 2004
- First Human microbiome publication - 2006
- Diploid human genome - 2007
- Genome transplantation - 2007
- Global Ocean Survey, GOS - 2007
- Synthetic microbial genome - 2008
- >11,000 influenza genomes (75% of total worldwide and ongoing)
- Sequenced most major pathogens (e.g. TB, malaria, cholera, *T. parva*, *T. cruzi*)
- Vaccine development program – 2010
- Bacterial cell controlled by a synthetic chromosome – 2010
First Genome Sequenced 1995

25 thousand sequences

6.25 x 10⁸ pairwise comparisons
DNA synthesis Makes “Impossible” Genetic Manipulations Doable in Real Time

- We can synthesize genes and chromosomes cheaply and rapidly
- Enormous potential for new health and industrial applications
  - Production of biofuels
  - Small molecule therapeutics
  - New vaccines, antibiotics
  - Therapeutic microbes
  - Chloroplasts as plant factories
- Understand basic biology

Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome


We report the design, synthesis, and assembly of the 1.08-mega-base pair Mycoplasma mycoides JCVI-syn1.0 genome starting from digitized genome sequence information and its transplanted into a M. capricolum recipient cell to create new M. mycoides cells that are controlled only by synthetic chromosome. The only DNA in the cells is the designed synthetic DNA sequence including “watermark” sequences and other designed gene deletions and polymorphisms acquired during the building process. The new cells have expected phenotypic properties and are capable of continuous self-replication.
We are capable of sequencing and analyzing the genomes of culturable species. These species are estimated to represent less than 1% of total microbial diversity.

**Culture dependent analysis:**
Culture and obtain pure colonies
Complete genome sequencing of DNA
Organism has to be cultured in the laboratory

**Culture-independent analysis**
16S ribosomal RNA (rRNA) sequencing
Whole genome sequencing, assembly, annotation

**Metagenomics:** sequence based analysis of complete microbial communities without need for culturing

Made possible by number of parallel developments:
- **Assembly and data analysis capabilities** developed to being able to tease apart these large datasets
- **Sequencing capabilities** capable of achieving great depths of coverage at reduced cost
- **Demonstrated proof of concept** via Sargasso Sea study
- **Global Ocean Sampling (GOS)** largest protein dataset in existence

**Other “omics” technologies.** Proteomics, metatranscriptomics, metabolomics
Changes in Sequencing Technologies

- **ABI 3730xl**: 1-2 Mb/day
- **Illumina GA IIx**: 50 Gb/12 day run
- **ABI SOLiD**: 100Gb/12 day run
- **Ion Torrent**: 1Gb/2hr run (available at the end of 2012)
- **454 GS FLX +**: 0.6Gb/23hr run
- **Illumina HiSeq 2000 (2500†)**: 600 Gb/11 day run
- **HiSeq 2500 upgrade**: up to 120Gb/27 hour run (available now for $50K)
- **Ion Proton‡**: up to 100Gb/4 hour run (available at the end of 2012)
Changes in genomics sciences

<table>
<thead>
<tr>
<th>Year</th>
<th># of Sequencers</th>
<th>FTE</th>
<th>Sequence data produced (Gbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>100</td>
<td>77</td>
<td>24</td>
</tr>
<tr>
<td>2006</td>
<td>100</td>
<td>70</td>
<td>27</td>
</tr>
<tr>
<td>2007</td>
<td>78</td>
<td>64</td>
<td>38</td>
</tr>
<tr>
<td>2008</td>
<td>55</td>
<td>60</td>
<td>36</td>
</tr>
<tr>
<td>2009</td>
<td>29</td>
<td>30</td>
<td>350</td>
</tr>
<tr>
<td>2010</td>
<td>9</td>
<td>17</td>
<td>900</td>
</tr>
<tr>
<td>2011</td>
<td>9</td>
<td>13</td>
<td>14000</td>
</tr>
</tbody>
</table>
Sargasso Sea study

- Venter and colleagues at the JCVI
- Generated 1,987,936 DNA reads
- Approximately 1,625 Mb of DNA
- 1.2 million new genes identified
- ~1,412 rRNA genes
- Estimated 1,800 species
- 12 complete genomes recovered
- Demonstration of the power of genomics
Global Ocean Sampling Expedition
Global Ocean Sampling and Analysis

Sampling and Sequencing

Tool Development

Data Analysis
GOS increases size and diversity of known protein families
(Yooseph et al, 2007 *PLoS Biol*)

GOS: prokaryotes, eukaryotes
Known: prokaryotes, eukaryotes

RuBisCO

Glutamine synthetase (type II)
Spin off “omics” studies
transcriptomics – metabolomics
Metagenomic projects

Types include

- Various animal species, insects, non-human primates
- Global Ocean Sampling and other marine environments
- Lake water
- Air
- Bioremediation Sites
- Soil
- Humans
Mucosal samples were obtained during colonoscopy from healthy-appearing sites within the six major subdivisions of the human colon: cecum ascending colon transverse colon descending colon sigmoid colon rectum.

Fecal samples were collected from each subject 1 month following colonoscopy.

From 11,831 bacterial and 1524 archaeal 16S sequences, identified 395 phylotypes.

Eckburg et al., 2005 Science
Stomach

1,833 full-length 16S sequences

Described 128 16S rDNA phylotypes

Derived from 23 human subjects

First human metagenomic paper
- Investigated the gastrointestinal tract (via fecal samples) of two healthy adults
- 78 Mbp
- 2062 amplified 16S rDNA

Metagenomic Analysis of the Human Distal Gut Microbiome


The human intestinal microbiota is composed of $10^{13}$ to $10^{14}$ microorganisms whose collective genome ("microbiome") contains at least 100 times as many genes as our own genome. We analyzed ~78 million base pairs of unique DNA sequence and 2062 polymerase chain reaction-amplified 16S ribosomal DNA sequences obtained from the fecal DNAs of two healthy adults. Using metabolic function analyses of identified genes, we compared our human genome with the average content of previously sequenced microbial genomes. Our microbiome has significantly enriched metabolism of glycanas, amino acids, and xenobiotics; methanogenesis; and 2-methyl-α-erythritol 4-phosphate pathway-mediated biosynthesis of vitamins and isoprenoids. Thus, humans are superorganisms whose metabolism represents an amalgamation of microbial and human attributes.

Our body surfaces are home to microbial communities whose aggregate metabolic outputs are human ≥100 times as many genes as our 2.85-billion base pair (bp) human genome (/1). Therefore, a comprehensive view of our genetic landscapes of single organisms, recent reports from Venter et al. (9) and Baker et al. (10) have demonstrated the utility of this approach for studying mixed microbial communities. Variations in the relative abundance of each member of the microbial community and their respective genome sizes determine the final depth of sequence coverage for any organism at a particular level of sequencing. This means that the genome sequences of abundant species will be well represented in a set of random shotgun reads, whereas lower abundance species may be represented by a small number of sequences. In fact, the size and depth of coverage (computed as the ratio between the total length of the reads placed into contigs and the total size of the contigs) of genome assemblies generated from a metagenomics project can provide information on relative species abundance.

A total of 65,059 and 74,462 high-quality sequence reads were generated from random DNA libraries created with fecal specimens of two healthy humans (subjects 7 and 8). These two subjects, ages 28 and 37, female and male, respectively, had not used antibiotics or metagenomic Analysis of the Human Distal Gut Microbiome.

Gill et al., Science 2006
Human Microbiome
Metagenomics, Health and Disease
Human Microbiome

~$10^{12}$ Human cells

~$10^{13}$ Bacterial cells

>600 oral bacterial species
• Collective of the human microbiome exceeds the number of human cells by at least an order of magnitude.
• Many of these microbial interactions endow or enhance human physiology including processes related to development, nutrition, immunity and resistance to pathogens.
• The majority of the human microbiome remains unknown.
• Many relationships between the human host and microbiome remain to be determined.

image courtesy of the NIH HMP website
http://nihroadmap.nih.gov/hmp/
The Human Microbiome

Significant role: Example in the Gastrointestinal tract

• They foster development of the mucosal wall.
• The development and maturation of the immune system is dependent on the presence of some members of the intestinal microbiota. Link to human health and disease.
• Essential for the metabolism of certain compounds as well as xenobiotics.
• Protection against epithelial cell injury.
• Regulation of host fat storage.
• Stimulation of intestinal angiogenesis.
Microbiota are acquired anew each generation.

1) Infants obtain microbes from mother or environment.

2) Microbial succession over ~1-2 yrs.

3) Microbiome becomes “adult-like” in ~1-2 yrs.
NIH Roadmap Human Microbiome Project

- Budget > $175 million 2007-2013
- Goal: Characterize the microbes that inhabit the human body and examine whether changes in the microbiome can be related to health and disease
- Feasibility project designed to determine the value of microbial metagenomics to biomedical research
- Community Resource Project-generate reagents and data sets; rapidly placed in public domain
- Continuous Scientific Community Input
  - External Scientific Advisory Group, Workshops.

http://nihroadmap.nih.gov/hmp
http://www.human-microbiome.org/#

Slide courtesy Maria Giovanni-NIAID
NIH HUMAN MICROBIOME PROJECT

- 3000 Reference Bacterial Genomes; Viral and Eukaryotic Genomes
- Metagenomic Data Set
  - 300 healthy humans
  - Diverse Body Sites
- Technology & Bioinformatic Tools Development; ELSI
- Database and Resource Center
- Reagent Repository
- Demonstration Projects
  - Changes in Microbiome Health & Disease

Slide courtesy Maria Giovanni-NIAID
“Healthy Cohort” Body Sites

- Saliva
- Tongue dorsum
- Hard palate
- Buccal mucosa
- Keratinized (attached) gingiva
- Palatine tonsils
- Throat
- Supragingival plaque
- Subgingival plaque
- Retroauricular crease, both ears (2)
- Antecubital fossa (inner elbow), both arms (2)
- Anterior right and left nares (pooled)
- Stool
- Posterior fornix, vagina
- Midpoint, vagina
- Vaginal introitus

Slide courtesy of NHGRI
Supplementary Figure 8. Phylum abundances per body site. For each of the body sites studied by both 16S rRNA gene sequencing (A) and whole-genome shotgun sequencing (B) the five most abundant phyla are shown. The small remaining fraction of the data is collapsed and labeled as other phyla (grey).
In adults, each part of the body supports a distinct microbial community.

With no apparent relationship with gender, age, weight, ethnicity or race.

Some results from HMP:

HMP estimates for global microbiome:

~ 10,000 microbial species

~ 8 million microbial genes
Sub-body sites have distinct communities

Soft

1. Cheek
2. Palate
3. Gums
4. Tonsils
5. Saliva
6. Subgingival Plaque
7. Supragingival Plaque

Hard

8. Throat
9. Tongue

Slide courtesy of HMP Consortium and Bruce Birren, Broad Institute
- Reference Strains: Generate complete genomes from > 3000 prokaryotes.
- Build our understanding of those recognized through 16S profiles
- Provide for interpretation of metagenomics and other “omics” data
- Sequence reference phage, viruses and eukaryotes

A Catalog of Reference Genomes from the Human Microbiome

178 genomes
~550,000 genes
Nelson et al., Science
May 21, 2010
Reference Genomes of the Human Microbiome Project

In order to facilitate the phylogenetic and functional analysis of the metagenomic sequences produced from human body sites, the HMP plans to sequence, or collect from publicly available sources, a total of 1000 reference genomes. The organisms included in this collection have all been isolated from a human body site. The information gained from the Reference Genomes will allow 16S RNA sequences and metagenomic sequence from uncharacterized microbiome organisms to be grouped phylogenetically with related organisms from the reference set providing information about the taxonomy of the unknown strains. Likewise, functional characterization of proteins in the reference organisms will aid in the functional annotation of related proteins contained in the sequence fragments derived from metagenomic samples.

Choosing Reference Organisms:
The HMP has developed a detailed set of guidelines for inclusion of a strain in the reference genome group. If you have suggestions for additional strains to include on the list or if you have a strain that you would like to contribute please use our feedback form to let us know.

- Guidelines for inclusion of a strain
- Feedback form - help us by recommending strains to include in the HMP reference genome collection
- Current breakdown of strains according to body site
- Phylogenetic Analysis - Below are phylogenetic trees of HMP organisms in the context of a wide sampling of sequenced and/or culturable bacteria:
  - All HMP reference genomes
    - Reference genomes isolated from airways
    - Reference genomes isolated from the gastrointestinal tract
    - Reference genomes isolated from the oral cavity
    - Reference genomes isolated from the skin
    - Reference genomes isolated from the vagina

HMP Catalog

For a full list of the HMP reference genomes please visit the HMP Project Catalog where you can search for strains by many features and characteristics, including body site and taxonomy. The collection of strains in the HMP Project Catalog represents projects at all stages including those that are planned (project status "targeted") as well as those that have reached completion (project status "complete"). Also included in the set are strains that are being sequenced by members of the International Human Microbiome Consortium (link to further down section of page). More information on this effort can be found further down this page.

Most of the HMP Reference Genomes will be sequenced only to the “standard draft” stage, a minimum standard for a draft genome that has been established by the HMP sequencing centers. Draft genome sequence does not include every base of the genome, rather they are assemblies of several large contiguous pieces of sequence (contigs) with subsequent gaps in sequence knowledge. About 15% of the reference strains will be taken closer to a “finished” or fully complete state. There are several finishing levels that genomes can be taken to, each with an associated cost. The same guidelines mentioned above for choosing which strains to include on the list are applied to decide which of the strains should be promoted to a higher state of finishing. A standardized set of Finishing Categories is currently under development by a multi-center, international group of researchers. Once finalized, they will be posted on this site and each strain will be assigned to one of the categories.
The Human Microbiome: Altering the future of medicine

- Microbiome influenced by many factors including environment and host genetics
  - Complex bio-feedback mechanism: host <-> microbiome
- This population can be studied and altered to benefit the host
  - Normal flora of healthy individuals can potentially be mined to identify new probiotics
  - Population changes/shifts can be used as indicators of deterioration/improvement of health
  - Can be used for disease surveillance
- Need for integration of multiple “omics” approaches to understand the complexity of the microbiome and its broader implications
Disease related microbiome studies at JCVI

- Progression of esophageal cancer (NYU)
- Bacterial vaginosis and pre-term delivery (Illinois/Mayo; NIAID)
- Nasopharynx microbiome and vaccination in children (Gates)
- Skin microbiome, acne and psoriasis (NYU)
- Oral diseases including periodontitis (NYU)
- Colon cancer (Howard University)
- Type 1 Diabetes pilot (TEDDY)
Can we use as a biomarker for:

- Development of new predictive biomarkers so that preventive strategies based on pre- and probiotics can be developed.
- New therapeutic strategies
- Increase our understanding of the etiologies of complex diseases and health
NIDDK funded - Type 1 Diabetes Study

Integrated “omics” approaches

Gut Microbiome/Virome
- Viral-Microbial specific Biosignatures

Urinary Proteome
- Protein patterns

Urinary Metabolome
- Metabolite analysis

Compare and correlate gut microbiome, proteomic and metabolomic datasets –
On host side - HLA genotype, islet autoantibody status and Type 1 Diabetes status

Identification of panel of Biomarkers Candidates
Still need:

- Technology development
- Informatics and data handling
- Education
- Well defined studies
Transcriptomic and Proteomic Analyses of the Microbiome and Infectious Diseases
“Omics” Technologies

- Metagenomic Analysis
- Transcriptomics
- Proteomics
- Glycomics
- Lipodomics
- Metabolomics
What is Transcriptomics?
Transcriptomics Technologies

**in vitro** Gene Expression
- RNA purification
- cDNA synthesis
- Microarray Hybridizations
- Data analysis

**Genome Sequencing**
- gDNA
- 454 Sequencing
- Newbler assembly
- 3’ qPCR confirmation

**in vivo** Gene Expression
- RNA
- cDNA synthesis
- 384-well qRT-PCR
- Data analysis
Characterization of *in vitro* Samples

- 2 hour incubation
- 30 minute incubation
Environmental Stress

LexA gene
SOS Box
RecA gene
SOS Enzymes

Excision repair
Effect of levofloxacin on *B. anthracis* $\gamma$-polymerase

$log_2$ expression change
SOS response in *B. anthracis*

"DNA-damage-inducible protein P, putative"
LexA repressor
"recA protein, group I intron-containing"
"excinuclease ABC, A subunit"
"excinuclease ABC, B subunit"
"excinuclease ABC, C subunit"
**RNA purification** → **cDNA synthesis** → **Microarray Hybridizations** → **Data analysis**

**in vitro Gene Expression**

**gDNA** → **454 Sequencing** → **Newbler assembly** → **3’ qPCR confirmation**
Mapping Promoters: Chip-chip Analysis of *in vitro/ ex vivo* Samples

1. **Cross-link protein to DNA**
2. **Sonicate to fragment to DNA**
3. **Immunoprecipitate**
4. **Purify and label DNA**
5. **Hybridize to microarray**
RpoE Binding Site Oligos
RNASeq Data
## Linezolid Lineage Clinical Isolates

<table>
<thead>
<tr>
<th>locus</th>
<th>start</th>
<th>end</th>
<th>symbol</th>
<th>WT</th>
<th>SNP</th>
<th>3577</th>
<th>5612</th>
<th>5892</th>
<th>7210</th>
<th>function</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA2212</td>
<td>2483696</td>
<td>2482581</td>
<td>N/A</td>
<td>G</td>
<td>C</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>positive</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>SA1577</td>
<td>1815520</td>
<td>1808960</td>
<td>N/A</td>
<td>T</td>
<td>G</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>positive</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>SA1118</td>
<td>1268775</td>
<td>1270448</td>
<td>N/A</td>
<td>T</td>
<td>C</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>positive</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>SA1924</td>
<td>2173700</td>
<td>2172273</td>
<td>N/A</td>
<td>A</td>
<td>G</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>positive</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>SA1669</td>
<td>1906904</td>
<td>1905519</td>
<td>fumC</td>
<td>A</td>
<td>G</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>positive</td>
<td>fumarate hydratase</td>
</tr>
<tr>
<td>SA0500</td>
<td>579620</td>
<td>583171</td>
<td>rpoB</td>
<td>T</td>
<td>A</td>
<td>negative</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>DNA-directed RNA polymerase subunit beta</td>
</tr>
<tr>
<td>SA0264</td>
<td>318934</td>
<td>319926</td>
<td>N/A</td>
<td>C</td>
<td>T</td>
<td>negative</td>
<td>negative</td>
<td>positive</td>
<td>positive</td>
<td>hypothetical protein</td>
</tr>
</tbody>
</table>

### Two-step Resistance
<table>
<thead>
<tr>
<th><strong>RNA purification</strong></th>
<th><strong>cDNA synthesis</strong></th>
<th><strong>Microarray Hybridizations</strong></th>
<th><strong>Data analysis</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>in vitro</em> Gene Expression</td>
<td><img src="image" alt="RNA purification" /></td>
<td><img src="image" alt="cDNA synthesis" /></td>
<td><img src="image" alt="Microarray Hybridizations" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>gDNA</strong></th>
<th><strong>454 Sequencing</strong></th>
<th><strong>Newbler assembly</strong></th>
<th><strong>3’ qPCR confirmation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Sequencing</td>
<td><img src="image" alt="gDNA" /></td>
<td><img src="image" alt="454 Sequencing" /></td>
<td><img src="image" alt="Newbler assembly" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>RNA</strong></th>
<th><strong>cDNA synthesis</strong></th>
<th><strong>384-well qRT-PCR</strong></th>
<th><strong>Data analysis</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>in vivo</em> Gene Expression</td>
<td><img src="image" alt="RNA" /></td>
<td><img src="image" alt="cDNA synthesis" /></td>
<td><img src="image" alt="384-well qRT-PCR" /></td>
</tr>
</tbody>
</table>
qRT-PCR Validation using Roche 480 LC
Characterization of Pathogen Gene Expression During Infection
Hypothetical genes differentially expressed *in vivo*
Figure 1

~2.9 Mb Genome View

Panel A.

Log₂ Scale

Panel B.

Log₂ Scale

ACME
Lipoproteins
Oxytans
Sap5
qScCasA
LasNuc
and Las30F
QrSc
UrSe
Crescines
What is Proteomics?
Additional Public Health Concerns

- UTIs are the most common cause of hospital-acquired infections accounting for approximately 40% of the total

- Many of these UTIs are caused by the ESKAPE pathogens. There is an increasing shortage of effective antibiotics against pathogens with multiple resistances
  - Enterococcus faecium
  - Staphylococcus aureus
  - Klebsiella pneumoniae
  - Acinetobacter baumanii
  - Pseudomonas aeruginosa
  - Enterobacter species

- Carbapenem (Kp, Ec); MDR (Pa); penicillins and vancomycin (Ef, Sa)

- Large number of immune-compromised patients: HIV/AIDS, transplant and cancer patients
More informative Methods for UTI and ASB Diagnosis?

• Vaginal and urinary tract microbiome profiling (sensitive detection of protective bacteria, ESKAPE pathogens, anaerobes missed in urine cultures): metagenomics

• Protein profiling to identify the bacteria and survey antimicrobial and immune responses: proteomics

Fouts et al., J Transl Medicine (2012) 10, 174:
“Integrated next generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury”
Proteomics: Analysis Stages

- Pellet isolation
- Protein denaturation
- Protein digestion
- Mascot data analysis
- Quantitative methods
- Sample acquisition
- LC-MS/MS
- Electrospray

- Annexin A1 n=14 Tax=Eutheria RepID=ANXA1_HUMAN
- Alkyl hydroperoxide reductase subunit C [Klebsiella pneumoniae 342]
The database search space comprises ~80,000 distinct proteins

Patent application: Pieper et al., January 2013
Acknowledgements

- All JCVI faculty, staff and collaborators
- Funding Agencies: NIH
  - NIAID
    - Genome Sequencing Center (NIH-HHSN272200900007C)
    - System Biology for EnteroPathogens (NIH-HHSN27220070058C)
    - Pathogen Functional Genomics Resource Center (N01-AI-15447)
  - NIDCR
  - NIDDK