Microbial Genomics

Michael J. Stanhope,

Fleischmann et al. 1995. Science 269: 496
Outline

- Introduction
  - Microbial diversity
  - Universal Tree of Life
- Bacterial genome size
  - Core and pan genomes
- Horizontal Gene Transfer (HGT)
  - Mechanisms of HGT
  - Detecting HGT
- Comparative genomics of *Streptococcus*
- Comment on genome sequencing technology
- E.g. of 454 bacterial genome sequence
- Applications of microbial genomics
Introduction

- Microbial diversity
Microbial diversity

- Superficial inspection, bacteria and archaea hardly seem diverse

http://www.ucmp.berkeley.edu/archaea/archaeamm.html
Microbial diversity

- But metabolic diversity great, particularly energy generating
  - Even within a species; e.g. *E. coli*:
    - Fermentation or respiration; respire aerobically or anaerobically; glucose or lactose as sole carbon source
      - transforming sugar into amino acids, vitamins, nucleotides
Microbial diversity

- Energy generating metabolism in bacteria:
  - Alcohol fermentation
  - Lactic acid fermentation (present in eukaryotes & prokaryotes)
  - Aerobic respiration
  - Oxygenic photosynthesis
  - Anaerobic degradation of carbohydrates through the Embden-Meyerhof pathway.
  - Other fermentation pathways e.g. phosphoketolase pathway
  - Anaerobic respiration
  - Lithotrophy (inorganics as source of energy)
  - Anoxygenic photosynthesis
  - Methanogenesis (H₂ as energy source and produces methane)
  - Light driven nonphotosynthetic photophosphorylation
Microbial diversity

- prokaryotic cells on Earth = $6 \times 10^{30}$
- Prokaryotic cellular carbon = 60-100% of estimated carbon in terrestrial and marine plants.
- Abundant in environments where eukaryotes are rare
- How many species?
  - Definition of species?
    - Lack diagnostic morphological characteristics
    - Exchange genetic material in unique and unusual ways
    - Same species = 70% DNA-DNA hybridization
      - Underestimating prokaryotic diversity
  - Practical limitations in counting
    - 1% cultivable

Underestimating prokaryotic diversity
Practical limitations in counting
1% cultivable
Introduction

- Universal Tree
Universal Tree of Life

- 1980’s Carl Woese, phylogenetic analysis of all forms of cellular life; ssrRNA
  - Found in all cells
  - Present in thousands of copies and easy to isolate
  - Complementary to sequence of gene
  - Sequence can be compared to reveal similarity and differences

- Defined three cellular domains of life:
  - Eukaryotes
  - Eubacteria (Bacteria)
  - Archaeabacteria (Archaea)
Pace, NR. 1997. Science 276:734
THE UNIVERSAL TREE OF LIFE

http://whyfiles.org/022critters/archaea.html
Genome Size
Genome size

- 405 complete bacterial genomes on NCBI
  - *Carsonella ruddii* (159,662) – *Burkholderia xenovorans* (9.73 Mb)

- Genome size / ecological niche
  - Smaller genomes, endocellular parasites or symbionts
Genome size

- Mutually obligate endosymbiotic associations with animal hosts
  - bacteriocytes

Genome size

Nakabachi et al. 2006. Science 314:267
Genome size

- Free living bacteria, genome size correlates with species metabolism & width of ecological niche
  - Pathogenic species, narrow range of hosts, small genomes; e.g. *Helicobacter, Streptococcus*
  - Anaerobic bacteria, restricted metabolism, e.g. methanogens, small genomes.
  - Aerobic organisms, and opportunistic pathogens, higher diversity of genome size; e.g. *Pseudomonas* (6 Mb)
pan and core genomes

- Core
  - Genes present in all strains
- Pan (from Greek meaning whole)
  - Dispensable genome composed of genes absent from one or more strains and genes unique to particular strains
Bacteria chromosomes

- Most, single circular chromosome, but exceptions:
  - E.g. *Streptomyces, Borrelia, Agrobacterium*, linear chromosomes
  - Linear plasmids – e.g. *Klebsiella, Escherichia, Thiobacillus*
  - Linearity: enhances genomic plasticity?
  - Multichromosome spp.; e.g. some proteobacteria with free living, opportunistic lifestyle
Horizontal Gene Transfer
Horizontal Gene Transfer

- Genetic exchanges between different evolutionary lineages
- 1944 Avery et al., DNA can be absorbed by microorganisms (Studies on the chemical nature of the substance inducing transformations of pneumococcal types. J. Exp. Med. 79:137)
- Extent or degree is much debated
Mechanisms of HGT

(a) Conjugation
F factor

Bacterium
with alleles
a, b, c

Transfer during
conjugation, after
integration

F factor origin

F factor terminus

Recombination

(b) Transformation

Cell lysis

Fragment
taken up by
new bacterium

Free DNA
fragments

(c) Transduction

Virus

Empty
viral coat

Bacterial
chromosome
fragment

New virus
particles

Infection

New bacterium

http://fig.cox.miami.edu/Faculty/Dana/bacfun.jpg
Detecting HGT

- Phylogenetics
  - Gene tree that differs significantly from species tree
  - Compare all gene trees; gene trees that are significantly different from majority are putative LGT
Detecting HGT
Detecting HGT

- Best sequence match detection (BLAST)
  - Rapid, but of limited use, since sequence similarity not necessarily correlated with evolutionary history.
Bacteria to Vertebrate Horizontal Gene Transfer??

“Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage.”

Bacteria to Vertebrate HGT -- Implications (*If True*)

- HGT bacterial genes became fixed in vertebrates through insertion into germ cells (because somatic cell HGT genes would be lost within a generation).
- Foreign bacterial genes can co-opt vertebrate regulatory regions and transcription factors.
- Humans could accumulate foreign, perhaps deleterious, genes from bacterial infections and/or GM foods.
International Human Genome Sequencing Consortium (IHGSC)

- 113 genes that are likely examples of bacteria to vertebrate HGT (horizontal gene transfer).
- Conclusion based on BLASTP alignment scores.
  Best sequence match detection (BLAST)
Phylogenetic evidence in support of bacteria - vertebrate HGT

Human
Vertebrate
Bacteria1
Bacteria2
Bacteria3
Bacteria4
Bacteria5
Bacteria6
Non-vert Eukaryote
Non-vert Eukaryote
Paralog
Phylogenetic evidence rejecting bacteria - vertebrate HGT

Bacteria1
Bacteria2
Bacteria3
Bacteria4
Bacteria5
Bacteria6
Human
Vertebrate
Non-vert Eukaryote
Non-vert Eukaryote
Paralog
Why did the IHGSC conclude bacteria to vertebrate HGT?

- Equated BLAST ranking with evolutionary relatedness.
Detecting HGT

- **Nucleotide compositional analysis**
  - Based on premise that DNA fragments obtained through HGT retain sequence characteristics of donor genome
  - Advantage is it only requires genome sequence from 1 spp.
Comparative Genomics of Streptococcus
Streptococcus genomes

- 26 genomes (public) from 6 spp
Adaptive potential of bacteria

1. Darwinian or positive selection, favoring the fixation of advantageous mutations
2. Acquisition of new genetic material by lateral DNA exchange
3. Gene regulation
Core genome

- LGT of bacterial genomes, possibly key factor in adaptation
  - Nonetheless, core genome, possibly relatively LGT free
- Focus on adaptation often centered on species specific loci
  - Selection pressure on core genome not explored
Molecular selection

- Powerful statistical methods for detecting adaptive molecular evolution (Yang and Nielsen)
  - Nonsynonymous substitution rate elevated above the synonymous rate as evidence for positive selection
    - Fixation of advantageous mutations, driven by NS => evolutionary innovations
- Our goal: assess positive selection pressure across core genome components of Streptococcus, while concomitantly assessing levels of recombination within core genome
Pipeline (part 1)

Gene vs genomes table

<table>
<thead>
<tr>
<th>gene #1</th>
<th>genome #1</th>
<th>genome #2</th>
<th>genome #3</th>
<th>genome #4</th>
<th>genome #5</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene #1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene #2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene #3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene #4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>gene #5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene #6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>gene #7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>gene #8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Set containing genome #4

Set of strict orthologs

Alignment

gene #1

Gene by gene alignments

gene #2
Pipeline (part 2)

Intragenic recombination detection

Alignment without recombination

Recombination breakpoint

gene #1 fragment #1

gene #1 fragment #2

Phylogenetic reconstructions

gene #1 fragment #1

gene #1 fragment #2

Positive selection detection

Genes, lineages and sites showing evidence of evolution by positive selection

...
Estimated pan & core genome sizes

S. agalactiae
S. pyogenes

Alignable core genome size for interspecific analysis = 260
## Genes under positive selection: between species

<table>
<thead>
<tr>
<th>Lineage</th>
<th>nbr analyzed</th>
<th>nbr under PS</th>
<th>% under PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. <em>mutans</em></td>
<td>260</td>
<td>33</td>
<td>12.69</td>
</tr>
<tr>
<td>S. <em>pneumoniae</em></td>
<td>260</td>
<td>73</td>
<td>28.08</td>
</tr>
<tr>
<td>S. <em>suis</em></td>
<td>260</td>
<td><strong>89</strong></td>
<td><strong>34.23</strong></td>
</tr>
<tr>
<td>S. <em>thermophilus</em></td>
<td>260</td>
<td>61</td>
<td>23.46</td>
</tr>
<tr>
<td>S. <em>agalactiae</em></td>
<td>260</td>
<td>28</td>
<td>10.77</td>
</tr>
<tr>
<td>S. <em>pyogenes</em></td>
<td>260</td>
<td>44</td>
<td>16.92</td>
</tr>
<tr>
<td>(S. <em>pneumoniae</em>, S. <em>suis</em>)</td>
<td>221</td>
<td>71</td>
<td>32.13</td>
</tr>
</tbody>
</table>
## Genes under positive selection: S. *agalactiae*

<table>
<thead>
<tr>
<th>Lineage</th>
<th>nbr analyzed</th>
<th>nbr under PS</th>
<th>% under PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>COH1</td>
<td>1212</td>
<td>7</td>
<td>0.58</td>
</tr>
<tr>
<td>18RS21</td>
<td>1212</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>NEM316</td>
<td>1212</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>H36B</td>
<td>1212</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>A909</td>
<td>1212</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>2603V/R</td>
<td>1212</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>CJB111</td>
<td>1212</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>515</td>
<td>1212</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>
## Genes under positive selection: *S. pyogenes*

<table>
<thead>
<tr>
<th>Lineage</th>
<th>nbr analyzed</th>
<th>nbr under PS</th>
<th>% under PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGAS10270</td>
<td>1297</td>
<td>7</td>
<td>0.54</td>
</tr>
<tr>
<td>MGAS10394</td>
<td>1297</td>
<td>3</td>
<td>0.23</td>
</tr>
<tr>
<td>MGAS10750</td>
<td>1297</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>MGAS2096</td>
<td>1297</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>MGAS315</td>
<td>1297</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>MGAS5005</td>
<td>1297</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>MGAS6180</td>
<td>1297</td>
<td>2</td>
<td>0.15</td>
</tr>
<tr>
<td>MGAS8232</td>
<td>1297</td>
<td>4</td>
<td>0.31</td>
</tr>
<tr>
<td>MGAS9429</td>
<td>1297</td>
<td>2</td>
<td>0.15</td>
</tr>
<tr>
<td>M1 GAS</td>
<td>1297</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SSI-1</td>
<td>1297</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(MGAS9429, MGAS2096)</td>
<td>925</td>
<td>2</td>
<td>0.22</td>
</tr>
<tr>
<td>(MGAS5005, M1 GAS)</td>
<td>978</td>
<td>4</td>
<td>0.41</td>
</tr>
<tr>
<td>(SSI-1, MGAS315)</td>
<td>983</td>
<td><strong>9</strong></td>
<td><strong>0.92</strong></td>
</tr>
</tbody>
</table>
## Recombination

<table>
<thead>
<tr>
<th>Data set</th>
<th>SPI (strong phylogenetic incongruence)</th>
<th>PHI (intragenic method)</th>
<th>PHI ∩ MaxChi ∩ NSS (set of intragenic methods)</th>
<th>SPI ∩ PHI</th>
<th>SPI U intragenic set</th>
</tr>
</thead>
<tbody>
<tr>
<td>interspecific</td>
<td>26 (10%)</td>
<td>54 (21%)</td>
<td>35 (14%)</td>
<td>11 (4%)</td>
<td>53 (20%)</td>
</tr>
<tr>
<td>S. pyogenes</td>
<td>434 (33%)</td>
<td>284 (22%)</td>
<td>168 (13%)</td>
<td>186 (14%)</td>
<td>477 (37%)</td>
</tr>
<tr>
<td>S. agalactiae</td>
<td>222 (18%)</td>
<td>34 (3%)</td>
<td>7 (1%)</td>
<td>18 (1%)</td>
<td>223 (18%)</td>
</tr>
</tbody>
</table>
Recombination and positive selection

<table>
<thead>
<tr>
<th>Data set</th>
<th>Genes under PS</th>
<th>PS + recombinant</th>
<th>PS + SPI</th>
<th>PS + intragenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>interspecific</td>
<td>175</td>
<td>43 (25%)</td>
<td>20 (8%)</td>
<td>29 (11%)</td>
</tr>
<tr>
<td><strong>S. agalactiae</strong></td>
<td>10</td>
<td>4 (40%)</td>
<td>4 (40%)</td>
<td>0</td>
</tr>
<tr>
<td><strong>S. pyogenes</strong></td>
<td>32</td>
<td><strong>25 (78%)</strong></td>
<td>21 (65%)</td>
<td>17 (53%)</td>
</tr>
</tbody>
</table>
Pan genome and recombination

- Habitat differences for *S. pyogenes* and *S. agalactiae*
  - Reduced gene pool environment for *S. pyogenes*, could result in smaller pan genome and potentially more homologous recombination
Statistical analysis of PS data

- Significant affect of lineage (ANOVA; p<0.0001):
  - Majority of pairwise multiple comparisons significantly different
- Significant affect of biochemical category (p<0.0001)
  - Amino acid biosynthesis; Biosynthesis of cofactors, prosthetic groups, and carriers; Cell envelope; Cellular processes; Central intermediary metabolism; **DNA metabolism**; Energy metabolism; Fatty acid and phospholipid metabolism; Hypothetical proteins; Protein fate; Protein synthesis; Purines, pyrimidines, nucleosides, and nucleotides; Regulatory functions; Signal transduction; **Transcription**; Transport and binding proteins; Unknown function
- Significant interaction between lineage and biochemical category (p=0.003)
  - (S. pneumoniae, S. suis) DNA metabolism, Transcription, Protein fate
Genes selected per lineage

![Histogram showing frequency of genes selected per lineage](image)

- **X-axis:** Number of lineages
- **Y-axis:** Frequency

The histogram displays the distribution of genes selected across different numbers of lineages. The majority of genes are selected in 1 lineage, with a significant drop in frequency for 2 or more lineages.
19 unique loci for *S. suis*; 15 for *S. thermophilus*; 14 for *S. pneumoniae*
Lineages with unusual selection pressure

- **S. suis**
  - Both gene gain and loss and PS; suggesting evolutionary flexibility – host jumping?

- **S. agalactiae, COH1**
  - Significantly associated with neonatal disease, and of recent bovine ancestry

- **S. pyogenes, M3 serotype**
  - M3 cause more cases of invasive disease, higher rate of lethal infections, epidemic tendencies

- **S. thermophilus, LMD-9**
  - ?
Streptococcus comparative genomics tentative conclusions

- Considerable recombination and positive selection pressure in *Streptococcus* core genome
- Several loci identified for *S. agalactiae* and *S. pyogenes* that could be linked to the specific pathogenic features of these strains
- Identification and cataloguing of these loci, serve as an evolutionary short-cut for laboratory mutation experiments, to assess specific functional significance of these genes.
Sequencing Technology
First “shot gun” microbial genome sequence

*Haemophilus influenzae* 1.8 Mb

Library of plasmid clones, 1600-2000 bp fragments; sequences of these clones with their many overlaps represent the raw data entered into computer programs (e.g. TIGR assembler) which assemble the genome;

remaining gaps closed with other strategies (e.g. long range PCR)

Fleischmann et al. 1995. Science 269: 496
Race for the $1000 genome

- First to produce $1000 human genome
  - J. Craig Venter Science Foundation: $500,000
  - X Prize Foundation: $5 million
- 2004; NIH; $70 million grant program
## Next generation of sequencers

### Searching for Cheaper Genome Sequencers

<table>
<thead>
<tr>
<th>Company</th>
<th>Format</th>
<th>Read Length (bases)</th>
<th>Expected Throughput Mb (million bases)/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 Life Sciences</td>
<td>Parallel bead array</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>Agencourt Bioscience</td>
<td>Sequencing by ligation</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Applied Biosystems</td>
<td>Capillary electrophoresis</td>
<td>1000</td>
<td>3-4</td>
</tr>
<tr>
<td>LI-COR Biosciences</td>
<td>Electronic microchip</td>
<td>20,000</td>
<td>14,000</td>
</tr>
<tr>
<td>Microchip Biotechnologies</td>
<td>Parallel bead array</td>
<td>850-1000</td>
<td>7</td>
</tr>
<tr>
<td>Network Biosystems</td>
<td>Biochip</td>
<td>800+</td>
<td>5</td>
</tr>
<tr>
<td>NimbleGen Systems</td>
<td>Map and survey microarray</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>Solexa</td>
<td>Parallel microchip</td>
<td>35</td>
<td>500</td>
</tr>
<tr>
<td>VisiGen Biotechnologies</td>
<td>Single-molecule array</td>
<td>NA</td>
<td>1000</td>
</tr>
</tbody>
</table>

*from: Service, RF 2006. Science 311:1544*
454 sequencing

- Sequencing by synthesis (tracks bases as they are added); pyrosequencing
- 300-500 bp pieces, denatured;
- link one strand to plastic bead
- copy using emulsion PCR
- beads are separated on a fibre-optic plate containing approx. 1.6 million wells;
- add sequencing reagents

from: Margulies et al. 2005 Nature 437:376
454 sequencing

- Nucleotides added release pyrophosphate, prompting luciferase & flash of light.
- Correlating flashes from each cell with nucleotides presented in flow through, computer tracks sequence growth.

From: Margulies et al. 2005 Nature 437:376
Nanopore sequencing

Example of 454 Bacteria Genome Sequence
Streptococcus canis

- Genome sequence data for putative sister groups to major pathogens often not available
  - e.g. S. pyogenes; putative sister group S. canis

- S. canis from 454 Life Sciences
  - 103 contigs, 2,191,310 bp, 98.5% coverage, 39.6% GC
  - 100% of the bases with Q40+ rating (99.99% accuracy)
canis / pyogenes genome wide alignments

match found by NUCMER:

match found by PROMER:
S. canis vs S. pyogenes
canis / pyogenes genomic content comparison

S. pyogenes (SSL1)

S. canis

S. pyogenes (5005)
Applications of Microbial Genomics
Comparative genomics and drug discovery

- Genes need prioritization
- Drug development against a single bacterial species usually impractical
- Gene products, with orthologs in humans, may lack selectivity
  - => compare genomes, find potential drug targets shared by clinically important range of taxa, & absent or divergent from human host
Molecular Epidemiology

- MLST = multi locus sequence typing; sequence of portions of 7 (or more) housekeeping genes; combination of alleles = sequence type (ST); closely related STs (differ by one or two alleles) = clonal complex

http://eburst.mlst.net/6.asp
Microarray gene / presence absence

- Genome sequence allows gene presence / absence detection across strains using microarrays
  - E.g. Combimatrix 4 X 2K microarrays
Gene / presence absence hybridization