Comparative Genomics – Presentation 1

Ivan Antonov, Robert Arthur, Erin Cook, Paul Cooper, Gabriel Mitchell, Shrutii Sarda, Aswathy Sebastian, Racchit Thapliyal
Outline

• What is comparative genomics?
• Biology background
• Motivation/goals
• Classification systems
• Some interesting biological questions
What is comparative genomics?

- The study of the relationship of genome structure and function across different biological species or strains.
- Through comparing the finished reference sequence of the target genome with genomes of other organisms researchers can identify regions of similarity and difference (proteins, RNA, etc). This can be used to infer how natural selection has acted upon these organisms.
What is comparative genomics?

• This is used to help us better understand structure and function of genes which thereby enables better strategies to be developed to combat illness and/or pathogens.

• Also used for building phylogenies and classification/taxonomy.
Comparative genomics

Functional genomics

Therapeutic screening

Congenic strains

QTL mapping

Consomic panel

Physiology

Association studies

Clinical research

Linkage studies

Knockouts

ENU mutagenesis

Functional genomics

Nature Reviews | Genetics
Haemophilus Haemolyticus

- Gram – negative
- coccobacillus
- Capable of ß-hemolysis (lost in certain strains)
  - Genes encoding hemolysis not known
- Commensal but can be pathogenic
**Haemophilus Haemolyticus**

- No growth on MacConkey agar
- Growth in culture requires exogenous hemin (oxidized ferroprotoporphyrin) (X factor) and/or nicotinamide adenine dinucleotide (NAD) (V factor)
- Growth on chocolate agar with 5-10% CO2 (capnophilia)
  - Chocolate agar contains NAD, hemoglobin, assorted vitamins (cobalamin, thiamine, hydrochloride), minerals, cysteine, glutamine, and glucose.
Similarities to NTHi

- Morphology
- Metabolism
- Genetic background
- Will not react to a-f antiserum for Hi
## Our samples – Biology

<table>
<thead>
<tr>
<th>CDC ID</th>
<th>Species</th>
<th>Disease</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>M19107</td>
<td>H. haemolyticus</td>
<td>asymptomatic</td>
<td>OP swab</td>
</tr>
<tr>
<td>M19501</td>
<td>H. haemolyticus</td>
<td>asymptomatic</td>
<td>OP swab</td>
</tr>
<tr>
<td>M21127</td>
<td>H. haemolyticus</td>
<td>pathogenic</td>
<td>blood</td>
</tr>
<tr>
<td>M21621</td>
<td>H. haemolyticus</td>
<td>pathogenic</td>
<td>knee fluid</td>
</tr>
<tr>
<td>M21639</td>
<td>H. haemolyticus</td>
<td>pathogenic</td>
<td>blood</td>
</tr>
<tr>
<td>M21709</td>
<td>H. influenzae</td>
<td>pathogenic</td>
<td>blood</td>
</tr>
</tbody>
</table>
Tree of H. genus is highly polyphyletic
(Part of family – Pasteurellaceae)
Phylogenetic tree of the Pasteurellaceae family

Suggests a more fluid/open genome – indicating that a significant amount of genetic diversity is probably due to horizontal gene transfer between the various species/strains

Why do we care?

• *H. haemolyticus* is an emerging pathogen
• The first task from a public health perspective is to accurately identify pathogens
• There is some deliberation over whether or not existing typing schemes can distinguish *H. haemolyticus* from *H. influenzae*
Typing schemes of yore

A series of experiments induces a walk down a binary tree until one arrives at a satisfactory answer. This can be time consuming, expensive and, critically, not accurate. This is certainly the case for *H. haemolyticus* strains, for which it is now apparent that the absence of hemolytic activity is not a reliable indicator for the absence of the strain.

Chart from: http://education.med.nyu.edu/alex coursework/microbiology/infect-disease/Gram_Neg_Bacilli5.html
Dealing with constraints

colony morphology, metabolic assays
Sequencing specific targets
whole genome sequencing

Whole cell gamma ray chemical tomography

more feasible

more resolution
Dealing with constraints

colony morphology, metabolic assays
Sequencing specific targets
whole genome sequencing
Whole cell gamma ray chemical tomography

more feasible

more resolution

genA, genB, genC, genD, genE
How would classification work?

Based on gene sequence divergence

Based on presence or absence of genes
MLST Outline

• Why use multiple-loci approach?
• What is it’s pre-decoessor?
• Why should it work?
• How does it work? What does it require?
• Examples of General MLST.
• Example of *Haemophilus Influenzae*.
• MLST Considerations
Why use Multiple-loci approach?

• Typing of multiple-loci based on housekeeping genes (typically enzymes) with database

• a nucleotide sequence based approach for unambiguous characterization

• Allows for various evolutionary predictors to separate strains
Predecessors?

- Multilocus Enzyme Electrophoresis (MLEE) - based off conserved enzymes with little variation (allozymes). *Phenotype based on electrophilic motility*
  - Enzymes are used because of their strict requirements (mobility, electrostatic forces...)

- Previous techniques: serotyping, monoclonal antibody typing, biotyping, bacteriophage typing, fimbriation typing, resistotyping, gel electrophoresis, whole protein extract electrophoresis, outer membrane protein electrophoresis...

- DNA-DNA hybridization (DDH) is not database driven, time consuming, expensive, uses dangerous chemicals, and typically requires 60-70% similarity.

- rRNA has been shown to be too conserved at times.

M. Richter, hisfting the genomic gold standard for the prokaryotic species definition. PNAS 2009.
What does it require?

• Semi-conserved sequences with enough variation to differentiate species.

How does it work?

• Uses primers of conserved sequences to extract and compare the sequences based on historical data.
Why would it work?

- **Point Mutations**

- **Point Accepted Mutation (PAM Matrix) and BLOSUM Matrix**
How do they traditionally derive the loci?

- Historically (database with near 50 years worth of alleles)
- Conserved metabolic pathways
- Proximity to conserved regions
General Loci Examples

• Staphylococcus aureus:
  – carbamate kinase (arcC)
  – shikimate dehydrogenase (aroE)
  – glycerol kinase (glpF)
  – guanylate kinase (gmk)
  – phosphate acetyltransferase (pta)
  – triosephosphate isomerase (tpi)
  – acetyl coenzyme A acetyltransferase (yqiL)

• Vibrio vulnificus:
  – Glucose-6-phosphate isomerase (glp)
  – DNA gyrase, subunit B (gyrB)
  – Malate-lactate dehydrogenase (mdh)
  – Methionyl-tRNA synthetase (metG)
  – Phosphoribosylaminoimidazole synthetase (purM)
  – Threonine dehydrogenase (dtdS)
  – Diaminopimelate decarboxylase (lysA)
  – Transhydrogenase alpha subunit (pntA)
  – Dihydroorotase (pyrC) Tryptophanase (tnaA)
MLST- Haemophilus Influenza

<table>
<thead>
<tr>
<th>Locus</th>
<th>Gene product</th>
<th>Length of sequenced fragment (bp)</th>
<th>No. of alleles</th>
<th>No. of variable sites</th>
<th>% Sequence diversity [range (mean)]</th>
<th>dS/dN ratio</th>
<th>TIGR identifier</th>
<th>Chromosomal location (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adk</td>
<td>Adenylate kinase</td>
<td>477</td>
<td>23</td>
<td>50</td>
<td>0.2–5.8 (2.5)</td>
<td>23.2</td>
<td>HI0349</td>
<td>376</td>
</tr>
<tr>
<td>apG</td>
<td>ATP synthase F1 subunit gamma</td>
<td>447</td>
<td>26</td>
<td>35</td>
<td>0.2–4.2 (1.9)</td>
<td>9.1</td>
<td>HI0480</td>
<td>502</td>
</tr>
<tr>
<td>frdB</td>
<td>Fumarate reductase iron-sulfur protein</td>
<td>489</td>
<td>26</td>
<td>50</td>
<td>0.2–5.1 (2.6)</td>
<td>119</td>
<td>HI0834</td>
<td>883</td>
</tr>
<tr>
<td>fucK</td>
<td>Fucokinase</td>
<td>345</td>
<td>22</td>
<td>32</td>
<td>0.3–5.5 (2.5)</td>
<td>5.3</td>
<td>HI0613</td>
<td>644</td>
</tr>
<tr>
<td>mdh</td>
<td>Malate dehydrogenase</td>
<td>405</td>
<td>36</td>
<td>50</td>
<td>0.2–6.7 (3.3)</td>
<td>29.5</td>
<td>HI1210</td>
<td>1,277</td>
</tr>
<tr>
<td>pgi</td>
<td>Glucose-6-phosphate isomerase</td>
<td>468</td>
<td>32</td>
<td>73</td>
<td>0.2–6.8 (4.3)</td>
<td>15.7</td>
<td>HI1576</td>
<td>1,645</td>
</tr>
<tr>
<td>recA</td>
<td>RecA protein</td>
<td>426</td>
<td>23</td>
<td>36</td>
<td>0.2–5.9 (2.6)</td>
<td>&gt;100</td>
<td>HI0600</td>
<td>622</td>
</tr>
</tbody>
</table>

Used to show phylogenetic relationships between encapsulated and noncapsulated HI

MLST Considerations

• What are the pathway commonalities?
• Previous studies have had problems with using MLST due to the fucK operon being missing. Other studies have omitted results without this operon.
• Recent papers often use many methods to characterize HI (rRNA, DNH, MLST)
• How will our small sample size effect our ability to properly use MLST (previous HI paper has 131 isolates?)
• What other sequence based traits can be used to differentiate our species?
• Possibly use in combination with other computer based approaches like (modern DDH called ANI)
The second goal: understanding functions
Genetic Diversity

- **Genetic diversity of NTHi**
  Out of 242 strains analyzed by Multilocus enzyme electrophoresis, all 65 NTHi have a distinct electrophoretic type (ET), 177 typeable Hi belong to 29 ETs.

- **Intra- and inter- species genetic diversity**

- **Naturally competent** = Ability of bacteria to take up extracellular DNA
Mechanisms of genetic diversity: driven by vertical and horizontal gene transfer

**Vertical**
- Point mutation (light or chemical induced)
- Inversion
- Spontaneous deletion

**Horizontal**
- Transformation (chromosomal DNA)
- Transposon mutagenesis
- Transduction (phage)
- Conjugation (plasmid)
- Transformation (plasmid)
Transformation

- A process of direct uptake, incorporation and expression of exogenous genetic material from its surrounding.

- Plasmid, exogenous chromosomal DNA fragments, and transposon.

- Natural transformation/chemical or electrical transformation.
Process of Horizontal Gene Transfer (HGT)

1. Entry into the transfer process
   - Release of naked DNA
   - Integration of plasmid into chromosome
   - Interaction with mating-pair formation apparatus
   - Presence of pac sites
   - Packaging into phage particle

2. Selection of recipient
   - Uptake sequences in DNA
   - Binding of naked DNA
   - Phage receptor specificity
   - Pilius specificity
   - Surface exclusion

3. Uptake + successful entry
   - Restriction
   - Antirestriction systems
   - Selection against restriction sites

Recipient

4. Establishment
   - Replication
   - Integration
   - Homologous recombination
   - Illegitimate recombination

Christopher M Thomas and Kaare M Nielson, Nature Reviews 2005
Questions of Interest

- Hi is naturally competent. Genes involved in this process have been identified. Competency genes in Hhae?
- Level of natural competence in Hhae?
- Putative sources of Hhae genome Content?
- What % of Hhae genome comes from other bacterial species?
Barriers to HGT – Species Specificity

Restriction Enzymes (RE)

• REs/RM (Restriction -Modification) - mechanisms of genome defense to recognize and destroy (break into pieces) foreign DNA

• Two opposing functions:
  - protecting the host DNA against restriction by methylating the DNA within specific target sites
  - “restricting”, i.e. degrading any unmodified DNA that may enter the cell
• Allelic diversity and phase variable nature of RM has implications in the fitness of *H. influenzae*

• DNA restriction and modification systems in Hhae?
Hemolysis

- Hh has beta-hemolysis capabilities
- One of the means of differentiation between Hh and Hi
- Literature says Hh can lose hemolytic capability
Hemolysis Genes

- Unknown in Hh!
- *Staphylococcus aureus* and *S. pneumoniae* may provide clues
- Homology to genes in these species?
- Hemolysis capability might be encoded on plasmid?  --Leonard Mayer
Plasmids

• Can carry genes that give specific functional capabilities
• How will we find them?
  – Low coverage? --Lee Katz
  – Smaller contigs? --Leonard Mayer

• PLASMID (online database):
  http://plasmid.med.harvard.edu/PLASMID/
Strategy

• Naïve strategy
  – Find all X in Hh that is not present in Hi
  – Find all X in Hi that is not present in Hh
  – Can it be this easy?

• Higher order?
  – Find all (X,Y) in Hh...

• Sequence variation and/or homology?

• Difference in pathways?
Tools

• Pathway tools – output from annotation group
• Jspecies – sequence comparison tool
• GenomeBlast – Small genome comparison
• SynBrowse, SynView – Synteny
• eBURST – uses MLST database
• Alien Hunter, Darkhorse – HGT
• LOTS MORE!!