Comparative Genomics

Preliminary Results
April 4, 2016

I. Methods
   A. Phylogeny Methods
   B. Whole Genome Methods
   C. Horizontal Gene Transfer

II. Preliminary Results
What is the goal?

Develop a typing scheme for various strains of Nontypeable *Haemophilus influenzae* (NTHi)
Methods

● Phylogeny based
  ○ rpoS and other phenotypic markers
  ○ SNP (core and accessory)
  ○ Whole genome
  ○ MLST (housekeeping and accessory genome)

● Whole genome based
  ○ ANI / MASH
  ○ MUMmer dotplots
  ○ Pan-, Core-, and Shell-genome
  ○ Hierarchical clustering of orthologs

● Horizontal Gene Transfer
Phylogeny Methods

- rpoS marker
- kSNP
- Whole genome
- MLST
16S rRNA tree

In the laboratory of Woese, it was shown that phenotypic relationship between all life-forms can be determined by comparing the stable part of the genetic code- with 16S rRNA bring one of the candidates.

Constructing the tree:

- RNAmer was run to identify the 16S rRNA genes
- Corresponding sequences were aligned using CLUSTALw
- Tree was constructed using MEGA
kSNP- SNP tree

1. List canonical k-mers and counts for each genome
2. Remove singleton k-mers from unassembled genomes
3. Remove k-mers with allele conflicts within a genome from k-mer list for that genome
4. Merge sort remaining k-mers across genomes
5. Find SNP loci, k-mers with allelic variation across genomes
6. Find allele variant within each genome by comparing k-mer lists in steps 5 and 3
7. Find SNP position in finished genomes using MUMmer
8. Create SNP matrix and build trees (parsimony, MJ, ML, core, majority fraction)
9. Cluster SNPs by genome groups and label tree nodes with allele counts
10. Annotate SNPs with protein and other Genbank information

Gardner and Hall, 2013
Whole Genome Tree

- Pan-genome tree: full complement of genes in a set of genomes
- Method used: Get_homologues
  - Builds on orthology calling methods for the analysis
  - Based on sequence similarity
  - Clusters homologous gene families using bidirectional best hit, COG Triangles or OrthoMCL clustering methods
- Set of genomes: 64 assembled + all *Haemophilus* references
Multilocus Sequence Typing (MLST)

- Molecular typing technique whereby a number of housekeeping genes (loci) are sequenced.
- Each sequence for a given locus is screened for identity with already known sequences of that locus.
- In case seven housekeeping genes are studied, each strain is characterized by a profile of seven allele numbers.
Whole Genome Methods

- ANI / MASH
- MUMmer dotplots
- Pan-, Core-, and Soft-core genome
- Hierarchical clustering of orthologs
ANI Methods

- ANI = Average Nucleotide Identity
  - Compares genetic relatedness of two genomes
  - *in silico* alternative to DNA-DNA Hybridization (DDH)
  - Genome sequences are split into 1020 bp regions and then fragments of each genome are compared to each other to determine mean identity of matches
  - Based on BLASTn method (ANIb) or MUMmer algorithm (ANIm)
  - ANI values >95% are considered equivalent to the 70% DDH threshold
MASH Methods

- **MASH**: Fast genome and metagenome distance estimation using MinHash
- MinHash locality-sensitive hashing reduces large sequences to compressed representative sketches to estimate the similarity of the original sequences
- Mash distance correlates strongly with ANI
- Provides two main functions to compare sequences:
  - **Sketch**: Converts sequences into a MinHash sketch
    - Sketch: Equivalent to 1020 bp fragments used in ANI
  - **Distance**: Compares two sketches and gives estimate of Mash distance
    - Mash distance: Estimates evolutionary distances between sequences
MUMmer Methods

- Whole genome alignment tool used to compare closely related sequences

- Detects differences between two microbial genomes: SNPs, insertions/deletions, differences in number and location of repeat elements and tandem repeats, and regions repeated in only one of the two genomes

- MUMs = Maximal Unique Matches shared between two sequences

- MUMmer generates a list of all matches shared between two genomes
● MUMmer plot: Alignment dotplot where the two sequences are displayed on each axis and points are plotted at the positions where the sequences exhibit similarity

● Perfect alignment between two genomes would completely fill positive diagonal

http://mummer.sourceforge.net/manual/AlignmentTypes.pdf
Ortholog Clustering Methods

● Orthologous genes:
  ○ Genes that diverged from a common ancestor after a speciation event
  ○ Conserve functions across organisms => helpful in comparative studies

● Cluster protein/nucleotide sequences into orthologous groups based on sequence similarity

● LS-BSR: Large Scale Blast Score Ratio
  ○ Compares bacterial genomes and returns a matrix containing CDS name and BSR value
  ○ BSR values for each CDS indicate relative level of relatedness of CDS between genomes
  ○ Matrix can be parsed to study genetic relationships between bacterial genomes of interest

● GET_Homologues
  ○ Clusters genes using bidirectional best-hit, COGtriangles, or OrthoMCL algorithms
Detection of Horizontal Gene Transfer

- Input proteins from Prodigal while we wait for prediction group’s final GFF files
  - Oversampling, which is fine
Preliminary Results
Horizontal Gene Transfer and you

- Prediction group predicted two HGT outliers, M10540 (0 HGT) and M36564 (349 HGT)
  - We find both M10540 and M36564 to be approximately normal with 73 and 65 genes, respectively
  - Sample mean = 68
- We predict M07055 and M25364 as outliers with 443 and 312 genes, respectively
Horizontal Gene Transfer and you
Horizontal Gene Transfer and you
Horizontal Gene Transfer and you
Horizontal Gene Transfer and you

Pasteurellales ($\mu=56$)

Lactobacillales

Pseudomonadales

Burkholderiales

Enterobacteriales

Neisseriales ($\mu=8$)

Bacillales
Horizontal Gene Transfer and you

<table>
<thead>
<tr>
<th>Pasteurellales (µ=56)</th>
<th>Bacillus</th>
<th>Lactobacillales</th>
<th>Enterobacteriales</th>
<th>Burkholderiales</th>
<th>Pseudomonadales</th>
<th>Neisseriales (µ=8)</th>
<th>Bacillales</th>
</tr>
</thead>
<tbody>
<tr>
<td>M07095</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M12636</td>
<td>M26547</td>
<td>M28239</td>
<td>M12882</td>
</tr>
<tr>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
<td>M12636</td>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
</tr>
<tr>
<td>M18263</td>
<td>M26547</td>
<td>M12636</td>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
</tr>
<tr>
<td>M26547</td>
<td>M12636</td>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
</tr>
<tr>
<td>M12636</td>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
<td>M12636</td>
</tr>
<tr>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
<td>M12636</td>
<td>M28239</td>
</tr>
<tr>
<td>M12882</td>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
<td>M28239</td>
<td>M12636</td>
<td>M12882</td>
</tr>
<tr>
<td>M23922</td>
<td>M28621</td>
<td>M18263</td>
<td>M26547</td>
<td>M12636</td>
<td>M28239</td>
<td>M12882</td>
<td>M23922</td>
</tr>
</tbody>
</table>
Horizontal Gene Transfer and you

358 genes

Pasteurellales (μ=56)
Lactobacillales
Pseudomonadales
Burkholderiales
Enterobacteriales
Neisseriales (μ=8)
Bacillales
Different species
Same species
Very closely related
ANI - M19501 and M25364

Different species
Same species
Very closely related

H. haemolyticus

Minor Haemophilus species

M19501

M25364
ANI - Serotype F-like isolates

- Different species
- Same species
- Very closely related

M07066
M10540
KR494 (*Hi* serotype F)
M07055
M04827
M04744
M15301
M36564
Tree color schemes

*H. haemolyticus*
  (ANI Group 1)

Other *Haemophilus*
  (ANI Group 1)

*Hi* serotype F-like
  (ANI Group 2)

*NTHi* group 1
  (ANI Group 3)

*NTHi* group 2
  (ANI Group 4)

*Hi* serotype b

Strains without clear grouping

M25364 - Unknown species
16S rRNA Tree
MLST tree

*H. haemolyticus* - (ANI Group 1)

Other *Haemophilus* - (ANI Group 1)

*Hi* serotype F-like - (ANI Group 2)

NTHi group 1 - (ANI Group 3)

NTHi group 2 - (ANI Group 4)

*Hi* serotype b

Strains without clear grouping
**SNP tree**

*H. haemolyticus* - (ANI Group 1)

*Other Haemophilus* - (ANI Group 1)

*Hi serotype F-like* - (ANI Group 2)

NTHi group 1
(ANI Group 3)

NTHi group 2
(ANI Group 4)

*Hi serotype b*

Strains without clear grouping
SNP tree, all *Haemophilus*

RefSeq NTHi and CDC strains
NTHi are intermingled with typeable strains
Whole genome tree

H. haemolyticus -
(ANI Group 1)

Other Haemophilus -
(ANI Group 1)

Hi serotype F-like -
(ANI Group 2)

NTHi group 1
(ANI Group 3)

NTHi group 2
(ANI Group 4)

Hi serotype b
Strains without clear grouping
Dot Plots

Reference and CDC assembly selected from different groups along ANI

M29697: *NT H. influenzae*  
(ANI Group 3)

M07055: Serotype F-like isolate  
(ANI Group 2)

M11818: *H. haemolyticus*  
(ANI Group 1)

M25364: Different species  
(ANI Group 1)

M25384: Different species  
(ANI Group 1)
MASH

- Different species
- Same species
- Very closely related
MASH

Different species
Same species
Very closely related

Grouping are similar to ANI, but not as distinct.

Related to MinHash subsampling?

*H. haemolyticus* -
(ANI Group 1)

*Other Haemophilus* -
(ANI Group 1)

*Hi* serotype F-like -
(ANI Group 2)

*NTHi* group 1
(ANI Group 3)

*NTHi* group 2
(ANI Group 4)

*Hi* serotype b
Strains without clear grouping
In silico phenotyping

Heatmap colorkey
- double positive
- slighly positive
- negative

Sample colorkey
- dark red
- orange
- yellow
- green
- dark blue

Phenotype colorkey
- glucose
- sucrose
- maltose
- inulin
- melibiose
- galactose
- raffinose
- starch
- xylan
- pullulan
- carboxymethyl cellulose
- pullulanase

Alkaline phosphatase
Oxidase
Colla-stain-polymyxin susceptible
Fermentative
Glucose fermenter
Bacillus subtilis
Gram negative
Nitrate to nitrite
Catalase
Urea Hydrolysis
Capnophilic
Malolactic
Growth in KCN
Indole
Guanidine decarboxylase
Growth at 42°C
Ethanol susceptible
Xylose
Caffeine
GNPG (beta galactosidase)
Gluconate from glucose
Hydrogen sulfide
Lactose
L-Rhamnose
L-Arabinose
Bifidobacterium
D-Sorbitol
Casein hydrolysis
Coagulate production
Arginine dihydrolase
Starch hydrolysis
Cysteine - parasites or chains predomin
Methyl red
Trehalose
Escherichia hydrolysis
Beta hemolysis
Coccus - clusters or groups on plate
Aerobic
Growth in 6.5% NaCl
mycobacterial
Madura
Acetate utilization
Nitrate to gas
Lipase
Pyruvate carboxylase
Nitrates in milk
Glycerol
Growth on MacConkey agar
Mucic acid utilization
Gelatine
Tartaric utilization
Glucose or dextrin
Anaerobe
Gram positive
Yellow pigment
Escherichia formation
Methaneroxidase
Methyl red
Urease decarboxylase
Coccus
Gelatin hydrolysis
Salicine
D-Mannitol
Salmonella
O-Rhamnose
Growth in ordinary blood agar
In silico phenotyping

Shared by all

Shared by many not-Hi
Pan-, Core-, and Soft-core genome

- **Pangenome**
  - All the unique genes found in a set of genomes
  - Gives an idea of diversity

- **Core genome**
  - Genes found in every sample
  - Targets for core MLST

- **Soft-core genome**
  - Core genome + genes present in a super-majority of genomes (e.g. present 90% of samples)
  - Targets for core+accessory MLST
Clusters by method - CDC strains

COG
- 542
- 11
- 0

OMCL
- 4345
- 0
- 594

BDBH
- 887
Ortholog clusters - CDC strains

Total clusters = 5797

- cloud
- shell
- soft core
- core
Unique clusters per genome - CDC strains
Clustering methods - CDC + reference set

COG: 1287
OMCL: 1220
BDBH: 14

Intersection:
- COG and OMCL: 578
- COG and BDBH: 11
- OMCL and BDBH: 13
- All three: 14
Hierarchical clustering of orthologs from the genus Haemophilus
Conclusion

- Previously described global and local measures are insufficient for differentiation of NTHi from Hi
- Current work:
  - Virulence genes related to phenotypes are being explored
  - Shared regions unique to NTHi extracted from orthologous clusters
  - HMM-SOM mediated selection of distinguishing regions
Exercise


Short link: https://goo.gl/gL5zmR

Due: Monday, April 17 2016 by midnight
References


Supplementary data