TAG  
(Transposon Associated Genome)

Rapidly mapping the bacterial genetics of vertebrate intestinal colonization

Zac Stephens
Institute of Molecular Biology - Guillemin Lab
Microbes and Humans

- 100 x 10^{12} Microbial Cells
  - ~8 meters long intestine.

- ~150 X our own protein-coding genes.

- >160 spp. / individual
  - Great variation.

- Terrible models for vertebrate microbiotas.
Fish Guts

Great models for vertebrate microbiotas:

- Similar architecture to mammalian intestines.
- Rapidly develop. Transparent in early stages.
- Innate and adaptive immunity.
- Tractable:
  - A 6 day old larval fish gut (~0.5 x 2 mm) contains ~100,000 bacteria.
The Intestinal Microbiota

- Immune system development/education.
- Gut development and function.
- Dietary - Complex polysaccharide fermentation.
The Aim of TAG

- How do these bacteria get there? What common features allow different species to survive inside a host?

- Systematically disrupt the 1,000s of genes in each of 100s of bacteria to identify genes required for colonization and survival with a vertebrate host.
TAG Libraries

Disruption by transposon.

Gene A

OE

OE

Gene A

AGCTCGTTGGGCCGGGCAGATG TGTATAAGAGACA NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

60 bp
TAG Libraries

AGCTCGTGGGCCGCGCCAGATGTTGATAAGAGACANNNNNNNNNNNNNNNNNNNNNNNNNNN

Molecular Barcode

13 bp target verification

25 bp Genomic Tag

Informs us which gene is disrupted.
<table>
<thead>
<tr>
<th>ID</th>
<th>@GRC13_0025_FC:1:1:1578:1023#0/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>ATGTGTGTGGCAGATTGATAAGAGACAAACTGTCGCAAAAGTTGTTAAAC</td>
</tr>
<tr>
<td>ID</td>
<td>+GRC13_0025_FC:1:1:1578:1023#0/1</td>
</tr>
<tr>
<td>Quality</td>
<td>BQQPPXWWX_bQQQQ_bb_bb__b__b________bb_b_____b_YMYYY________________</td>
</tr>
</tbody>
</table>

**Raw Illumina Sequence Data (fastq)**
Illumina Sequence Data (fastq)

@GRC13_0025_FC:1:1:1578:1023#0/1
ATGTGTGTTGGGCGCCGCCAGATGTATAAGAGACA
+GRC13_0025_FC:1:1:1578:1023#0/1
BQQPPXWWXX_bQQQQ_bb_bb_b_b_b___bb_b___b_YMYYY__________

@GRC13_0025_FC:1:1:1649:1023#0/1
GGCGCTGTTGGGCGCCGCCAGATGTATAAGAGACA
+GRC13_0025_FC:1:1:1649:1023#0/1
BRQOOVVVVW________QQ____________________

@GRC13_0025_FC:1:1:3313:1022#0/1
ATCTCTGTTGGGCCGCAGATGTCTCAGTGGGCAACCATGATCGCCTCATC
+GRC13_0025_FC:1:1:3313:1022#0/1
BOHHKMLNNQ____________________W_________________V_

@GRC13_0025_FC:1:1:3565:1023#0/1
ATCTCTGTTGGGCGCCGCCAGATGTGTGCAAACGTGAAGCAATCAGAACTGTGCATAGAT
+GRC13_0025_FC:1:1:3565:1023#0/1
BOHHKMLNNQ____________________WBBBBB_____YWPPB

@GRC13_0025_FC:1:1:3565:1023#0/1
ATCTCTGTTGGGCGCCGCCAGATGTATAAGAGACAGCAATCAGAACTGTGCA
+GRC13_0025_FC:1:1:3565:1023#0/1
BOHHKMLNNQ____________________WBBBBB_
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.

1) ATGTGTGTTGGGCGCGCCAGATGTGATAAGAGACAACTGTGCAAAAGTTGTTATCG
2) ATGTGTGTTGGGCGCGCCAGATGTGATAATAGACAGCTGCTCTGACTGTTGTTAAAC
3) ATGTGTGTTGGGCGCGCCAGATGTGATAACAGACAGCAATCAGAACTGTCA

1) GTGGGTGTTGGGCGCGCCAGATGTGATAAGAGACAACTGTGCAAAAGTTGTTATCG
2) GTGGGTGTTGGGCGCGCCAGATGTGATAAGAGACAGCTGCTCTGACTGTTGTTAAAC
3) GTGGGTGTTGGGCGCGCCAGATGTGATAAGAGACAGCAATCAGAACTGTCA
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.

<table>
<thead>
<tr>
<th></th>
<th>ACTGTCGCAAAAAGTTGTTATCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AGCTGCTCTGACTGTTGTTAAAC</td>
</tr>
<tr>
<td>2</td>
<td>GCAATCAGAACTGTCA</td>
</tr>
</tbody>
</table>

1) ACTGTCGCAAAAAGTTGTTATCG
2) AGCTGCTCTGACTGTTGTTAAAC
3) GCAATCAGAACTGTCA
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.
2) BLAST “tags” against genome.

Gene A

ACTGTCGCAAAAAGTTGTTATCGGTCTCATCTGTCATAATAGCTTTCGGTTCTGTG
TGACAGCGTTTTCAACAATAGCCAGAGTAGACAGTATTATCGAAGCCCCACAAGAC

Gene B

ACTGTCGCAAAAAGTTGTTATCGGTCTCATCTGTCATAATAGCTTTCGGTTCTGTG
TGACAGCGTTTTCAACAATAGCCAGAGTAGACAGTATTATCGAAGCCCCACAAGAC

ACTGTCGCAAAA
AGCCAGAGTAGACAG
CAGCGTTTTCAAC
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.

2) BLAST “tags” against genome.

3) Determine genes enriched in particular pools.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Input Pool Tags</th>
<th>Output Pool Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>12%</td>
<td>13%</td>
</tr>
<tr>
<td>Gene B</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.

2) BLAST “tags” against genome.

3) Determine genes enriched in particular pools.

4) Extract orthologous gene information and compare to other species that have been TAGged.

- Common proteins required? Specific to each species?
- Common processes required (cell wall synthesis, metabolic pathways, quorom-sensing, etc.)?
TAG Pipeline

1) Basic string manipulations – filter (allow some mismatch), sort and extract “tag”.

2) BLAST “tags” against genome.

3) Determine genes enriched in particular pools.

4) Extract orthologous gene information and compare to other species that have been TAGged.

5) Visualize insertions clusters as tracks against genome.