PanCake: A Data Structure for Pangenomes

Corinna Ernst & Sven Rahmann

Genome Informatics
Institute of Human Genetics
University of Duisburg-Essen

GCB 2013
Terminology

Tettelin et al., 2005

Pangenome $\equiv$ global gene/subsequence repertoire within a collection of phylogenetically related genomes

- Which features appear in all available sequences?
  $\Rightarrow$ **core genome**

  > Genome1
  AGCCGATGAAATAGTGA
  > Genome2
  GATCAACGATTCCGATA
  > Genome3
  ACGATAGGATGAAAGGAT

- Which features appear exclusively in one genome?
  $\Rightarrow$ **singletons**

  > Genome1
  AGCCGATGAAATAGTGA
  > Genome2
  GATCAACGATTCCGATA
  > Genome3
  ACGATAGGATGAAAGGAT
Goals of Pangenome Analysis

- delineation of prokaryotic species
- phylogenetic analysis
- identification of genes indispensable to cell survival
- detection of strain-specific characteristics, e.g. virulence, synthesis of certain metabolites, . . .
- annotation refinements
Existing Pangenome Analysis Tools

- **EDGAR**
  - Blom et al., 2009

- **Prokaryotic Genome Analysis Tool**
  - Brittnacher et al., 2011

- **Comprehensive Microbial Resource**
  - Davidsen et al., 2010

- **PGAP**
  - Yongbing et al., 2012

- **CaBLAST**
  - Loh et al., 2012

- database approaches limit analysis to sets of pre-processed strains
- require annotation data
PanCake – Development Goals

- representation of pangenomes as a persistent data structure
- singleton and core identification independent of pre-annotation
- similarity information arises from incremental inclusion of pairwise alignment information
The Idea behind PanCake

- pangenomes organized as a graph of **Shared Features**
- each Shared Feature is given by a set of related subsequences (**Feature Instances**) originating from diverse chromosomes
- chromosome reconstruction by iteration over sequence of succeeding Feature Instances
Sequence Retrieval by Edit Operations
Brandon et al., 2009
Loh et al., 2012

Example of sequence encoding by reference sequence and edit operations.

<table>
<thead>
<tr>
<th>Whole Numbers</th>
<th>Negative Numbers</th>
<th>Upper-case letter</th>
<th>Lower-case letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Whole Numbers</th>
<th>–</th>
<th>perfect matches of given length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Numbers</td>
<td>–</td>
<td>deletions of given length</td>
</tr>
<tr>
<td>Upper-case letter</td>
<td>–</td>
<td>substitution</td>
</tr>
<tr>
<td>Lower-case letter</td>
<td>–</td>
<td>insertion</td>
</tr>
</tbody>
</table>
Building the PanCake Data Structure

Adding alignment

chr1: 4 CCC 6
chr2: 5 CCC 7
Building the PanCake Data Structure

Adding alignment

chr1: 4 CCCAC 8
chr2: 12 CCGAC 8  ⇒  4 CCC 6 & 7 AC 8
12 CCG 10 & 9 AC 8
Core & Singleton Identification

Core Regions corresponding to chr1

chr1 4-8

ATG
chr1 1-3 [3]

TCCT
chr1 4-6 [3]

AC
chr1 7-8 [2]

GTGA
chr2 1-4 [4]

TTAA
chr1 9-12 [4]

CCC
chr2 5-7 [3]

AC
chr2 8-9 [2]

TAA
chr2 13-15 [3]

chr2 10-12 [2,G]
Core & Singleton Identification

Core Regions corresponding to chr2

chr2 5-7
chr2 8-12
## Results

<table>
<thead>
<tr>
<th>Genus</th>
<th>FASTA size [MB]</th>
<th>number of alignments(^1)</th>
<th>PanCake size [MB]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pseudomonas</strong> (3 strains)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. aeruginosa</em> PAO1</td>
<td>19.7</td>
<td>1405</td>
<td>7.7</td>
</tr>
<tr>
<td><em>P. aeruginosa</em> LESB58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. aeruginosa</em> UCBPP-PA14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yersinia</strong> (8 strains)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> Angola,</td>
<td>38.5</td>
<td>324925</td>
<td>8.9</td>
</tr>
<tr>
<td><em>Y. pestis</em> Antiqua</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> CO92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> KIM 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> Microtus 91001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> Nepal 516</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> Pestoides F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Y. pestis</em> Z176003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Burkholderia</strong> (10 strains)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. mallei</em> ATCC 23344,</td>
<td>66.2</td>
<td>147344</td>
<td>22.0</td>
</tr>
<tr>
<td><em>B. mallei</em> NCTC 10229</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. mallei</em> NCTC 10247</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. mallei</em> SAVP1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. pseudomallei</em> K96243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. pseudomallei</em> 1026b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. pseudomallei</em> 1106a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. pseudomallei</em> 1710b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. pseudomallei</em> 668</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B. thailandensis</em> E264</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^1\) computed by nucmer version 3.07 with parameter --maxmatch
Results

Yersinia, 8 strains
Annotation Effects on Core & Singleton Analysis

Diagram:
- Genome 1
  - Gene 1.A
- Genome 2
  - Gene 2.A
  - Gene 2.B
Results

Yersinia, 8 strains

Singletons

- PanCake
  - 0.00% 0
  - 0.24% 652
  - 37.85% 101592

Core Regions

- PanCake
  - 0.00% 0
  - 0.24% 652
  - 37.85% 101592
Dealing with Gene Duplications

Genome 1

Gene 1.A
Gene 1.B

Genome 2

Gene 2.A
Conclusion

PanCake offers

- a persistent data structure for pangenomes
- core & singleton analysis independent of pre-annotation
- considerably reduced storage requirements in comparison to pure sequence data

https://bitbucket.org/CorinnaErnst/pancake
AT-GTGTTTCATGT
ATAGTGTT-\_A-GT
A-TAGTGTTTAGT
AGTCGT-TTAGT
⇒
A-T-GTGTTTCATGT
A-TAGTGTT-\_A-GT
A-TAGTGTTTAGT
AGTCGT-TTAGT
⇒
A-T-GTGTTTCATGT
AGTCGT-TT-\_A-GT