Computational assembly for prokaryotic sequencing projects

Lee Katz, Ph.D.
Bioinformatician, Enteric Diseases Laboratory Branch (EDLB)
Enteric Diseases Bioinformatics Team (EDBiT)
January 30, 2017

Disclaimers
The findings and conclusions in this presentation have not been formally disseminated by the Centers for Disease Control and Prevention and should not be construed to represent any agency determination or policy. The findings and conclusions in this [report/presentation] are those of the author(s) and do not necessarily represent the official position of CDC.
2011 to present
Vibrio, Campylobacter, Escherichia, Shigella, Yersinia, Salmonella

ENTERIC DISEASES LABORATORY BRANCH
Thanks to John Besser for letting me borrow this slide
Thanks to John Besser for letting me borrow this slide
Outline

• Illumina sequencing
• Reads
  – Quality control (Q/C)
    • Read metrics
  – Read-cleaning
• Assembly
  – Algorithms
  – Assembly metrics
• Genome-based typing

Please also see my talk last year at http://compgenomics2016.biology.gatech.edu/index.php/Lectures
I have no real-world experience and I am incompetent at everything.

But unlike any of you, I have a Ph.D., and that means you have to take me seriously.

Is pretending allowed? Totally. It all looks the same to me.
Prokaryotic Sequencing Projects

**Stages**
- Sequencing
- Assembly
- Feature prediction
- Functional annotation
- ...analysis...
- Display (Genome Browser)

**Examples**
- *Haemophilus influenzae*
- *Neisseria meningitidis*
- *Bordetella bronchiseptica*
- *Vibrio cholerae*
- *Listeria monocytogenes*


Illumina sequencing (2nd Gen)

- You have sequences from the Illumina HiSeq
  - 250bp
  - Paired end
Paired end Illumina reads

Ideal

Still good

Bad
Q/C + cleaning

READS
Q/C

• You need to know if your data are good!
• Example software that give you read metrics
  – FastQC
  – FastqQC (from AMOS)
  – Computational Genomics Pipeline (CG-Pipeline)
    • `run_assembly_readMetrics.pl`
  – BamUtils stats
Quality Control

Sequence content across all bases
The CG-Pipeline way

run\_assembly\_readMetrics.pl

<table>
<thead>
<tr>
<th>File</th>
<th>avgReadLength</th>
<th>totalBases</th>
<th>minReadLength</th>
<th>maxReadLength</th>
<th>avgQuality</th>
</tr>
</thead>
<tbody>
<tr>
<td>tmp.fastq</td>
<td>80.00</td>
<td>177777760</td>
<td>80</td>
<td>80</td>
<td>35.39</td>
</tr>
</tbody>
</table>
READ CLEANING
Why clean my reads?

• Your reads have errors
• These errors can hurt downstream analysis

• Can be categorized into
  – Trimming
  – Filtering
  – Correcting
More read cleaning and correction

• Software
  – CGP documentation: http://cg-pipeline.sourceforge.net/wiki
  – Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit
  – AMOS amos.sourceforge.net
  – Sickle https://github.com/najoshi/sickle
  – Quake http://www.cbcb.umd.edu/software/quake
  – BayesHammer http://bioinf.spbau.ru/spades/bayeshammer
  – PrinSeq
  – ... and more is out there!

• Evaluation
  – Fabbro et al 2013, “An extensive evaluation of read trimming effects on Illumina NGS data analysis”
ASSEMBLY CONCEPTS
Assembly

- Overlaps between reads or between nodes in a graph
- Generate **contigs** (contiguous sequences)
- Generate **scaffolds**
Further reading: Zhang et al 2011 *Plos One*; Miller et al 2011 *Genomics*
Overlap-layout-consensus

- Great at longer, fewer, imperfect reads, e.g., Ion Torrent, 454, PacBio
Derive consensus sequence

Derive each consensus base by weighted voting

De bruijn (kmer)

- Great at shorter, numerous, less error-prone reads, e.g., Illumina

http://www.homolog.us/blogs/blog/2012/06/17/an-intuitive-explanation-for-running-de-bruijn-assembler-with-varying-k-mer-sizes/
Another way to look at de Bruijn assembly.

Unipath graph of the 1.8-Mb genome of *C. jejuni*.

Possible paths:
- ABCDBCEFCEG
- ABCEFCDDBCEG
Another way to look at de bruijn assembly

Unipath graph of the 1.8-Mb genome of *C. jejuni*

Possible paths:
- ABCDBCEFCEG
- ABCEFCDBCEG
Recap of assembly

reads

Paired end reads

contigs

Scaffold
ASSEMBLY EVALUATION
# Assembly Metrics

How do you tell if your assembly is good?

<table>
<thead>
<tr>
<th>Metric</th>
<th>description</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly Length</td>
<td>The size of the concatenated assembly</td>
<td></td>
</tr>
<tr>
<td>Number of Contigs</td>
<td>The count of contigs</td>
<td></td>
</tr>
<tr>
<td>N50</td>
<td>The size of the contig at where half the genome is located in size &gt;N50 and half is located in size &lt;N50</td>
<td></td>
</tr>
<tr>
<td>Longest Contig</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average contig length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kmer21</td>
<td>Frequency of kmers with k=21</td>
<td><a href="http://www.homolog.us/blogs/blog/2012/06/26/what-is-wrong-with-n50-how-can-we-make-it-better-part-ii/">http://www.homolog.us/blogs/blog/2012/06/26/what-is-wrong-with-n50-how-can-we-make-it-better-part-ii/</a></td>
</tr>
<tr>
<td>GC-content</td>
<td>Percentage of the genome that is either G or C</td>
<td></td>
</tr>
</tbody>
</table>
| Assembly score    | \[
\log \frac{N50 \times percent\text{GenomeCovered}}{\text{numContigs}}\]    | CG-Pipeline/Lee Katz                                                     |
Please also see my talk last year at http://compgenomics2016.biology.gatech.edu/index.php/Lectures

Some assemblers to try

- Open source
  - SPAdes
  - Velvet
  - Mira
  - Edena
  - CG-Pipeline (uses SPAdes + Velvet)
- Proprietary
  - Geneious
  - CLC Genomics Workbench
  - Lasergene
  - Bionumerics

- These are some ideas but it is up to you, the class, to do a more thorough review!

Comparisons of Illumina assemblers

- Zhang et al 2011 “A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies”
- Assemblathon - http://assemblathon.org/
- Lin et al 2011, “Comparative studies of de novo assembly tools for next-generation sequencing technologies”
- GAGE-B: http://ccb.jhu.edu/gage_b/
- GABenchToB: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107014#pone-0107014-g003

One problem: randomly low coverage (Lander-Waterman)

- Assuming random distribution of reads and ignoring repeat resolution issues,
  - \( G = \text{genome length} \)
  - \( L = \text{length of a single read} \)
  - \( N = \text{number of reads sequenced} \)
  - \( T = \text{minimum overlap to align the reads together} \)
  - Then overall coverage is \( C = LN/G \)
  - Coverage for any given base obeys the Poisson distribution:
    \[
    P(C = y) = \frac{C^y e^{-c}}{C_y!}
    \]
  - The number of gaps (bases with 0 coverage) is:
    \[
    G = Ne
    \]

Software to try

- Mapping software
  - Small
  - BWA
  - Bowtie/Bowtie2
- SNP-calling software
  - Samtools mpileup or bcftools
  - Freebayes
  - GATK
  - VarScan
  - SoiSNP
- Must know how to use samtools for this route

QUAST: http://bioinf.spbau.ru/quast

Lander-Waterman animation

Some places in the genome will randomly not be covered. You can overcome this obstacle with large depth of reads.
Totally changing the topic

**GENOME-BASED TYPING**
Can we use WGS with epidemiology?

Phylogeny* approximates** epidemiology***!

* Depending on your evolutionary model
** In a monoclonal outbreak
*** A trace back diagram

Epidemiology - linking cases through a common food vehicle

Phylogeny – how things are genetically related through evolution
Listeria real-time project

- Sequenced every *Listeria monocytogenes* in the US
  - Starting August 2013
  - Clinical & environmental (CDC/FDA/USDA)
- Put into large tree (NCBI)
- Mile-high view
- 7,800 *Listeria monocytogenes* genomes in this tree

NCBI tree and Genome Workbench.
The flood of data is upon us

Sept 23, 2016: 91k genomes
10 taxonomic groups

Jan 27, 2015: 121k genomes
19 taxonomic groups

Isolates by Organism Type

<table>
<thead>
<tr>
<th>Organism Group</th>
<th>All Isolates</th>
<th>New Isolates</th>
<th>Clinical Isolates</th>
<th>Environmental Isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella</td>
<td>51025</td>
<td>63</td>
<td>21755</td>
<td>29270</td>
</tr>
<tr>
<td>Escherichia coli/Shigella</td>
<td>18086</td>
<td>61</td>
<td>12855</td>
<td>8131</td>
</tr>
<tr>
<td>Listeria</td>
<td>11039</td>
<td>3</td>
<td>2538</td>
<td>8101</td>
</tr>
<tr>
<td>Campylobacter</td>
<td>3781</td>
<td>123</td>
<td>1687</td>
<td>2094</td>
</tr>
<tr>
<td>Acinetobacter</td>
<td>2500</td>
<td>0</td>
<td>1819</td>
<td>769</td>
</tr>
<tr>
<td>Klebsiella</td>
<td>2125</td>
<td>5</td>
<td>1593</td>
<td>542</td>
</tr>
<tr>
<td>Vibrio Parahaemolyticus</td>
<td>790</td>
<td>131</td>
<td>349</td>
<td>449</td>
</tr>
<tr>
<td>Serratia</td>
<td>357</td>
<td>12</td>
<td>107</td>
<td>250</td>
</tr>
<tr>
<td>Elizabethkingia</td>
<td>97</td>
<td>1</td>
<td>83</td>
<td>14</td>
</tr>
<tr>
<td>Providencia</td>
<td>77</td>
<td>1</td>
<td>63</td>
<td>14</td>
</tr>
</tbody>
</table>

* for foodborne and hospital-acquired pathogens
Are two isolates in the same outbreak?

• Need to refine NCBI analysis
• Can use many methods
  – MinHash clustering with kmers
  – wgMLST
  – SNPs
    • Assembly-based (faster)
    • Raw-read based (resolution)
MinHash

- **Kmer**: a length of DNA $k$ nucleotides long

1. Shred all reads into kmers
2. How many kmers are in common?
3. Transform into a jaccard distance
   - ⇒ Gives a genomic distance
   - ⇒ With the distances, make tree

*Fast*, not phylogenetic

Analogy: comparing equal-sized shreds of paper from two books

Mashtree: https://github.com/lskatz/mashtree
Mash: http://mash.readthedocs.io/en/latest
Sourmash: https://github.com/dib-lab/sourmash
MinHash algorithm explained: Ondov et al (2016) Mash: fast genome and metagenome distance estimation using MinHash” *Genome Biology*
**MLST**

**MLST:** multilocus sequence typing

**Locus:** a place in a genome. Plural: *loci*

Identify a set of loci in the genome
Compare each locus in a genome against the set of loci
Count differences and the number of loci compared
Analogy: comparing two books, page by page

Different kinds
- 7-gene MLST
- rMLST
- *wgMLST* (whole genome MLST)
- cgMLST (core genome MLST)
- ... and more

StringMLST https://github.com/anujg1991/stringMLST
SRST2 https://github.com/katholt/srst2
BioNumerics http://www.applied-maths.com/applications/mlst
mlst https://github.com/tseemann/mlst
Salmonella 7-gene scheme
http://mlst.warwick.ac.uk/mlst/dbs/Senterica

SNPs

- **SNP**: Single nucleotide polymorphism
- **hqSNP**: high-quality SNP

- Compare individual letters in a **query** genome against the **reference** genome
- Analogy: comparing individual letters between two books

Lyve-SET github.com/lskatz/Lyve-SET
SNP-Pipeline (CFSAN) http://snp-pipeline.readthedocs.org
kSNP https://sourceforge.net/projects/ksnp
SPANDx https://github.com/dsarov/SPANDx
Snippy https://github.com/tseemann/snippy
NASP http://tgennorth.github.io/NASP
Red Dog https://github.com/katholt/RedDog
Harvest https://harvest.readthedocs.io
MUMmer http://mummer.sourceforge.net/manual
SNVPhyl http://snpphyil.readthedocs.io
RealPhy http://realphy.unibas.ch
Which algorithm should you use?

<table>
<thead>
<tr>
<th></th>
<th>MinHash</th>
<th>wgMLST</th>
<th>hqSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity</td>
<td>✓✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Outbreak-level resolution</td>
<td>✗</td>
<td>✓✓</td>
<td>✓✓</td>
</tr>
<tr>
<td>Further genomic information</td>
<td>✗</td>
<td>✓✓</td>
<td>✓✓</td>
</tr>
<tr>
<td>Minimal upfront effort</td>
<td>✓✓</td>
<td>✗✗</td>
<td>✓</td>
</tr>
<tr>
<td>Fast</td>
<td>✓✓</td>
<td>✓✓</td>
<td>✗</td>
</tr>
<tr>
<td>Easy-to-use for anyone</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Easy-to-use for bioinformaticians</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Lyve-SET

Lyve – *Listeria, Yersinia, Vibrio*, and *Enterobacteriaceae* reference lab

SET – *Snp Extraction Tool*

Some details on Lyve-SET:

For LINUX

Extensive documentation

Help options are embedded in each script

--fast option, takes ¼ the normal time

Easy to use

Modular

Most stable version: 1.1.4f

Newest version: 2.0

https://github.com/lskatz/lyve-SET
An animation of Lyve-SET

0. Pre-processing
   a) phage discovery/masking
   b) Manual identification of troublesome regions
   c) Read cleaning (Poster – Wagner et al)

1. Mapping - SMALT
   a) 95% read identity
   b) Unambiguous mapping

2. SNP calling - VarScan
   a) 75% consensus
   b) 10x depth

3. Phylogeny inferring – RAxML v8
   a) Removal of clustered SNPs
   b) Ascertainment bias model
   c) Maximum likelihood

https://github.com/lskatz/lyve-SET
https://www.sanger.ac.uk/resources/software/smalt
Lyve-SET method with more details

Comparison with other SNP pipelines

Each data point is a SNP distance as determined by Lyve-SET (x-axis) and the distance of an alternative SNP pipeline (y-axis). The slope indicates the number of SNPs per Lyve-SET SNP.

Comparison with whole-genome MLST (*Listeria monocytogenes* only)

Comparison with other SNP Pipelines and wgMLST

<table>
<thead>
<tr>
<th></th>
<th>Lyve-SET</th>
<th>kSNP</th>
<th>RealPhy</th>
<th>Snp-Pipeline</th>
<th>SNVPhyl</th>
<th>wgMLST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tree sensitivity (Sn)</strong> a</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Tree specificity (Sp)</strong> a</td>
<td>100.0%</td>
<td>90.2%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>average of Sn and Sp</strong></td>
<td>100.0%</td>
<td>95.1%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Kendall-Colijn (λ = 0)</strong> b</td>
<td>-</td>
<td>1.26E-02</td>
<td>7.51E-03</td>
<td>9.28E-03</td>
<td>9.15E-02</td>
<td>1.00E-04</td>
</tr>
<tr>
<td><strong>Robinson-Foulds</strong> b</td>
<td>-</td>
<td>3.16E-69</td>
<td>6.79E-40</td>
<td>5.39E-74</td>
<td>9.61E-49</td>
<td>1.55E-147</td>
</tr>
<tr>
<td><strong>Mantel</strong></td>
<td>-</td>
<td>0.60</td>
<td>0.77</td>
<td>0.77</td>
<td>0.79</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>SNP ratio</strong> c,d</td>
<td>-</td>
<td>0.53, 0.78</td>
<td>0.97, 0.84</td>
<td>1.61, 1.75</td>
<td>0.67, 0.84</td>
<td>0.69, 0.72</td>
</tr>
<tr>
<td><strong>goodness-of-fit (R²)</strong> d</td>
<td>-</td>
<td>0.46, 0.42</td>
<td>0.7, 0.75</td>
<td>0.77, 0.3</td>
<td>0.83, 0.68</td>
<td>0.75, 0.72</td>
</tr>
<tr>
<td><strong>Genome analyzed</strong> e</td>
<td>25.9%</td>
<td>0.1%</td>
<td>84.8%</td>
<td>0.3%</td>
<td>82.1%</td>
<td>88.2%</td>
</tr>
</tbody>
</table>

a) Average percentage from 11 outbreaks. The *S. enterica* outbreak 1203NYJAP-1 was removed as an outlier because all pipelines except wgMLST produced errors with grouping outbreak vs non-outbreak isolates. Therefore this dataset was removed from the Sn and Sp calculations as an outlier.

b) Geometric mean.

c) Number of SNPs per Lyve-SET SNP, averaged across 12 outbreaks. For wgMLST, this is the number of alleles per Lyve-SET SNP.

d) The average for 12 outbreaks. First value is for all data points; second value is for distances between only outbreak-associated genomes.

e) The average for 12 outbreaks. Percentage of the reference genome included for analysis. For wgMLST, the average percentage was calculated by obtaining each GenBank-formatted file with annotated wgMLST loci and calculating the breadth of coverage for all loci.


Comparison metrics:

Kendall and Colijn 2015, “A tree metric using structure and length to capture distinct phylogenetic signals.” arXiv

Mantel 1967, “The detection of disease clustering and a generalized regression approach” Cancer Res

Robinson and Foulds 1981, “Comparison of phylogenetic trees.” Mathematical biosciences
Lyve-SET

Lyve – *Listeria, Yersinia, Vibrio,* and *Enterobacteriaceae* reference lab
SET – Snp Extraction Tool

Who uses Lyve-SET?

• Used in labs at CDC:
  – Foodborne Diseases Laboratory Branch - EDBIT, Reference Lab (NERO), PulseNet, NARMS, Botulism Lab
  – Clinical and Environmental Microbiology Branch
  – Special Bacterial Pathogens Branch
  – Respiratory Diseases Branch
  – Meningitis and Vaccine Preventable Diseases Branch

• State health labs - Colorado, Washington

• USDA - FSIS, ARS

https://github.com/lskatz/lyve-SET
Mashtree

Useful for creating a dendrogram quickly

*Does not* create a phylogeny

https://github.com/lskatz/mashtree
Mashtree comparison

Lyve-SET v1.1.4f
kSNP v3.0.0
RealPhy v112
Snp-Pipeline v0.5.2
SNVPhyl v1.0
wgMLST BioNumeric v7.5
Mashtree v0.08
5 minutes,
41 seconds

Mashtree v0.08 with min_abundance_filter
35 minutes

Green indicates outbreak association
Removed one outlier genome in the min_abundance_filter tree that flattened the entire tree
Can’t we just estimate a trace back tree?

- TransPhylo (Imperial College/bcCDC): [http://biorxiv.org/content/biorxiv/early/2016/07/22/065334.full.pdf](http://biorxiv.org/content/biorxiv/early/2016/07/22/065334.full.pdf)
- Outbreaker (Imperial College): [http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003457](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003457)
Acknowledgements

• Every single compgenomics class since 2008
• My branch at CDC
Questions?

Lee Katz
lkatz@cdc.gov

@lskatz
github.com/lskatz

The findings and conclusions in this presentation have not been formally disseminated by the Centers for Disease Control and Prevention and should not be construed to represent any agency determination or policy.

The findings and conclusions in this presentation are those of the author and do not necessarily represent the official position of CDC.