Genome Assembly
Background and Strategy

February 6th, 2017

BIOL 7210 - Faction I (Outbreak) - Genome Assembly Group

Yanxi Chen  |  Carl Dyson  |  Zhiqiang Lin  |  Sean Lucking  |  Chris Monaco
Shashwat Deepali Nagar  |  Jessica Rowell  |  Ankit Srivastava
Camila Medrano Trochez  |  Venna Wang  |  Seyed Alireza Zamani
What... What... and How...

What do we have?
Reads for 24 *Salmonella enterica* serovar Heidelberg outbreak isolates from 2013
- Sequenced using Illumina HiSeq (2\textsuperscript{nd} Generation)
- Short paired-end reads

What do we want to do?
*De novo* assembly of the reads to obtain the whole genome for the isolates

How would we go about it?
Benchmark tools for different steps and then use one (or more) for final assembly
- Read quality control - FastQC, TrimGalore, etc.
- Assembly - Velvet, ABySS, etc.
- Assembly evaluation - QUAST, KAT, etc.
Salmonella Heidelberg

Gram negative, rod shaped bacteria

Salmonella Heidelberg: Isolate of *Salmonella enterica* that can be found in chicken, beef, horses, etc. Exposure to infected animals can lead to illness in humans.

Symptoms: People infected with *Salmonella* Heidelberg typically experience diarrhea, cramps, and fever 12 to 72 after exposure to the pathogen. The symptoms can be severe enough to require hospitalization, but with antibiotic treatment usually pass within 4 to 7 days.
Salmonella Heidelberg

Why is this isolate of interest to us?

Foodborne illness: illness resulting from exposure to contaminated food, pathogenic bacteria, viruses, or parasites.

This isolate is a commonly occurring outbreak
  - November 28, 2016
  - Dairy bull calves

Has shown signs of resistance to multiple antibiotics used in treatment
Salmonella *heidelberg* Outbreak in 2013

- 634 cases of illness reported
- 29 states including Puerto Rico affected
- 38% of cases resulted in hospitalization
- 78% of all cases were from the state of California
- Resistant to several commonly prescribed antibiotics
  - These antibiotics are not generally used to treat blood or severe salmonella infections
Salmonella *heidelberg* Outbreak in 2013

- In June 2013 PulseNet reported a cluster of illness associated with a single rare Pulse Field Gel Electrophoresis (PFGE) pattern of *Salmonella* Heidelberg
  - From non-human retail chicken
- Six more PFGE patterns were identified during the outbreak
  - Total of seven strains
  - Including one very rare pattern for *Salmonella* Heidelberg
- Outbreak was eventually associated with consumption of undercooked Foster Farms brand chicken
- Outbreak appeared to end in July 2014
DNA Sequencing

- DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule

- A: adenine  T: thymine  C: cytosine  G: guanine

- Sanger sequencing
- Developed in 1977
- Dideoxynucleotides
Next Generation Sequencing (NGS)

- High-throughput sequencing
- Low-cost
- Massive reads
- Parallelization

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing
Illumina Sequencing

Figure 2: Prepare Genomic DNA Sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Figure 3: Attach DNA to Surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Illumina Sequencing

**Figure 4: Bridge Amplification**

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

**Figure 5: Fragments Become Double Stranded**

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
Illumina Sequencing

Figure 6: Denature the Double-Stranded Molecules

Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
Illumina Sequencing

Figure 8: Determine First Base
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Figure 9: Image First Base
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.
Illumina Sequencing

**Figure 10: Determine Second Base**

The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

**Figure 11: Image Second Chemistry Cycle**

After laser excitation, the image is captured as before, and the identity of the second base is recorded.
Illumina Sequencing

Figure 12: Sequencing Over Multiple Chemistry Cycles

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Figure 13: Align Data

The data are aligned and compared to a reference, and sequencing differences are identified.
Quality Control: Why?

- Removing duplicate reads
- Trimming off barcodes
- Overlapping reads
- Checking GC content
- Library insert size
Quality Control: Tools

● **Quake**
  ○ Uses k-mer error correction framework to correct substitution sequencing errors in experiments with deep coverage

● **PRINSEQ (PREprocessing and INformation of SEQuence data)**
  ○ Filter, reformat and trim

● **NGS QC Toolkit**
  ○ Filter, trim, generate statistics and conversion between different file formats/variants of NGS data from Illumina and Solexa and Sanger format

● **Sickle**
  ○ A windowed adaptive trimming tool for FASTQ files using quality

● **KAT (Kmer Analysis Toolkit)**
  ○ Assesses data quality using Jellyfish2 counting method

● **Trim Galore**
  ○ A wrapper script that automates adapter trimming (using Cutadapt) and quality control (using FastQC)
Quality Control: Tools

- Trim Galore
  - Use Cutadapt and FastQC
  - Cutadapt searches for the adapter in all reads and removes it when it finds it
  - FastQC provides a set of analyses which you can use to give a quick impression of whether your data has any problems
  - Can trim paired-end files by 1 additional bp from the 3' end of all reads to avoid problems with invalid alignments with Bowtie 1
  - Last version released 25-01-2017
Per base sequence quality

Results from FastQC
Per base sequence content

Results from FastQC
Quality Control: Tools

- KAT (Kmer Analysis Toolkit)
  - October 2016
  - Use Jellyfish: fast, memory-efficient counting of k-mers in DNA
  - k-mer spectra: Quality of the data (level of errors, sequencing biases, completeness of sequencing coverage and potential contamination) and genomic complexity (size, karyotype, levels of heterozygosity and repeat content)
  - Pairwise comparison can be done to identify problematic samples by highlighting differences between spectra
Genome Assembly

Image source: http://www.genomicglossaries.com/presentation/SLAgenomics.asp
Genome Assembly

- Two broad classifications:
  - Reference-based assembly
    - Mapping to a reference genome
    - Each read is aligned to the reference genome
    - In this approach, the new completely different sequences are lost
  - De novo assembly
De Novo Assembly Algorithms

Reconstructing whole-genomes from the sequenced reads is usually done using one of the following three graph-based approaches

- **Greedy approach**
  - Makes locally optimal choices i.e. join reads that overlap the most
  - Ignores global information, such as paired-end reads
  - e.g. TIGR, VCAKE, etc.

- **Overlap-Layout-Consensus approach**
  - Builds overlap graph for reads, creates contigs from chains in the graph, and then chooses the most likely sequence for each contig using MSA
  - Overlap graph construction time is **slow** for a large number of reads
  - e.g. Celera, SGA, etc.

- **De Bruijn graph approach**
  - Predominantly used for assembling short reads
  - e.g. ABySS, Velvet, ALLPATHS, SOAPdenovo, etc.
De Novo Assembly Algorithms

De Bruijn graph approach

- Construction of graph
  - All possible substrings of length $k$ are extracted from the reads - referred to as $k$-mers
De Novo Assembly Algorithms

De Bruijn graph

- Construction of graph
  - All possible substrings of length $k$ are extracted from the reads - referred to as $k$-mers
  - A node is created for every $(k-1)$-mer
De Novo Assembly Algorithms

De Bruijn graph

• **Construction of graph**
  • All possible substrings of length $k$ are extracted from the reads - referred to as $k$-mers
  • A node is created for every $(k-1)$-mer
  • An edge is added between two nodes if they share a $(k-2)$-mer
De Novo Assembly Algorithms

De Bruijn graph

- Given perfect data, there would exist a path in the graph that traverses each edge exactly once, corresponding to the whole-genome sequence.

- Problems with real data
  - Repeats in the original sequence
    - Use paired-end information
  - Errors in reads
    - Use read error correction
Assemble Genome: Tools

- SPAdes
- IDBA
- Velvet Optimiser
- ABySS
- MaSurCa
SPAdes

- Comes with different modules:
  - BayesHammer
    - Read Error Correction tool for Illumina reads
  - SPAdes
    - Iterative short-read genome assembly module
    - Values of K are selected automatically based on the read length and data set type
  - MismatchCorrector
    - Improves mismatch and short indel rates in resulting contigs and scaffolds.
ABySS

- Assembly by Short Sequences

- Uses distributed representation of de Bruijn graph
  - Allows parallel computation
  - Addresses limitation of assembling very large genomes (e.g. Human genome sequences)

- Two-Stage Assembly Process
  - Generation of all possible substring length k (k-mers)
    - Then processed to remove errors and generate contigs
  - Mate pair information is used to extend contigs
    - Resolves ambiguities in contig overlap
IDBA

- Iterative De Bruijn Graph De Novo Assembler

- Uses range of k-values to build iterative de Bruijn Graph
  - Keeps all the information of the graphs with different k values
    - Claims will perform better than other de Bruijn graph based assemblers
Velvet Optimiser

- An algorithm optimized for high coverage, very short (25-50 bp) read data
- Using de Bruijn graphs, aims to eliminate errors and resolve repeats
- Removes errors based on 3 topological features:
  - “Tips” due to errors at the edges of reads
  - “Bulges” due to internal read errors
  - Cloning errors resulting in false connections
- Implements “Breadcrumbs” module to resolve repeated regions that create tangles in graph
Reference-based Assembly

- SMALT
  - SMALT employs a hash index of short words up to 20 nucleotides long and sampled at equidistant steps along the reference genome. For each sequencing read, potentially matching segments in the reference genome are identified from seed matches in the index and subsequently aligned with the read using dynamic programming.
  - The best gapped alignments of each read are reported including a score for the reliability of the best mapping. The user can adjust the trade-off between sensitivity and speed by tuning the length and spacing of the hashed words.
Polishing and Gap-filling

- Fixing miscalls
- Fixing indels
- Extending contig ends
Polishing and Gap-filling

● PILON
  ○ Attempts to correct the following:
    ■ Single base differences
    ■ Small indels
    ■ Larger indel or block substitution events
    ■ Gap filling
    ■ Identification of local mis-assemblies, including optional opening of new gaps

● GMcloser
  ○ Fills and closes gaps in scaffold assemblies
Assembly Quality Assessment

- QUAST
  - Evaluates a range of metrics to evaluate assembly
  - The interface and visualizations are easy to use, representative and informative (according to their publication)

- KAT
Strain Identification

- **stringMLST**
  - **Fast** k-mer based tool for multi locus sequence typing (MLST) stringMLST is a tool for detecting the MLST of an isolate directly from the genome sequencing reads.
  - stringMLST predicts the ST of an isolate in a completely assembly and alignment free manner.
  - The tool is designed in a light-weight, platform-independent fashion with minimum dependencies.


References