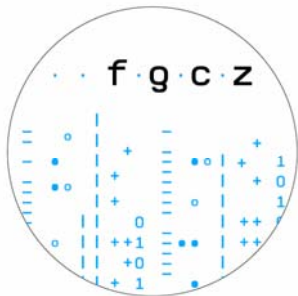


Genome Sequence Assembly

Weihong Qi, PhD
Functional Genomics Center Zurich
Uni./ETH Zurich

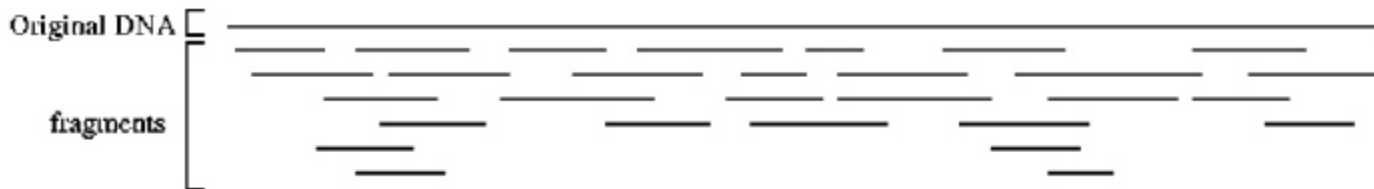


Outline of the program

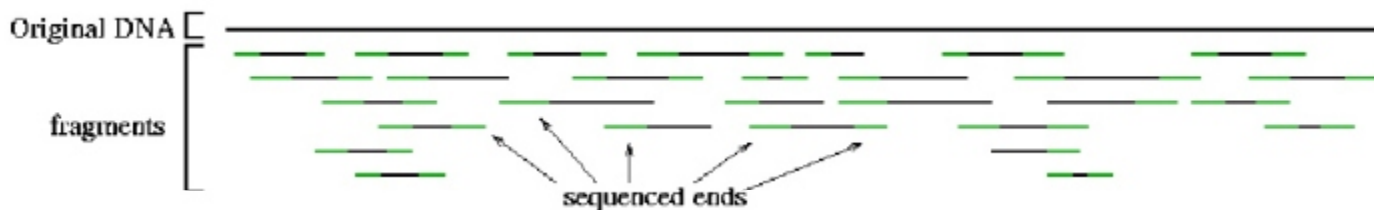
- Tuesday afternoon
 - Genome assembly theory
 - **Exercise 1:** Assemble 454 reads with 454 Assembler and Mapper
 - **Exercise 2 (optional):** Estimate sequencing coverage needed to sequence a genome based on the Lander and Waterman theory
 - **Homework:** Read about Mosaik assembler and Gigabayes
- Wednesday morning
 - **Exercise 3:** Analyze Solexa reads with Mosaik and Gigabayes
- Wednesday afternoon
 - **Summarize results for group presentation**

Whole genome shotgun sequencing (WGS) and assembly

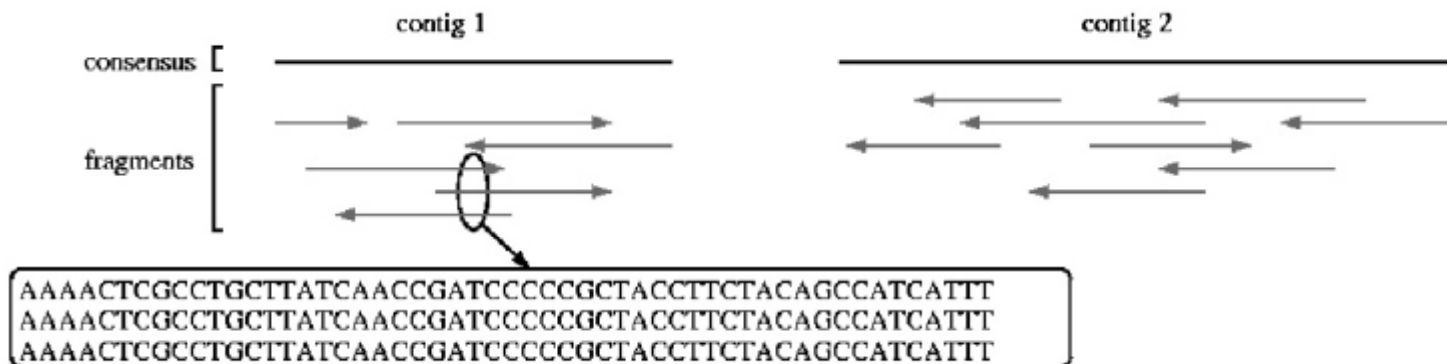
- Genomic DNA is broken into a collection of fragments



- The ends of each fragment are sequenced



- The sequence reads are assembled together based on sequence similarity



http://www.cbcb.umd.edu/research/assembly_primer.shtml

Coverage needed to sequence a genome

- We sequence enough random fragments so that we have an expected number of fragments containing each nucleotide. This expected number is the coverage **c**.
- **c** = **RL/G**, **R**: the number of reads sequenced, **L**: the average length of a read, **G**: the size of the genome
- Lander and Waterman theory
 - Sampling is perfectly random and uniform
 - Probability of a based is not sequenced = **e^{-c}**
 - Proportion of genome covered = **1-e^{-c}**
 - Total gap length = **Ge^{-c}**
 - Total number of gaps = **Re^{-c}**
 - Gaps of average length = **Ge^{-c} / Re^{-c} = L/c**
 - Contigs of average length = **G/Re^{-c} = (L/c)e^c**

Lander E and Waterman MS. 1998. Genomics 2:231-9

Coverage needed to sequence a genome

- We sequence enough random fragments so that we have an expected number of fragments containing each nucleotide. This expected number is the coverage **c**.
- **c** = **RL/G**, **R**: the number of reads sequenced, **L**: the average length of a read, **G**: the size of the genome
- Lander and Waterman theory

- Sampling is perfectly random and uniform
 - Probability of a based is not sequenced = e^{-c}
 - Proportion of genome covered = $1 - e^{-c}$
 - Total gap length = Ge^{-c}
 - Total number of gaps = Re^{-c}
 - Gaps of average length = $Ge^{-c} / Re^{-c} = L/c$
 - Contigs of average length = $G/Re^{-c} = (L/c)e^c$

Genome covered depends only on **c**, not **G** or **L**, 8 fold for 99% coverage

Gap length and number fall exponentially with **c**

Contig lengths rise exponentially with **c**

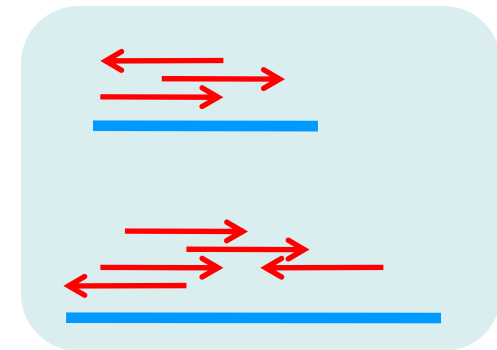
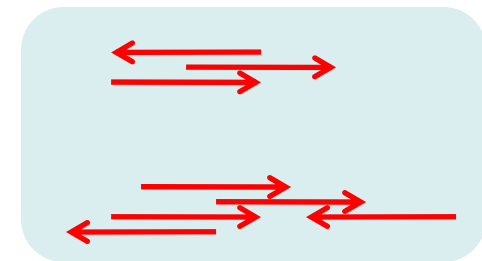
Lander E and Waterman MS. 1998. Genomics 2:231-9

Assembly algorithms

- **Overlap-layout-consensus (Align-layout-consensus)**
- Greedy algorithm
- Eulerian path

Overlap-layout-consensus

- Assemblers: 454 newbler, ARACHNE, PHRAP, CAP, TIGR, CELERA
- Overlap: find all overlapping reads
- Layout: merge optimal overlapping reads into contigs
- Consensus: derive the DNA sequence and correct read errors



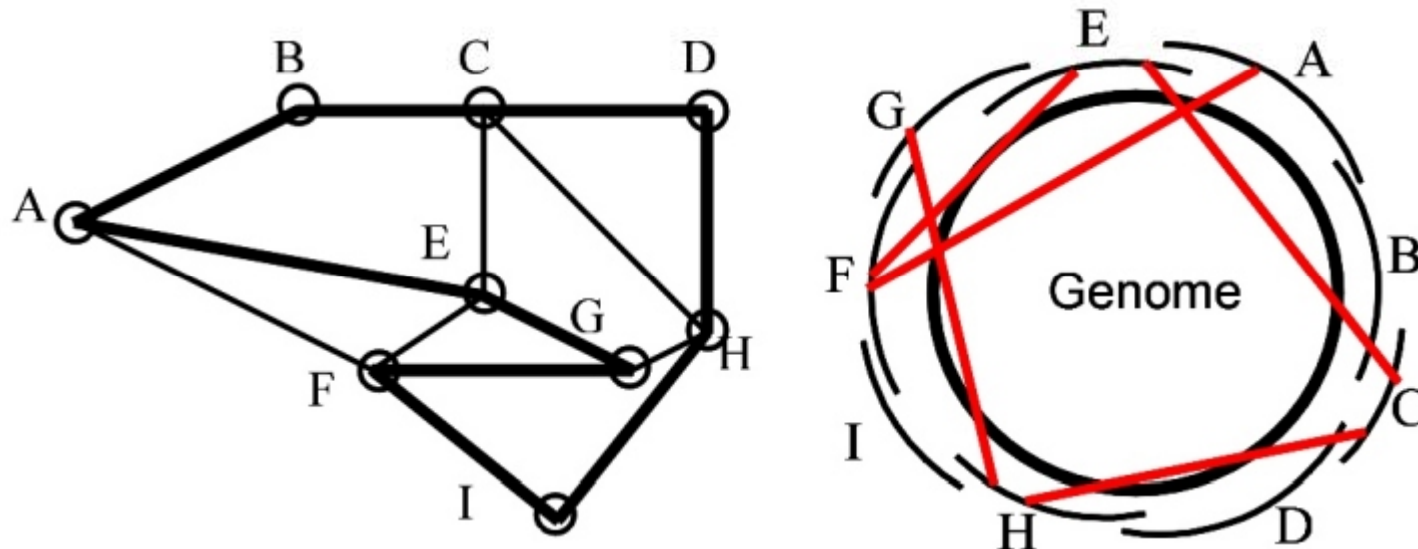
```
..ACGATTACAATAGGTT..  
..TAGATTACACAGATTACTGACTTGATGGCGTAA CTA..
```

Overlap

- Sort all k-mers in reads
- Find pairs of reads sharing a k-mer
- Extend to full alignment
- Throw away if not above a given threshold (X% similarity Y bp of length)
- Deal with repeats
 - Find areas covered by a significantly large number of reads
 - Discard all k-mers that appear more than $t \times \text{Coverage}$, ($t \sim 10$)

Layout

- Find optimal overlapping reads
 - Simple path: a path through each node just once



http://www.cbcb.umd.edu/research/assembly_primer.shtml

Crete local multiple alignment from overlapping reads

- Progressive alignment
 - Align the most similar pair
 - Progress to the most distantly related.
 - Performance is good when the sequences are closely related
 - Efficient for many (100s to 1000s) sequences

```
      TAGATTACACAGATTACTGA
    TAGATTACACAGATTACTGA
  TAG - TTACACAGATTATTGA
TAGATTACACAGATTACTGA
TAGATTACACAGATTACTGA
```

Consensus

- Nucleotide space

- Each consensus is derived by weighted voting

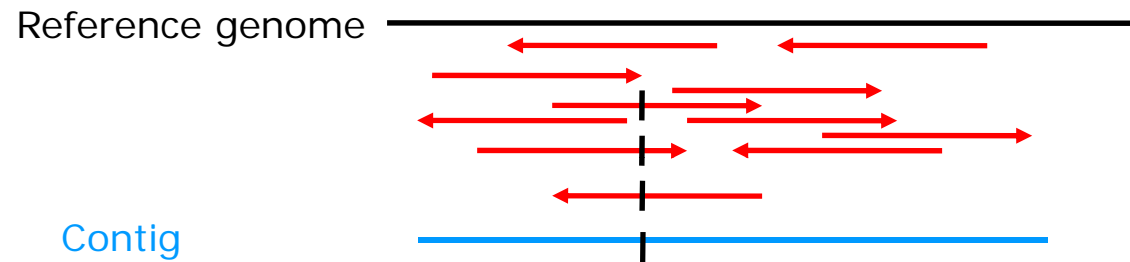
TAGATTACACAGATTACTGA	C	20	C	20
TAGATTACACAGATTACTGA	C	35	C	35
TAG - TTACACAGATTATTGA	T	30	C	0
TAGATTACACAGATTACTGA	C	35	C	35
TAGATTACACAGATTACTGA	C	40	C	40
	A	15	A	15
	A	25	A	25
	-		A	0
	A	40	A	40
	A	25	A	25

- Flowspace

- Averaging flow signals for each nucleotide flow included in the alignment
- Improve accuracy for the basecalls

Align-layout-consensus

- Align and layout reads to reference(s)



Exercise 1

- Assemble 454 reads with 454 newbler
- Assemble the same 454 reads by mapping to a reference with 454 mapper
- View assembly with EagleView

The screenshot shows the 454 Newbler software interface with the 'Parameters' tab selected. The interface is divided into several sections:

- Project:**
 - ☒ Incremental de novo assembler analysis
 - ☐ Large or complex genome
 - Expected depth: 0
- Overlap Detection:**
 - Seed step: 12
 - Seed length: 16
 - Seed count: 1
 - Minimum overlap length: 40
 - Minimum overlap identity: 90 %
 - Alignment identity score: 2
 - Alignment difference score: -3
- Configuration Files:**
 - Trimming database: [empty field]
 - Screening database: [empty field]
- Output:**
 - ☒ Include consensus
 - Pairwise alignment:
 - ☒ None
 - ☐ Simple format
 - ☐ Tabbed format
 - Ace/Consed:
 - ☐ No files
 - ☒ Single ACE file for small genomes
 - ☐ Single ACE file
 - ☐ ACE file per contig
 - ☐ Complete consed folder
 - Ace read mode:
 - ☒ Default
 - ☐ Raw
 - ☐ Trimmed
 - All contig threshold: 100
 - Large contig threshold: 500

On the right side of the interface, there is a vertical toolbar with buttons: Exit, New, Open, Start, and Stop.

Exercise 2

- Given a 5 Mb bacterial genome and an average read length of 600 bases, what will be the number of reads needed to reach fold coverage between 1 and 10?
- What will be the percentage of genome covered, average contig length, number of gaps, and average gap length?
- What will be these values if the average read length is 250 bases and 35 bases?