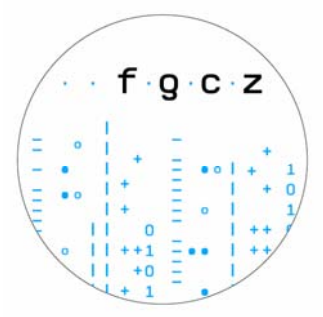
Genome Sequence Assembly

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

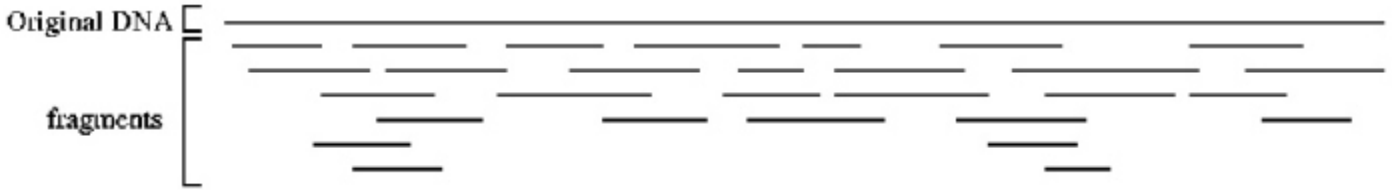


Outline of the program

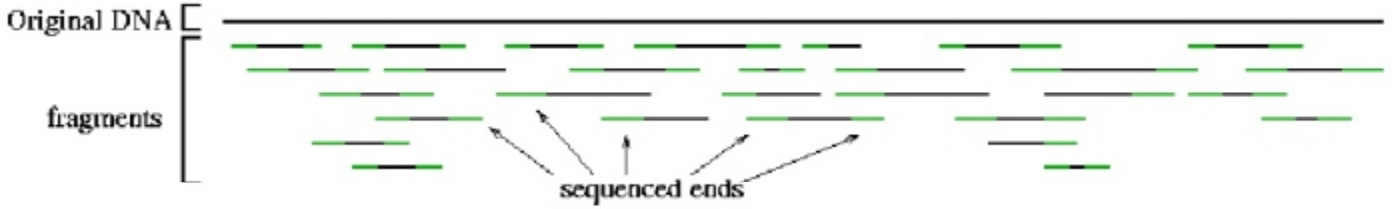
- Tuesday afternoon
  - Genome assembly theory
  - Exercise 1: Assemble 454 reads with 454 Assembler and Mapper
  - Exercise 2 (optional): Estimate sequencing coverage needed to sequence a genome based on the Lander and Waterman theory
  - Homework: Read about Mosaik assembler and Gigabayes
- Wednesday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

Whole genome shotgun sequencing (WGS) and assembly

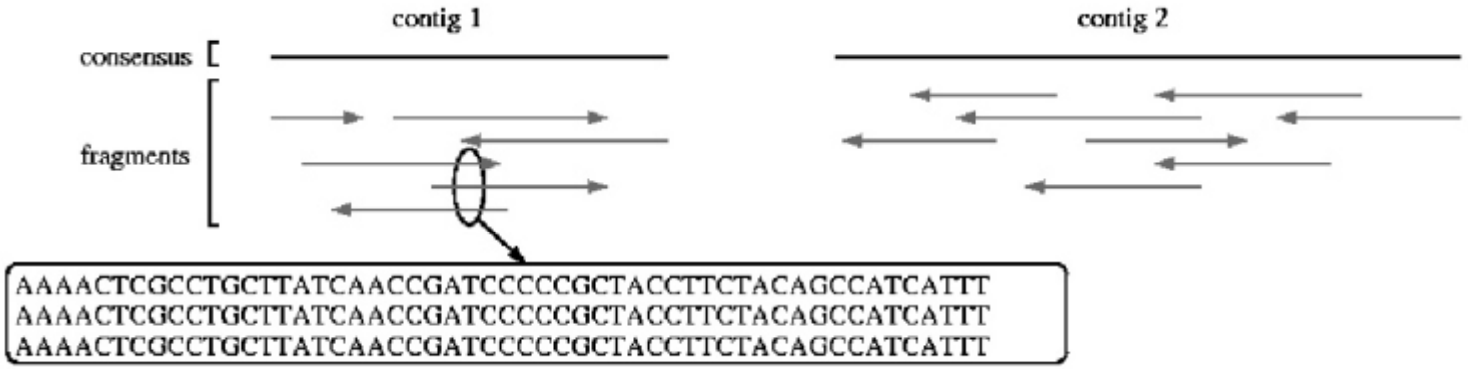
- Genomic DNA is broken into a collection of fragments



- 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}
    - Toatl 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}
    - Toatl 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**

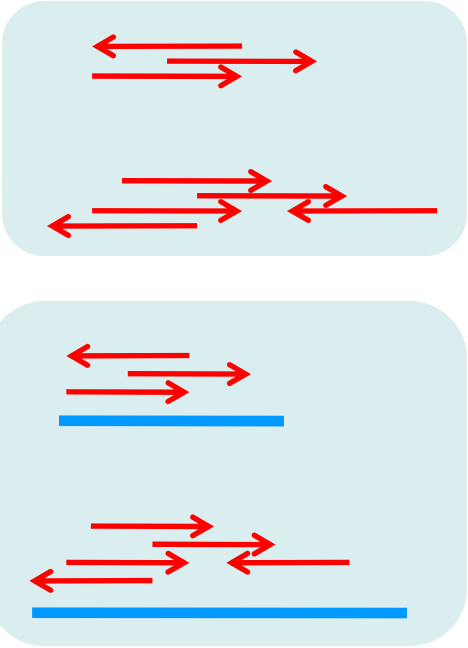
Lander E and Waterman MS. 1998. Genomics 2:231-9

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



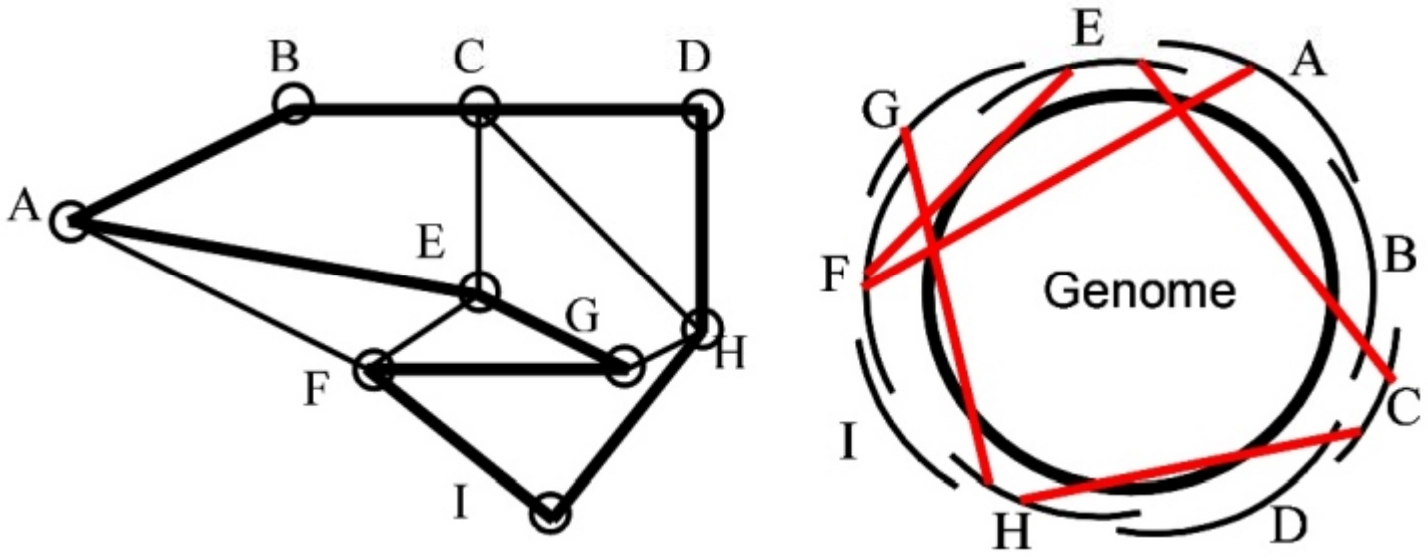
..ACGATTACAATAGGTT..
..TAGATTACACAGATTACTGACTTGATGGCGTAA CTA..

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 × Coverage, (t ~ 10)

Layout

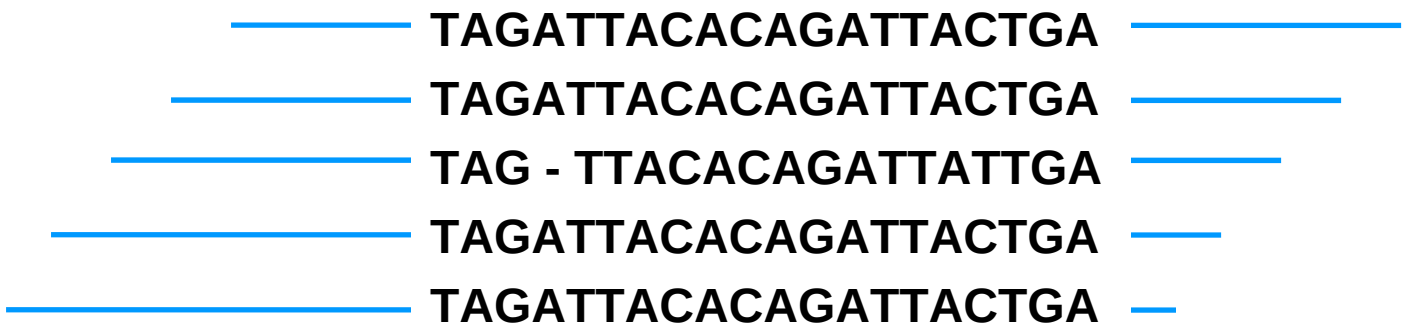
- Find optimal overlapping reads
  - Simple path: a path through each node just once



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

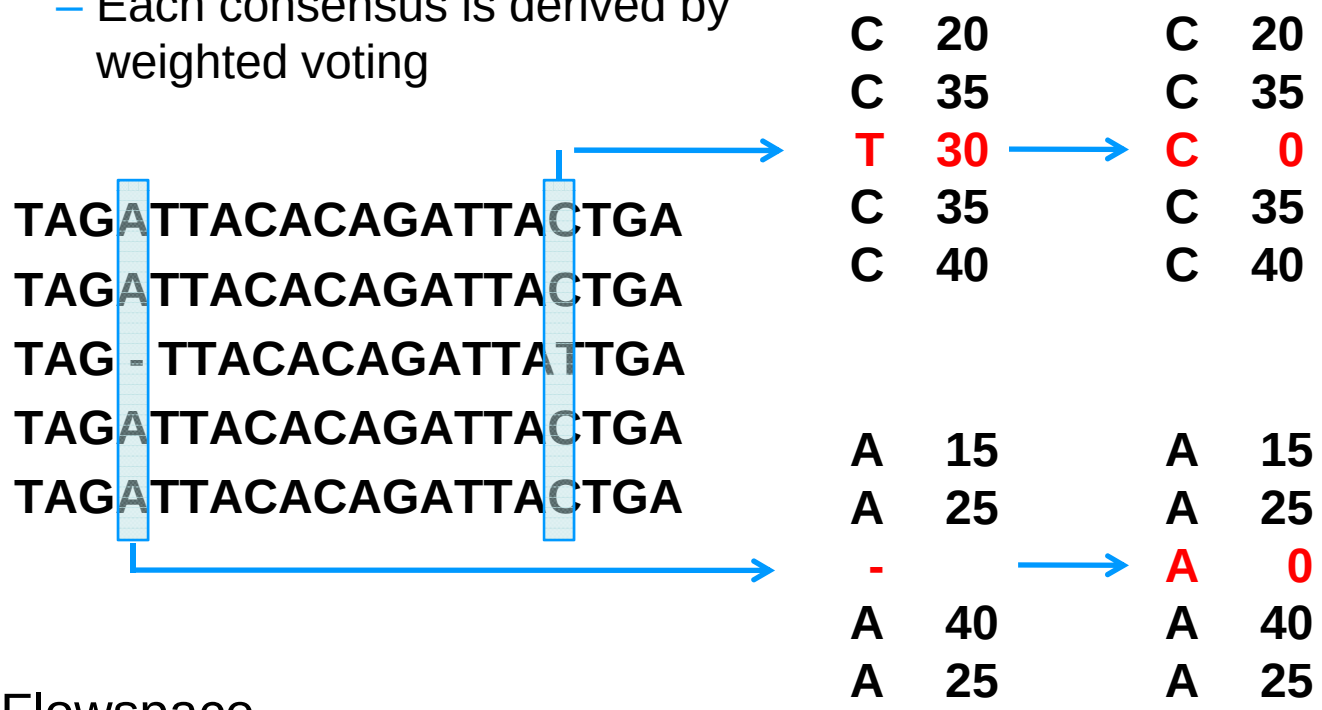
Creat local multiple alignment from overlapping reads

- 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



Consensus

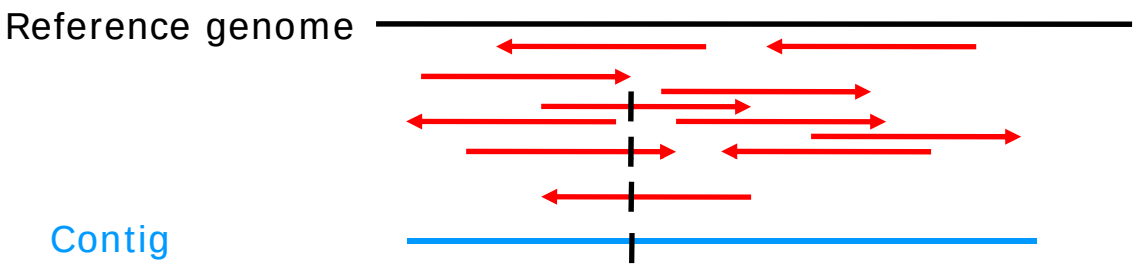
- Nucleotide space
  - Each consensus is derived by weighted voting



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

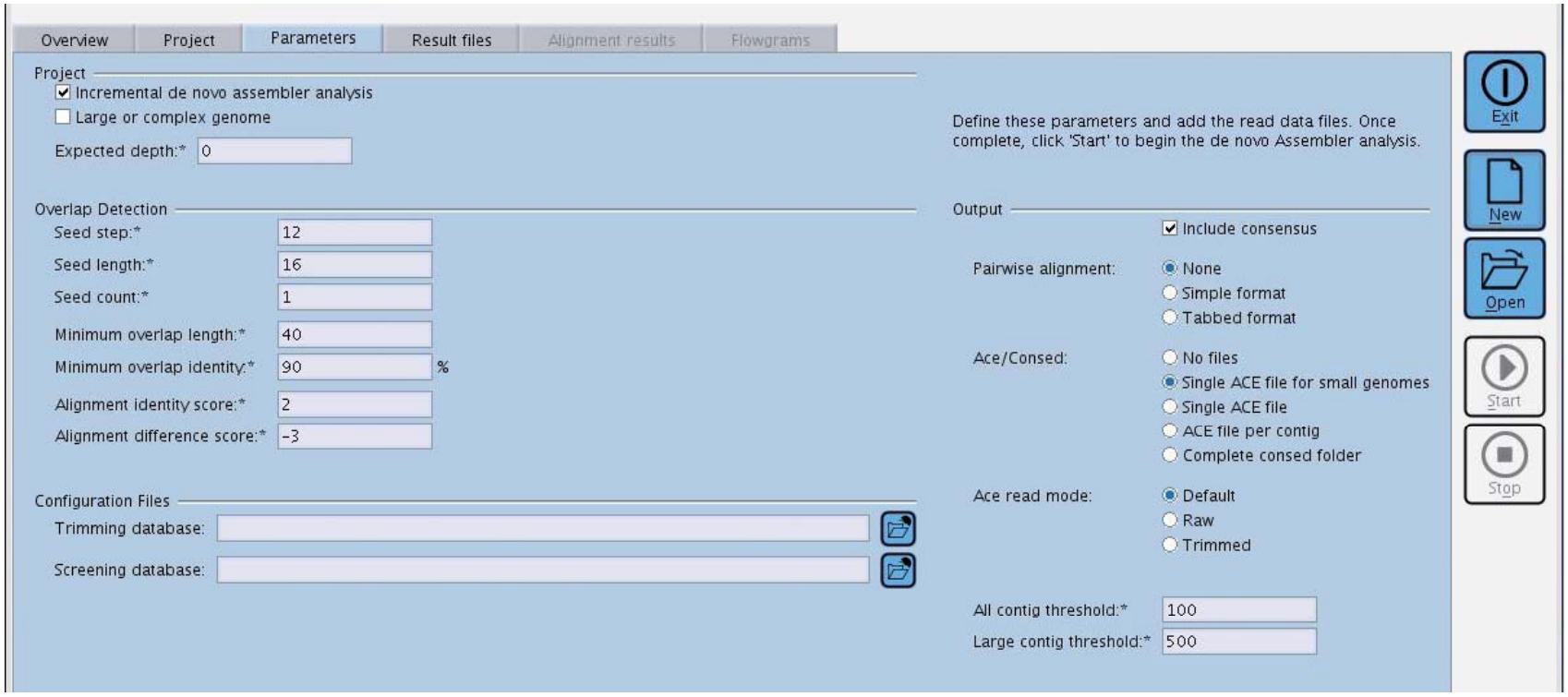
Align-layout-consensus

- Align and layout reads to reference(s)



Exercise 1

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



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?