Genome Assembly

“The art of possible”
- Can’t just read the DNA base by base.
- First technique: “Sanger ladder”, 1977
  - Cells copy DNA base by base.
  - Can modify this process by “starving” each of (A,G,T,C).
  - Replication would terminate (with some probability) when encountering “starved” base.
  - Separate sequences by length.
  - Measure the lengths.
  - Repeat for each of (A,G,T,C).
- Other techniques:
  - Sequencing by hybridization.
  - Sequencing by synthesis.
  - ...

Shotgun approach
- Problem: this process has limited duration.
  - Can read at most a few hundred bases (up to 1000).
  - Insert length estimation error grows with insert length.
  - Human genome consists of ~3 billion bases.
- Solution:
  - Cut the sequence into short fragments (“inserts”).
  - Sequence each fragment.
- How to put it together??

Shotgun approach, ctd.
- Actual solution:
  - Take several copies of the sequence.
  - Cut them into inserts.
  - Sequence each insert.
  - Let:
    - \( L \) = read length
    - \( G \) = genome length
    - \( N \) = number of reads sequenced
  - Coverage = \(LN/G\).
  - Use the coverage to recover the whole sequence.

Assembly
- Consider inserts:
  - TAATCTA
  - ACTAA
  - AGT
  - TCTA
  - AGTAC
- Can you recover the original sequence?
Assembly steps

- **Overlap:**
  - Identify potentially overlapping reads

- **Layout:**
  - Find the order of reads along the sequence

- **Consensus**

- **Issues:**
  - Sufficient coverage and read length
    - Lander-Waterman formula
  - Measurement errors:
    - A few percent of the bases will be incorrect
    - Non-exact overlap
  - Repeats!!
    - Can lead to multiple layouts
    - >50% of human genome consists of repeats

Dealing with repeats

- **Hierarchical shotgun sequencing:**
  - Partition the sequence into clones of ~100 kb
  - The order of clones is known
    - Requires additional information
    - Cumbersome to obtain
  - Sequence each clone separately
  - Combine

- **Approach used by the Human Genome Project**

Dealing with repeats II

- **Whole genome shotgun**

- **Inserts vs. reads**

- **We can have long inserts, read only partially from each end:** “mate pairs”
  - Known distance between the reads
  - Provide additional information

- **Reconstruction results in**
  - Contigs
  - Scaffolds

- **Approach used by Celera Genomics**

References

- **Human genome:**
  (for both papers, Google-Scholar “human genome”)