Sequence Assembly

Sequence assembly in practice

Outline

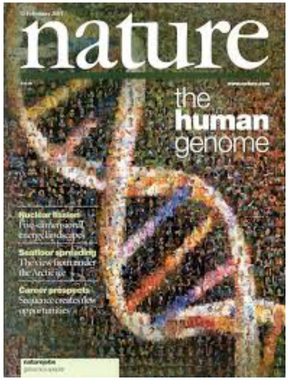
- Whole genome sequencing strategies
- Real-world assembly methodology and output
- Challenges of sequence assembly in practice
- The common Overlap-Layout-Consensus approach for fragment assembly
- Paired-end reads

Whole Genome Sequencing

- Two main strategies:

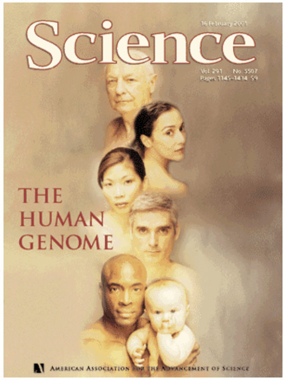
1. Clone-by-clone mapping

Lander et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409: 860–921.

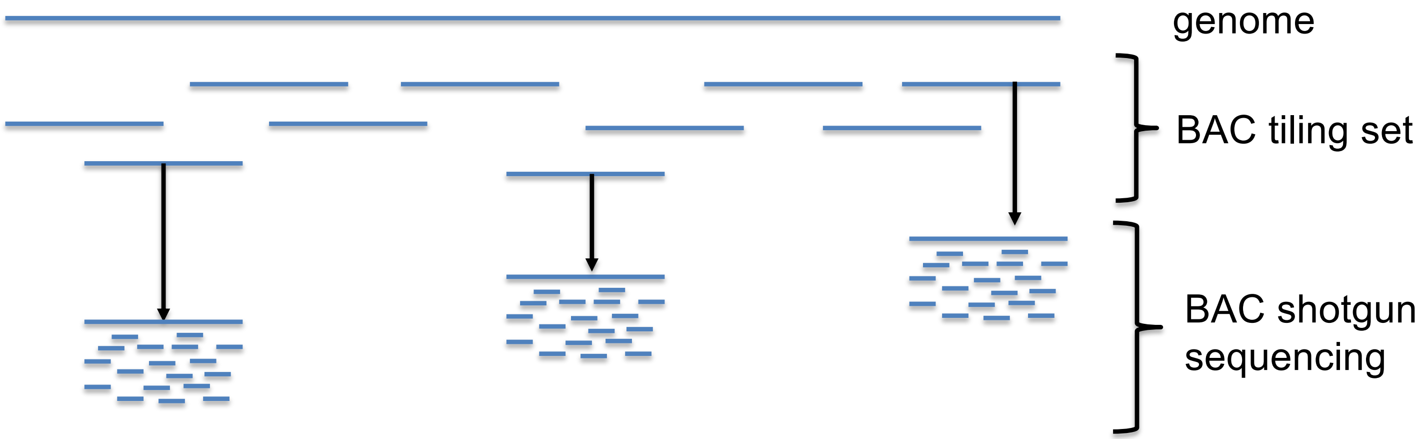


2. Whole-genome shotgun

Venter et al. The sequence of the human genome. *Science*. 2001;291: 1304–1351.

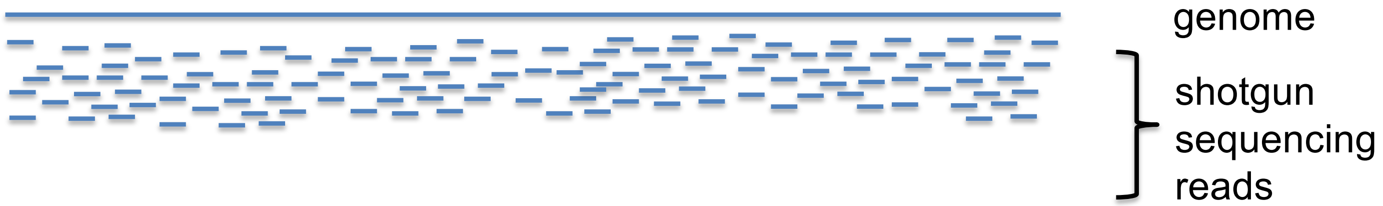


Clone-by-clone mapping



- Fragment genome into large pieces, insert into BACs (Bacterial Artificial Chromosomes)
- Choose *tiling set* of BACs: overlapping set that covers entire genome
- Shotgun sequence the BACs
- Merge assembled BACs into genome assembly

Whole-genome shotgun sequencing



- Shotgun sequence entire genome at once
- Larger computational problem to assemble all reads from entire genome at the same time

Assembly in practice

- Assembly methods used in practice are complex
  - But generally follow one of the two approaches
    - Reads as *vertices*
    - Reads as *edges (*or *paths* of edges)
- Assemblies do not typically give whole chromosomes
  - Instead gives a set of “contigs”
  - *contig*: contiguous piece of sequence from overlapping reads
  - contigs can be ordered into *scaffolds* with extra information (e.g., paired end reads)

Challenges with the fragment assembly approach

- Read errors
  - Complicates computing read overlaps
- Repeats
  - Roughly half of the human genome is composed of repetitive elements
  - Repetitive elements can be long (1000s of bp)
  - Human genome
    - 1 million Alu repeats (~300 bp)
    - 200,000 *LINE* repeats (~1000 bp)

Challenges with the spectral (de Bruijn) approach

- Not all *k*-mers may be contained within the reads even if reads completely cover the genome
- Reads often have sequencing errors!
  - False k-mers
  - Missing k-mers
- DNA repeats result in *k*-mers that are present in multiple copies across the genome

Overlap-Layout-Consensus

- Most common assembler strategy for long reads

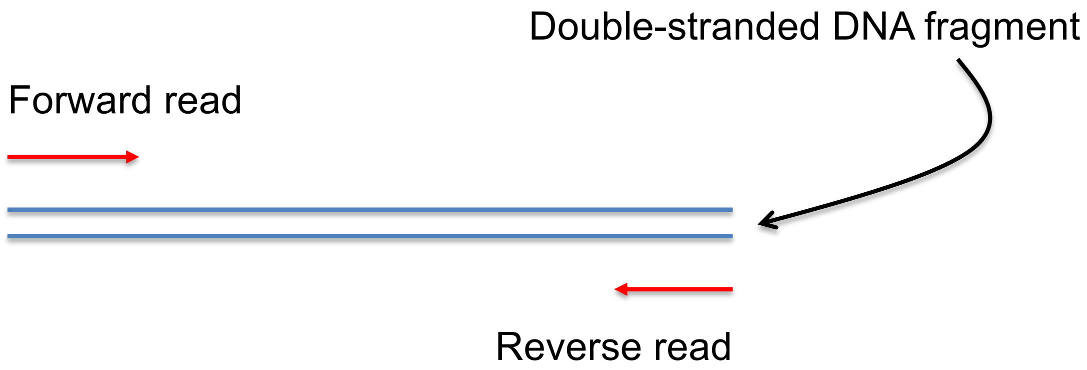
1. *Overlap*: Find all significant overlaps between reads, allowing for errors
2. *Layout*: Determine path through overlapping reads representing assembled sequence
3. *Consensus*: Correct for errors in reads using layout

Consensus

```
                                GTATCGTAGCTGACTGCGCTGC
Layout                       ATCGTCTCGTAGCTGACTGCGCTGC
                        ATCGTATCGAATCGTAG
            TGACTGCGCTGCATCGTATCGTATC

Consensus   TGACTGCGCTGCATCGTATCGTATCGTAGCTGACTGCGCTGC
```

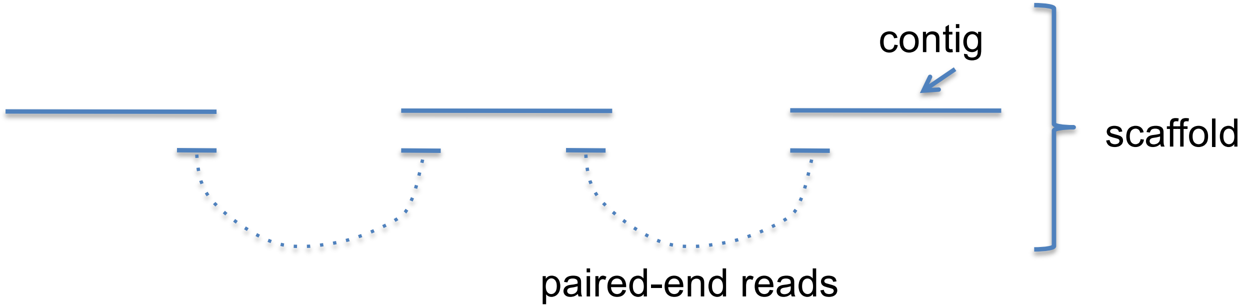
Paired-end reads



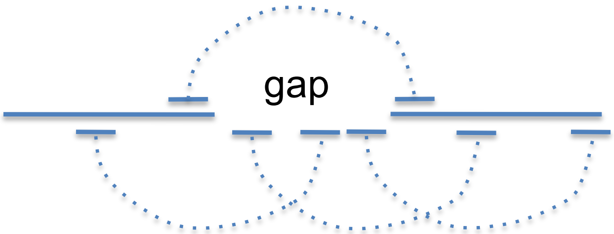
- Paired-end read data give sequence from both ends of a DNA fragment
- Read are “paired” in the sequencer output
- Which read is ‘forward’ and which is ‘reverse’ is unknown

Paired-end read advantages

- *Scaffolding*: layout of adjacent, but not overlapping, *contigs*



- *Gap filling*:



Summary

- Both approaches to sequence assembly face significant challenges
- Biggest challenge is repeats!
  - Large genomes have a lot of repetitive sequence
- Whole genome sequencing strategies
  - Clone-by-clone: break the problem into smaller pieces which have fewer repeats
  - Whole-genome shotgun: use paired-end reads to assemble around and inside repeats
- Consensus approaches are used to correct for sequencing errors