Conventional Sequencing

From “DNA Sequencing” - Wikipedia
Sequence assembly

Definitions:

- Contig – a contiguous block of sequence
- Scaffold – contigs joined with gaps of unknown sequence and length between them
- N50/NG50 – a measure of assembly length
- Total assembled bp
Assemblers

• Two types of assemblers:
  • De Bruijn graph assemblers
    ■ Newbler (454 only – may or may not be DBG)
    ■ Velvet (mixed assemblies)
    ■ ABySS
    ■ AllPaths
    ■ SOAPdenovo
  • Overlap graph assemblers
    ■ Mira
    ■ Celera assembler
Core assembly algorithms

A Read Layout

R_1: GACCTACA
R_2: ACCTACAA
R_3: CCTACAAG
R_4: CTACAAGT
A: TACAAGTT
B: ACAAGTTA
C: CAAGTTAG
X: TACAAGTC
Y: ACAAGTCC
Z: CAAGTCCG

B Overlap Graph

C de Bruijn Graph

From: Genome Res. 2010 September; 20(9): 1165–1173. Assembly of large genomes using second-generation sequencing
Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg
Assemblers are complicated, multi-stage pipelines, dealing with:

- Correcting measurement errors within the reads,
- Constructing contigs,
- Resolving repeats (i.e. disambiguating false positive alignments between reads) and
- Scaffolding contigs
- (optionally “gap filling”)

From: Assemblathon 1: A competitive assessment of de novo short read assembly methods, Genome Research, Sept. 2011
Assembly

The Celera Assembler works with fragmentary sequences, their detected overlaps, and their given mate pairs. Often, the data are mutually contradictory, as shown here. Yet, Celera Assembler reduces the data to a linear sequence whenever that is justified. (A) Sequence overlaps and mate pairs suggest several possible joins. Line segments represent fragments, vertical stacking represents overlaps, rectangles represent contigs, arrows represent links, and every element's thickness correlates to the amount of supporting data. (B) The assembler reduces the graph such that one contradiction remains. The sequence fragments were reduced to contigs based on overlaps. The mate pairs were reduced to contig links of various weights. Here, three contigs form a linear scaffold but the fourth contig is problematic. (C) The assembler has reduced the graph to a linear sequence. Its final step was to insert the 4th contig twice. Called a multiply placed surrogate unitig, the 4th contig appears to represent over-collapse of fragments induced by a near-perfect repeat in the genome.
Assemblers Comparison

The N50 length is defined as the length $N$ for which 50% of all bases in the sequences are in a sequence of length $L < N$. This can be found mathematically as follows: Take a list $L$ of positive integers. Create another list $L'$, which is identical to $L$, except that every element $n$ in $L$ has been replaced with $n$ copies of itself. Then the median of $L'$ is the N50 of $L$. For example: If $L = \{2, 2, 2, 3, 3, 4, 8, 8\}$, then $L'$ consists of six 2's, six 3's, four 4's, and sixteen 8's; the N50 of $L$ is the median of $L'$, which is 6.

Watch out! Some assemblies produce "contigs" which are actually scaffolds.