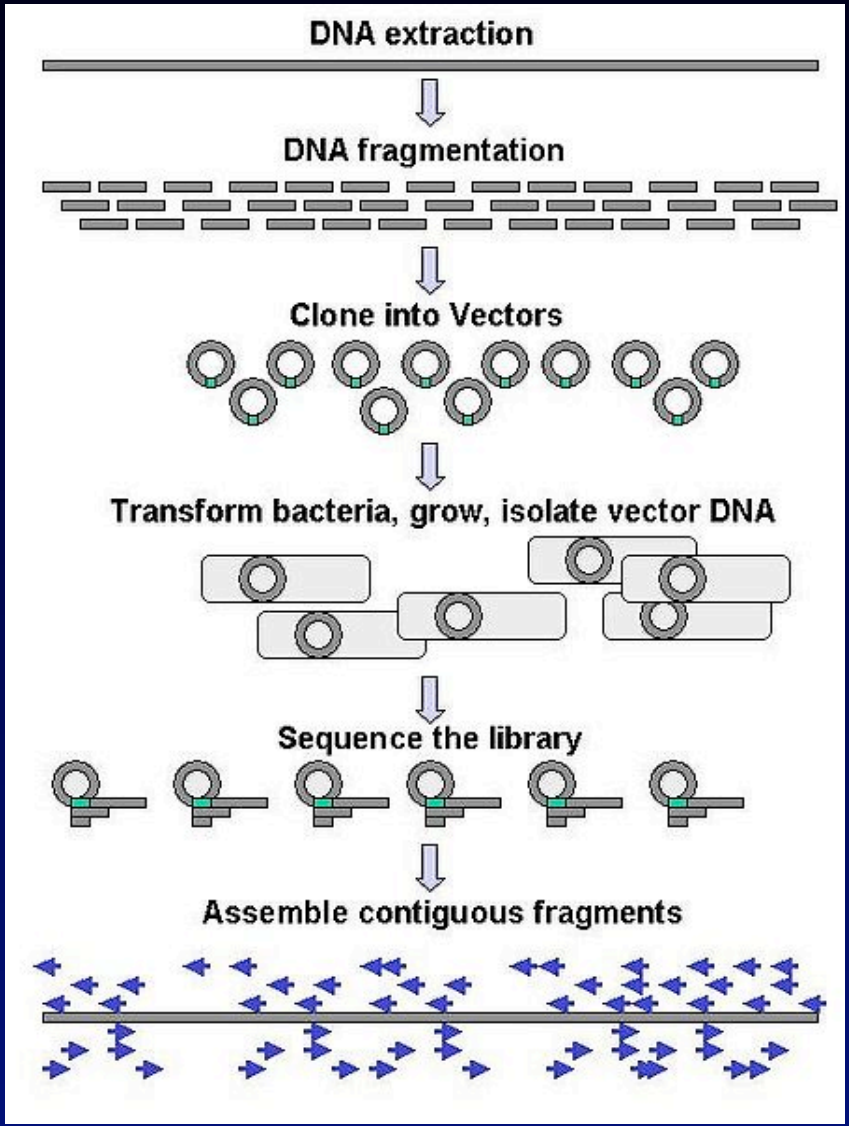
Conventional Sequencing



From “DNA Sequencing” - Wikipedia

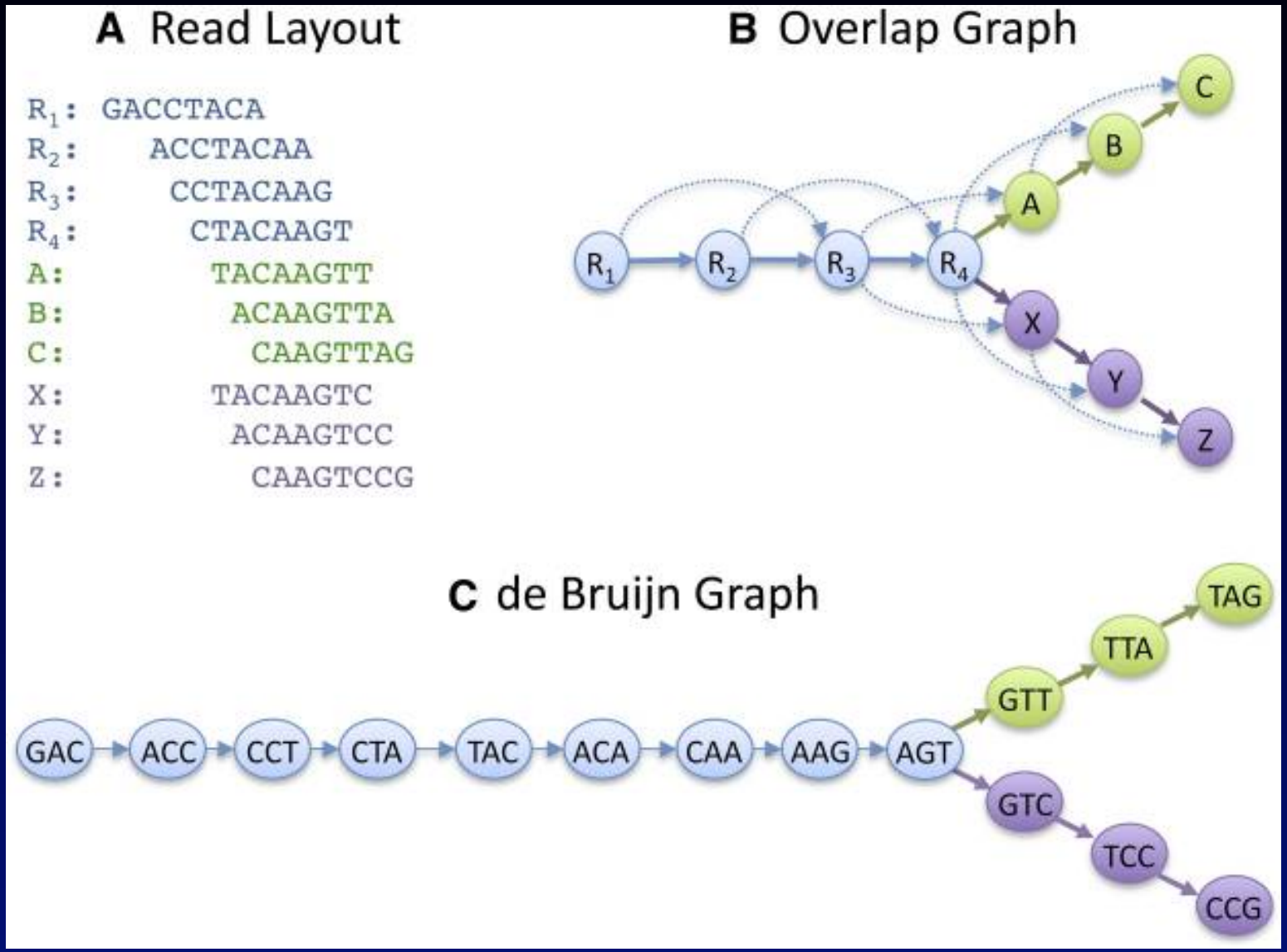
Sequence assembly

- 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 asembler

Core assembly algorithms

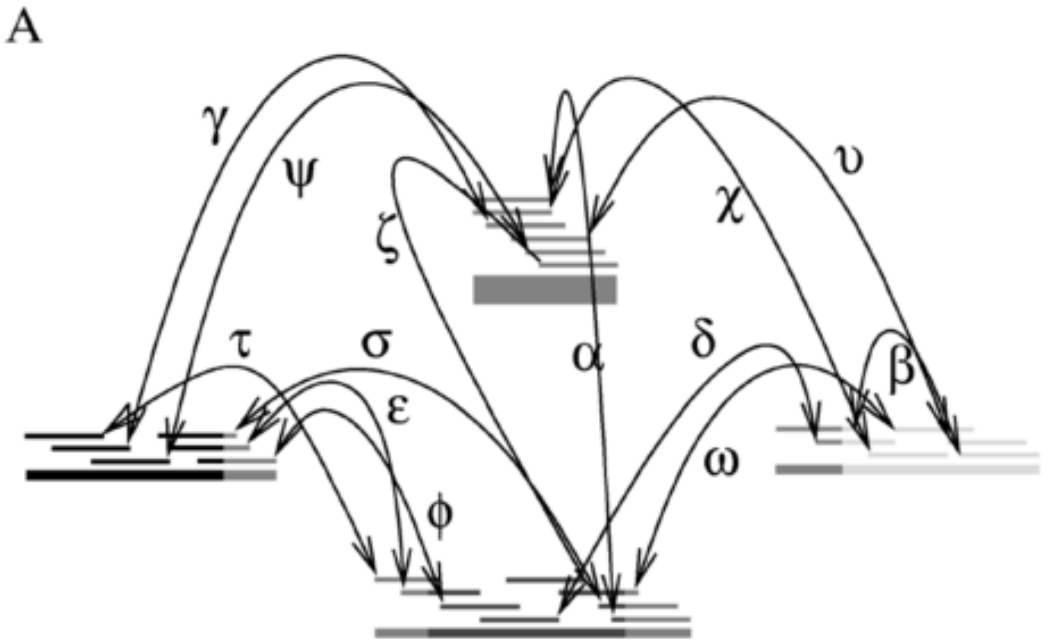


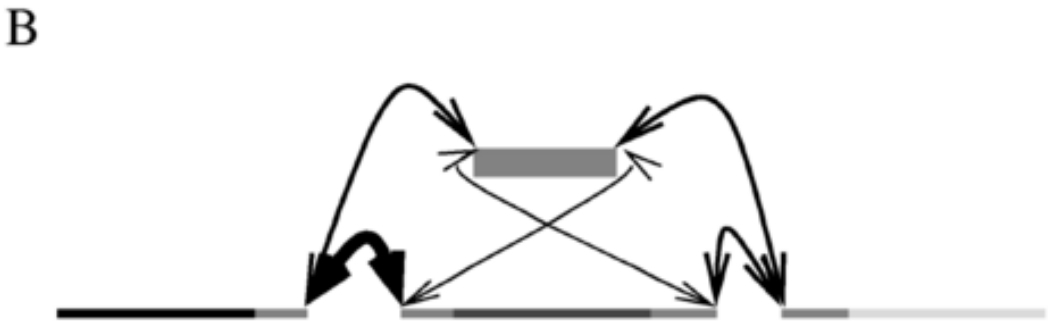
From: Genome Res. 2010 September; 20(9): 1165–1173. Assembly of large genomes using second-generation sequencing
Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg

Assemblers

- 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”)

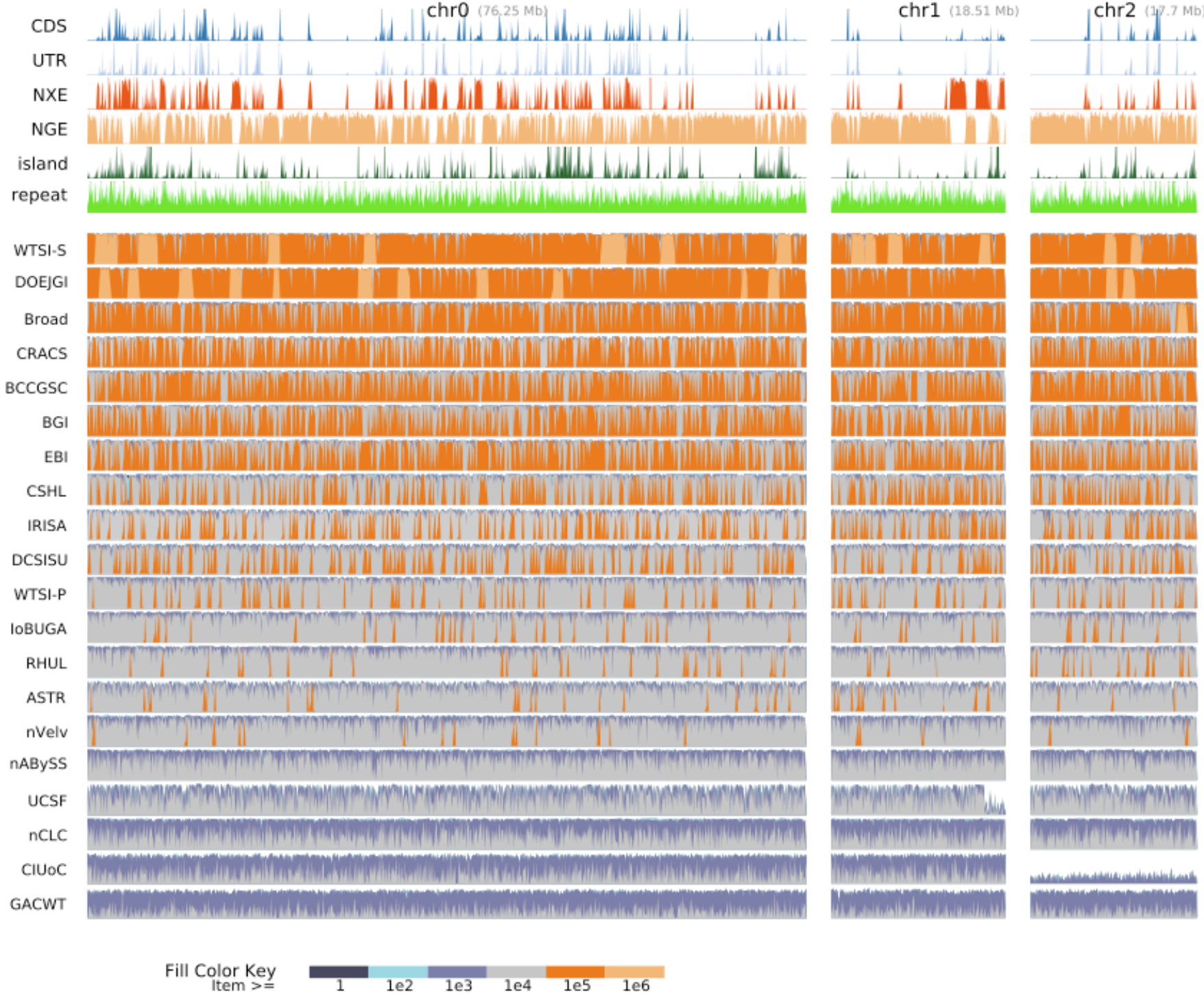
- From: Assemblathon 1: A competitive assessment of de novo short read assembly methods, Genome Research, Sept. 2011



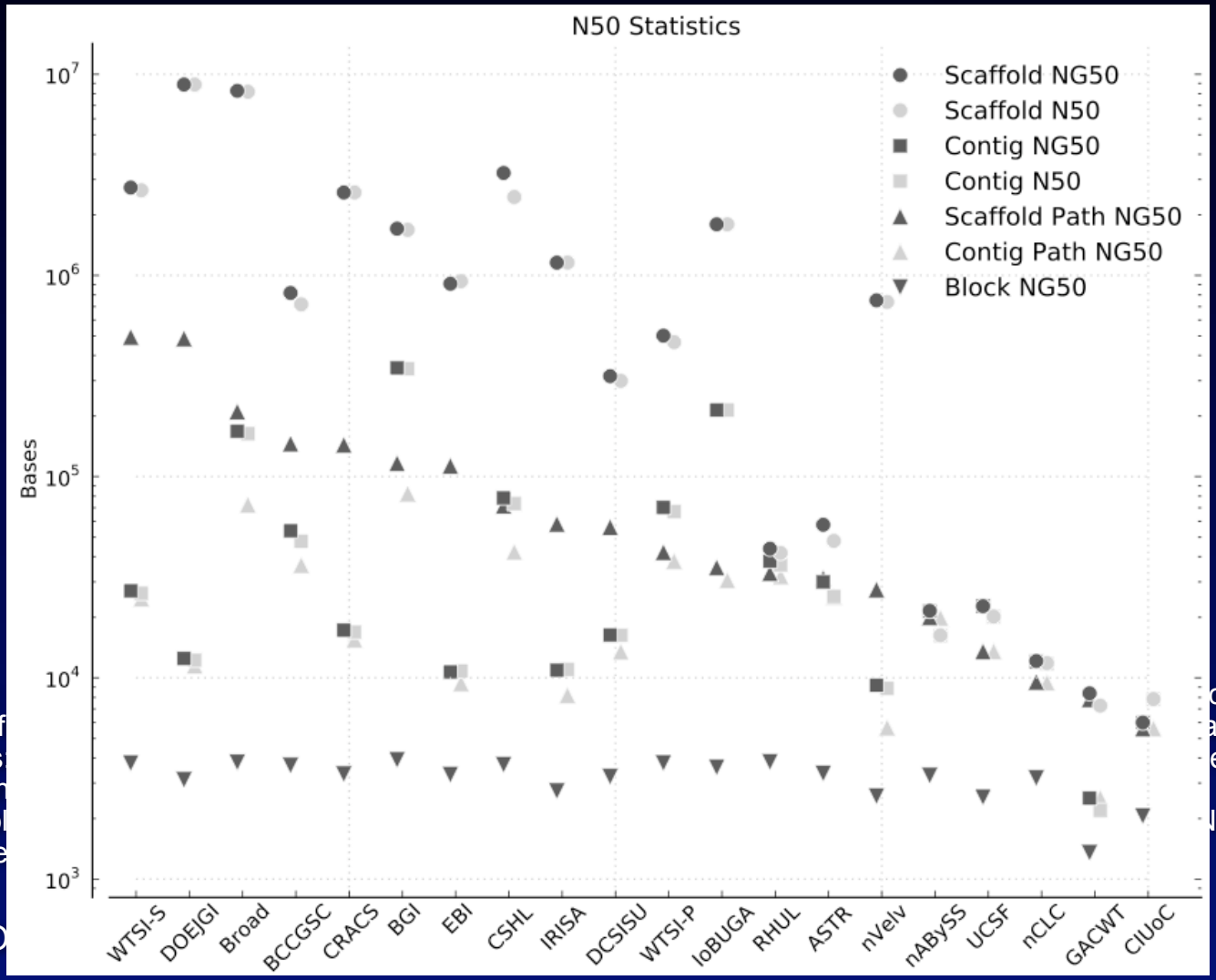


Assembly

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.


chr0 (76.25 Mb)
chr1 (18.51 Mb)
chr2 (17.7 Mb)
CDS
UTR
NXE
NGE
island
repeat
WTSI-S
DOEJGI
Broad
CRACS
BCCGSC
BGI
EBI
CSHL
IRISA
DCSISU
WTSI-P
IoBUGA
RHUL
ASTR
nVelv
nABySS
UCSF
nCLC
CIUoC
GACWT
Fill Color Key
Item >=
1
1e2
1e3
1e4
1e5
1e6

Assemblers Comparison



N50 length:
50% of
follows
elemen
exampl
L is the

WATCH O

ch
ally as
ery

N50 of