(More) Assembly QC

2016.12.14

Joe Fass
jnfass@ucdavis.edu

Screen

Outside of screen:

```
screen -S name     # start a named screen

screen -list     # lists running screens and their status

screen -r name    # should autocomplete (tab!) … reattaches named screen

screen -d name     # detaches named screen (in case it’s attached elsewhere)
```

In a screen:

```
<ctrl-a> “     # double quote typed as <shift-’> … lists panes

<ctrl-a> A     # (A=<shift-a>) … edit the name of the current pane

<ctrl-a> c     # creates a new pane … ‘exit’ or ‘<ctrl-a> k’ kills a pane

<ctrl-a> 0-9     # switches to pane 0-9

<ctrl-a> d     # detaches screen
```

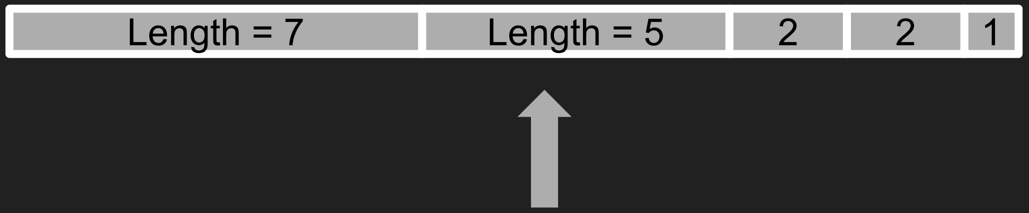
N50

“Length-weighted median length”

E.g. for lengths [7, 5, 2, 2, 1] … find median of:

[(7, 7, 7, 7, 7, 7, 7), (5, **5**, 5, 5, 5), (2, 2), (2, 2), (1)]

Equivalently, the length of the segment overlapping the midpoint of summed lengths:



N50

Allows one to claim that:

Half of the total (assembled) sequence is in pieces of at least [N50] in length.

N50’s *fatal flaw!*

The main problem with N50 comes from comparison. “Which assembly has better N50?”

Since N50 is defined with respect to a *total size* (of the sequence set), both the median length *and* total size will change N50.

E.g. remove shortest sequences:

[(7, 7, 7, 7, 7, 7, 7), (5, **5**, 5, 5, 5), (2, 2), (2, 2), (1)]

[(7, 7, 7, 7, 7, **7, 7**), (5, 5, 5, 5, 5)]

… different assemblers / runs may have different lower length cutoffs. See Keith Bradnam’s “N50 Booster” script … note the April 1st blog post!

NG50

Define a standard length for all comparisons, e.g. “50% of the expected genome size.”

E.g. for an expected “genome” size of 112, the median or 50th percentile (56) lies at the same spot in both these “assemblies:”

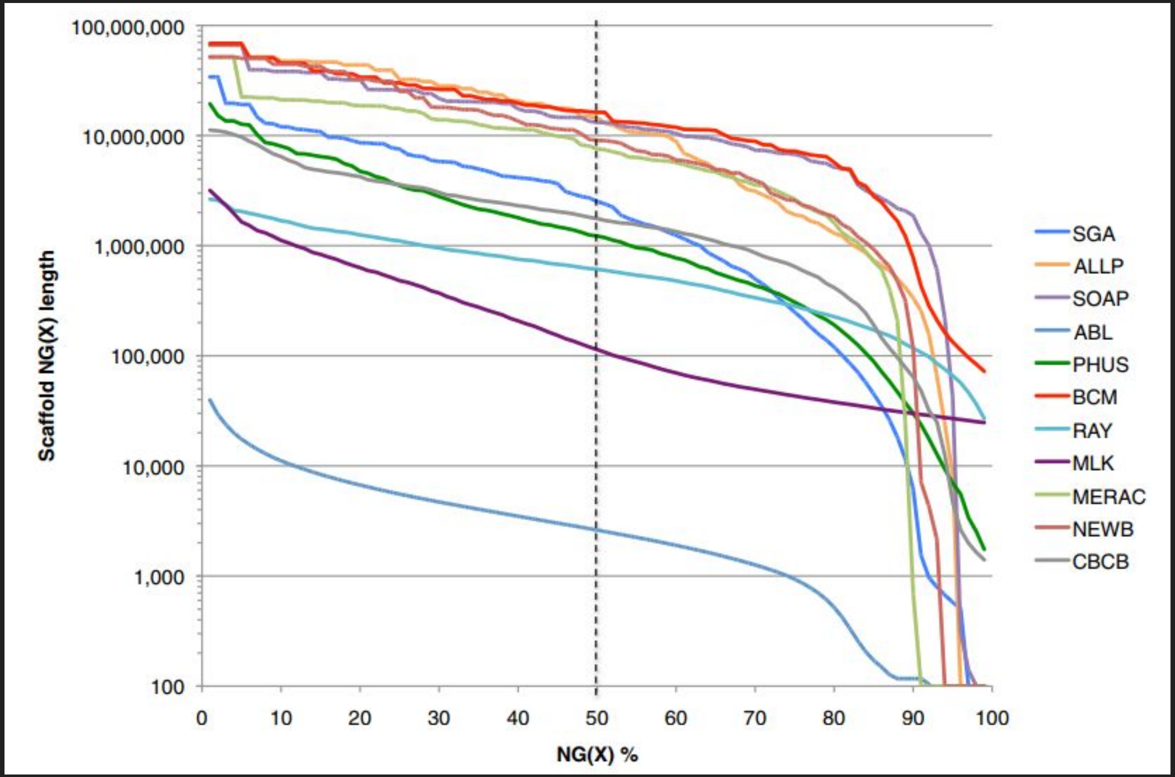
[(7, 7, 7, 7, 7, 7, 7), (5, **5**, 5, 5, 5), (2, 2), (2, 2), (1)]

[(7, 7, 7, 7, 7, 7, 7), (5, **5**, 5, 5, 5)]

This makes more sense, as the difference between these two assemblies has nothing to do with the median- or larger sized contigs.

NGx Plot

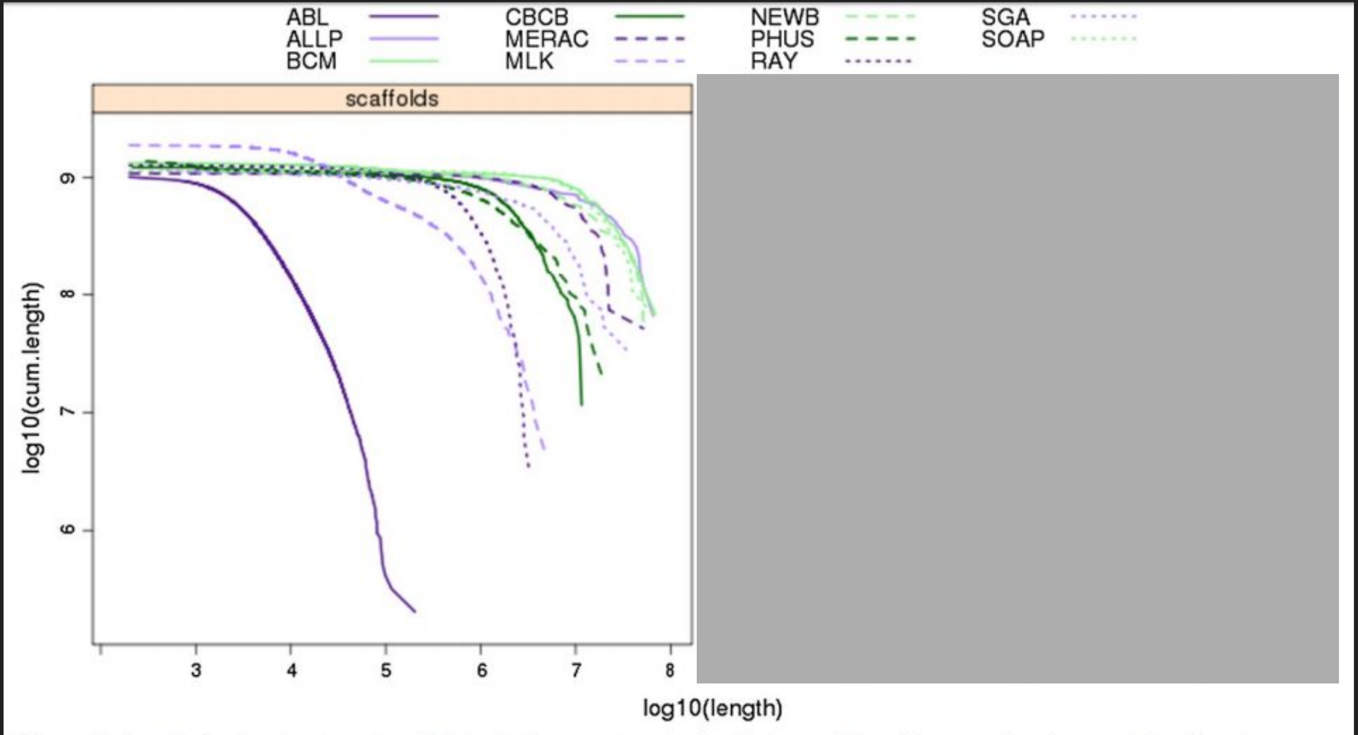
If NG50 is interesting, why not NG25? NG75? NG11? …



Bradnam (2013) *GigaScience* 2:10

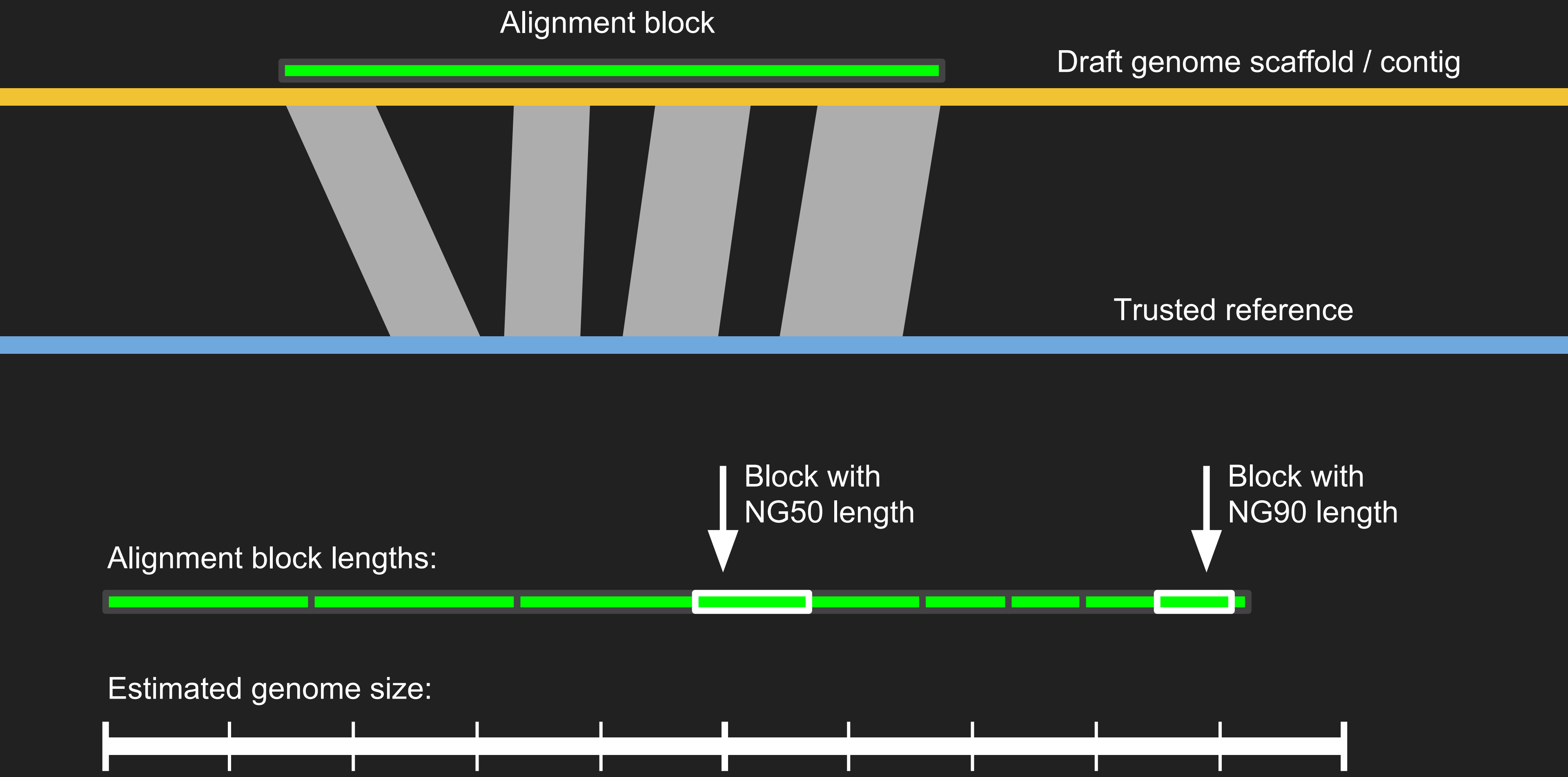
Cumulative Length Plots

… but, measures calculated at predetermined discrete intervals are “lossy.”
Instead calculate cumulative coverage:



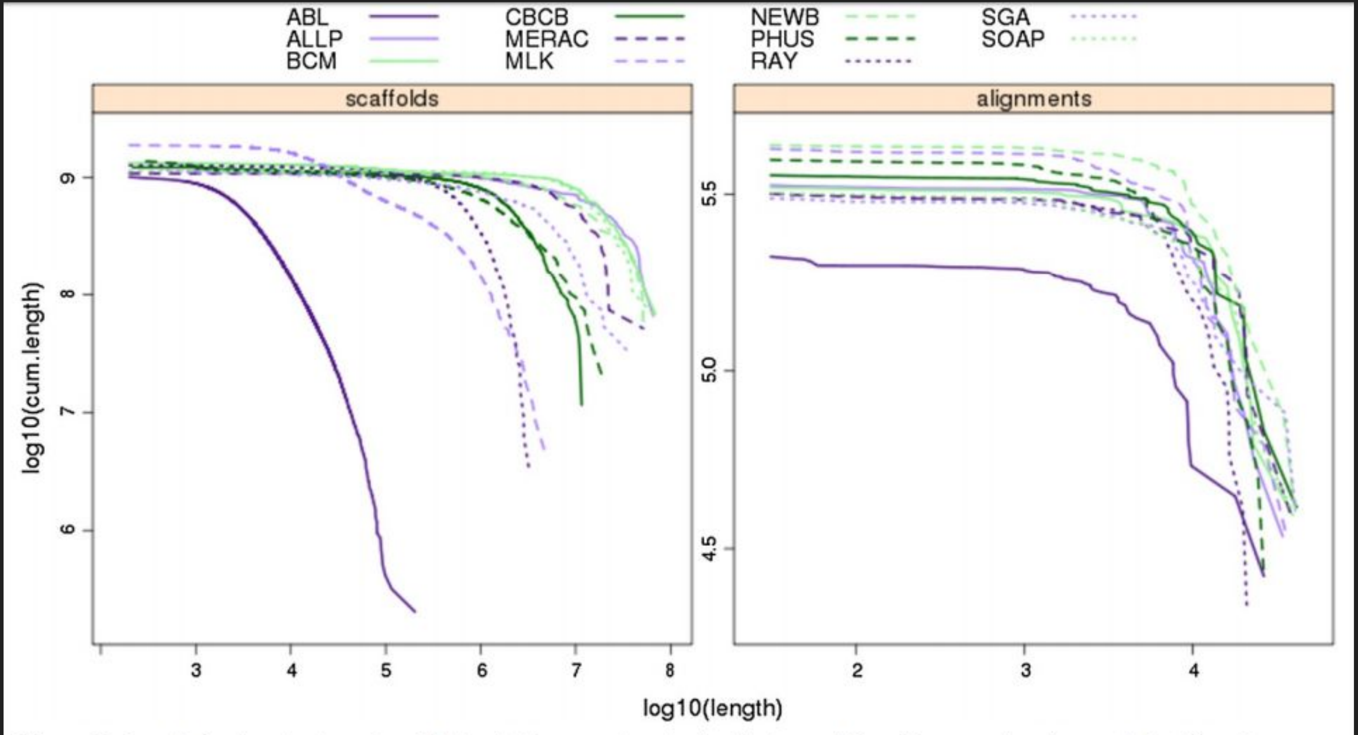
Bradnam (2013) *GigaScience* 2:10

Alignment Block NGx



Cumulative Alignment Block Length Plots

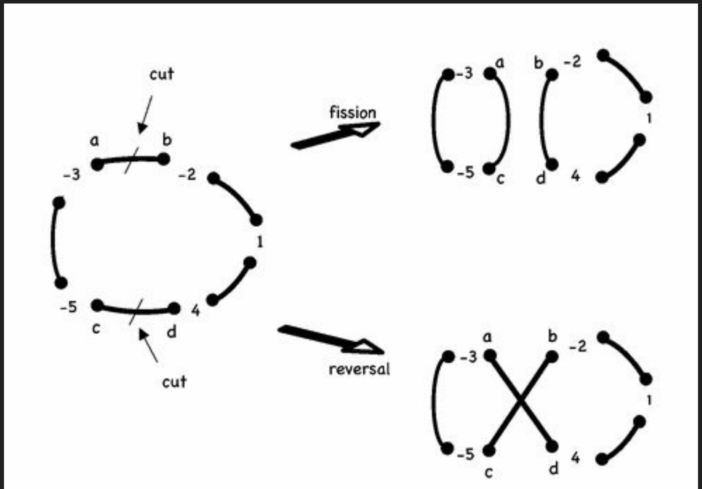
… but, measures calculated at predetermined discrete intervals are “lossy.” Instead calculate cumulative coverage:



Bradnam (2013) *GigaScience* 2:10

Mauve - Whole Genome Aligner

Calculates evolutionary distance between sequences in terms of Double-Cut and Join (DCJ) operations:



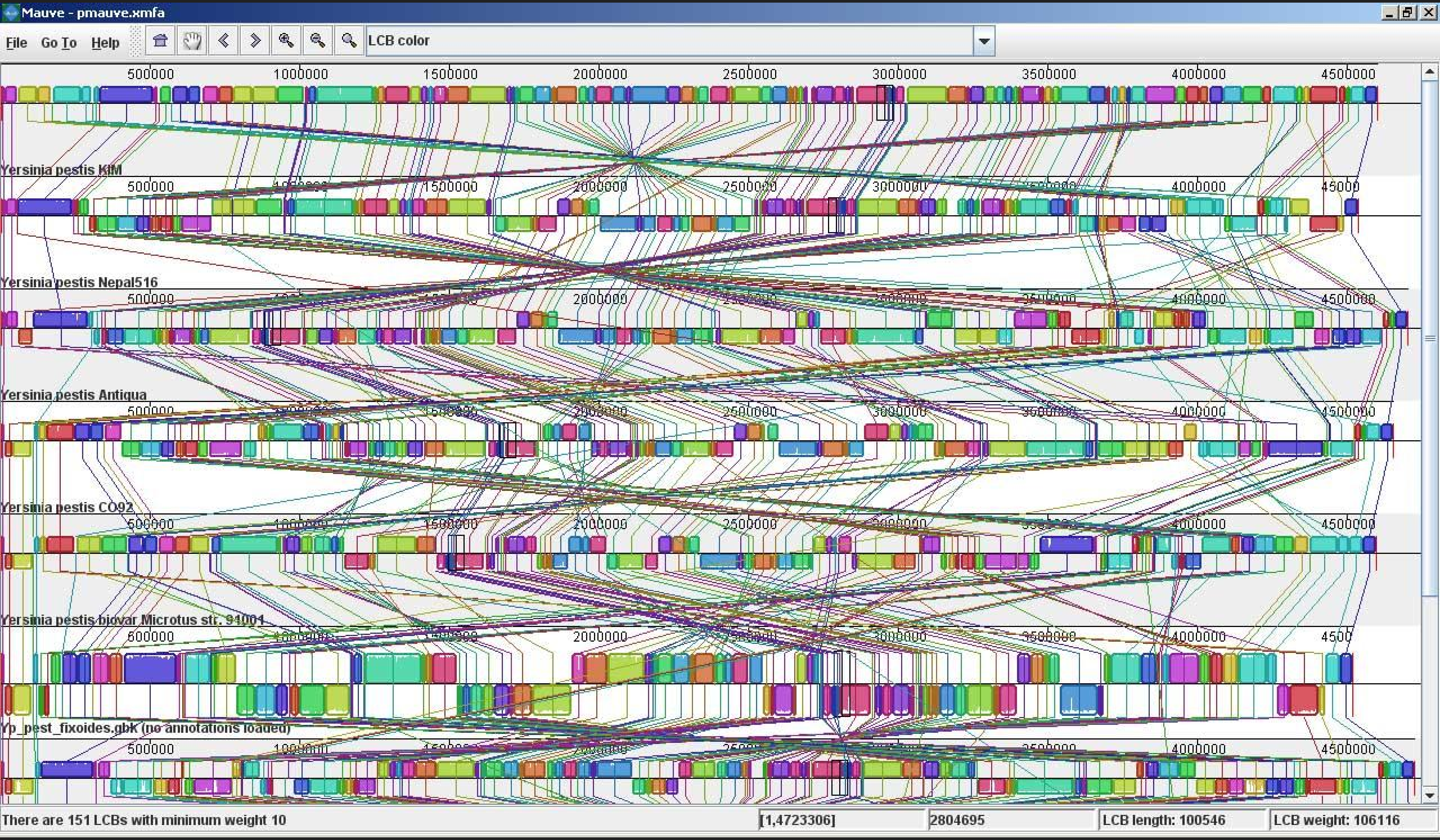
Friedberg, Darling, and Yancopoulos (2008) Bioinformatics ... Keith, ed. 452:385

Darling (2004) Genome Research 14:1394

Mauve - Whole Genome Aligner

Tries to find Locally Collinear Blocks (LCBs) that appear to be internally free of rearrangements (other than simple insertions or deletions).

Mauve - Whole Genome Aligner



There are 151 LCBs with minimum weight 10

[1,4723306]

2804695

LCB length: 100546

LCB weight: 106116

Mauve - Contig Reordering Tool

Change order and orientation of one multi-fasta file (draft genome) with respect to another multi-fasta file (“finished” genome) in order to find highest scoring Locally Collinear Blocks (LCBs)