Table S1. Regression analysis for predictors of scaffold NG50.

Predictor Variables	Model 1	Model 2
Intercept	0.822	0.850
	(0.768)	(0.745)
COVERAGE	1.63e-03	
	(4.90e-03)	
HETEROZYGOSITY	-0.420 ***	-0.413 ***
	(8.69e-02)	(8.17e-02)
REPEAT CONTENT	-0.747 **	-0.763 **
	(0.242)	(0.231)
Multiple R^2	0.76	0.76
Adjusted R^2	0.72	0.73

N=21 for all models. Standard errors in parentheses.
. $p \leq 0.10$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Scaffold NG50 = $\log_{10}$(Scaffold NG50)
COVERAGE = total sequenced bases (after decontamination) / estimated genome size
HETEROZYGOSITY = $\log_{10}$(frequency of variant branches in de Bruijn graph, k=41)
REPEAT CONTENT = $\log_{10}$(frequency of repeat branches in de Bruijn graph, k=41)

Estimated genome sizes and the frequency of variant / repeat branches were calculated by SGA Preqc (Simpson 2014).

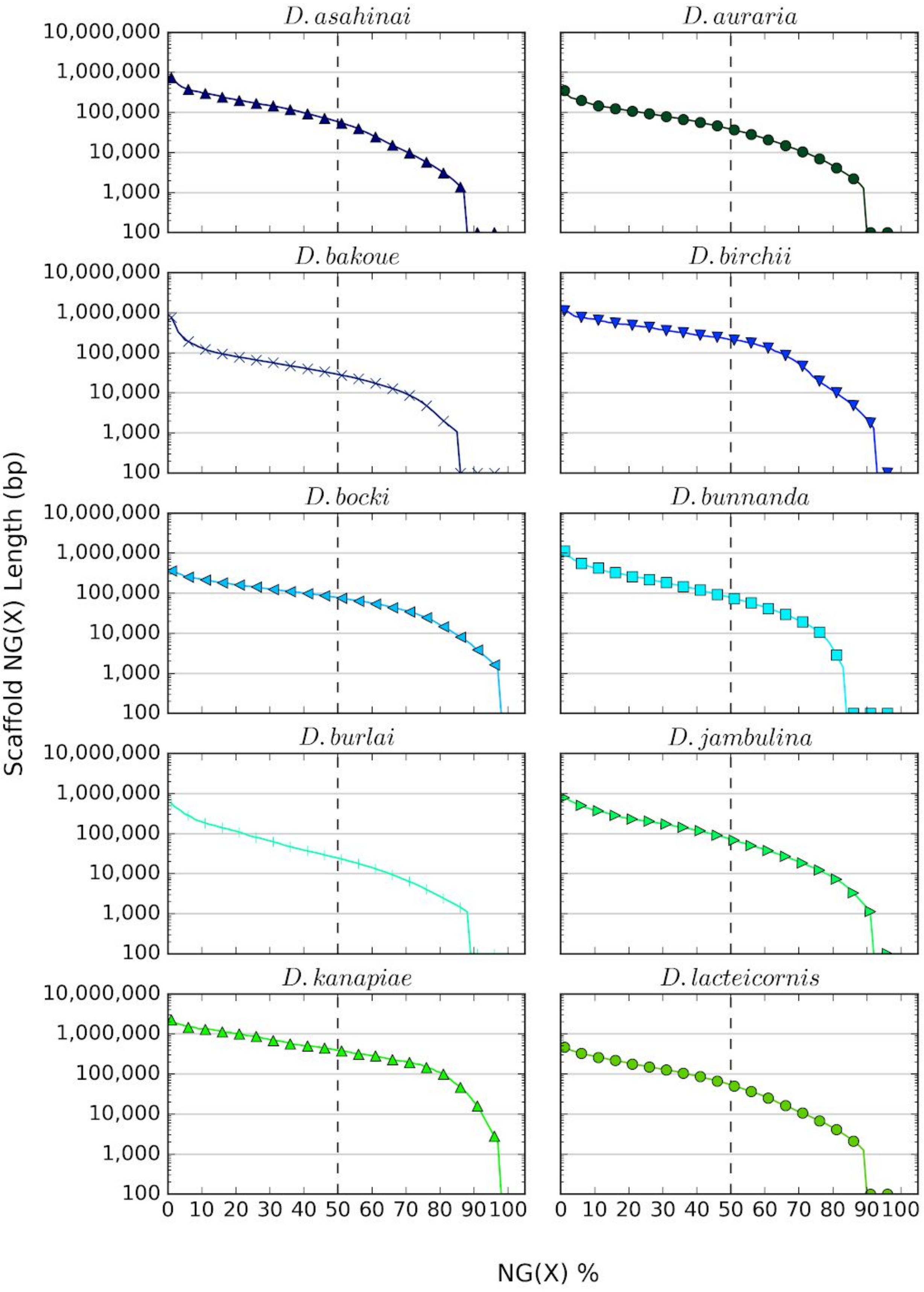
Table S2. Regression analysis for predictors of the percentage of the estimated genome size that was assembled.

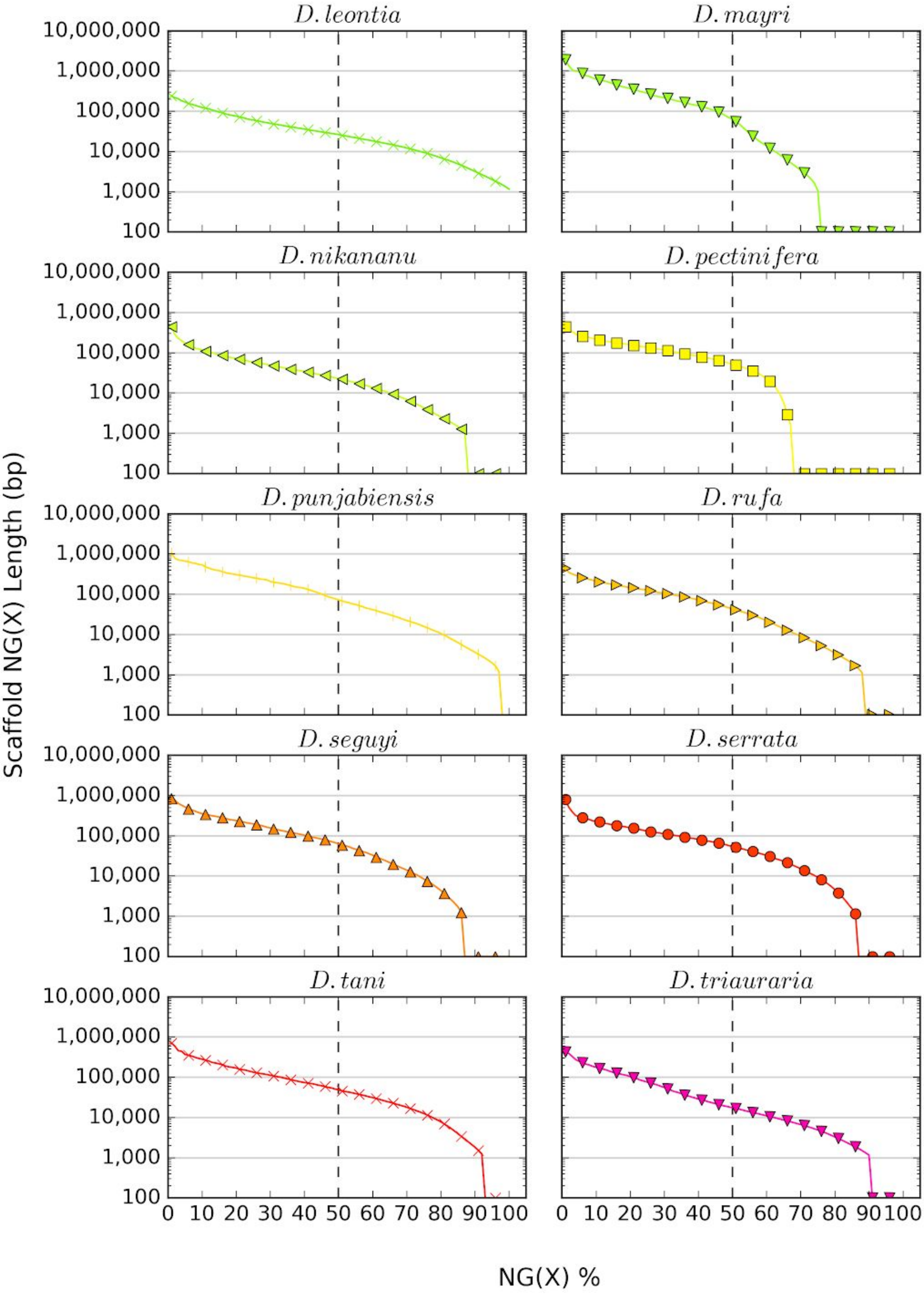
Predictor Variables	Model 1	Model 2
Intercept	5.77 (25.5)	4.65 (24.8)
COVERAGE	-6.50e-02 (0.163)	
HETEROZYGOSITY	5.43 **.** (2.89)	5.13 **.** (2.72)
REPEAT CONTENT	-30.1 ** (8.04)	-29.4 ** (7.69)
Multiple R^2	0.46	0.45
Adjusted R^2	0.36	0.39

N=21 for all models. Standard errors in parentheses.
. $p \leq 0.10$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

% of est. genome size assembled = (assembly length / estimated genome size) * 100
COVERAGE = total sequenced bases (after decontamination) / estimated genome size
HETEROZYGOSITY = $\log_{10}$(frequency of variant branches in de Bruijn graph, k=41)
REPEAT CONTENT = $\log_{10}$(frequency of repeat branches in de Bruijn graph, k=41)

Estimated genome sizes and the frequency of variant / repeat branches were calculated by SGA Preqc (Simpson 2014).


D. asahinai
D. auraria
D. bakoue
D. birchii
D. bocki
D. bunnanda
D. burlai
D. jambulina
D. kanapiae
D. lacteicornis
Scaffold NG(X) Length (bp)
10,000,000
1,000,000
100,000
10,000
1,000
100
0
10
20
30
40
50
60
70
80
90
100
NG(X) %



D. leontia
D. mayri
D. nikananu
D. pectinifera
D. punjabiensis
D. rufa
D. seguyi
D. serrata
D. tani
D. triauraria
Scaffold NG(X) Length (bp)
10,000,000
1,000,000
100,000
10,000
1,000
100
0
10
20
30
40
50
60
70
80
90
100
NG(X) %

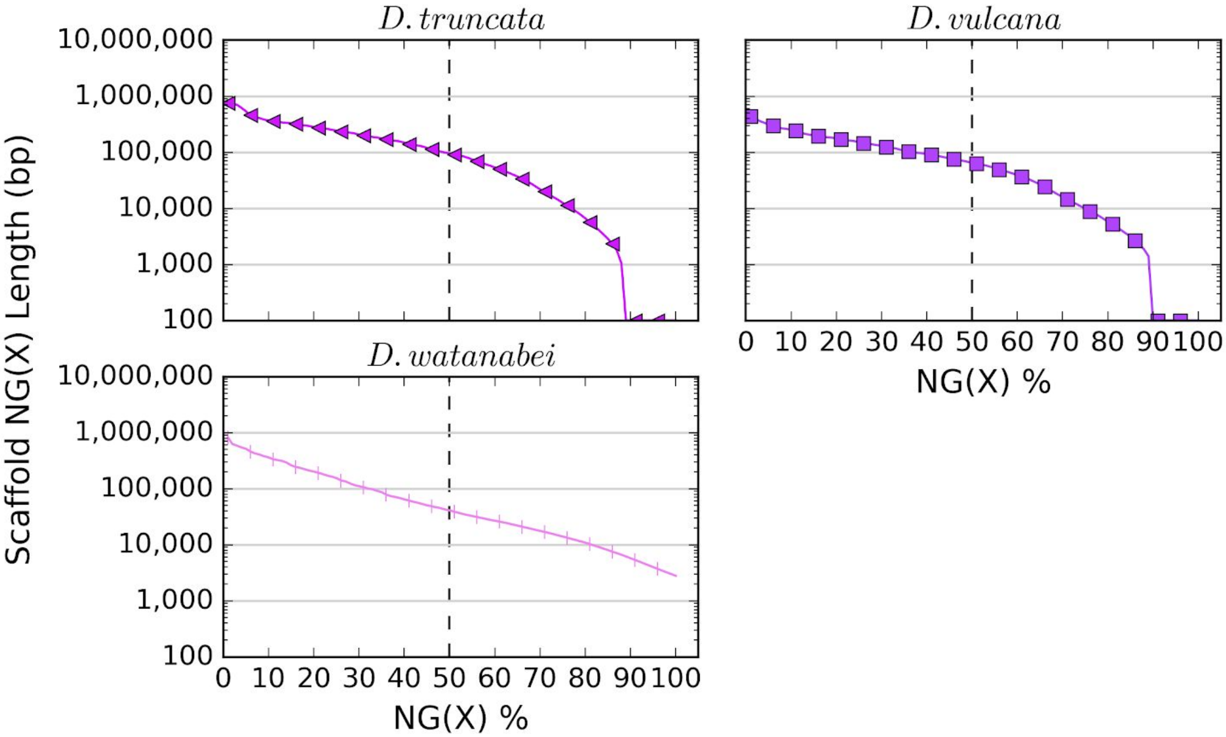


Figure S1. NG graphs showing the distribution of scaffold lengths for 23 *montium* assemblies.

To calculate the scaffold NG50 (Earl *et al.* 2011; Bradnam *et al.* 2013), scaffold lengths are ordered from longest to shortest and then summed, starting with the longest scaffold. The NG50 is the scaffold length that brings the sum above 50 % of the estimated genome size. When this calculation is repeated for all integers from 1 to 100, the result is an NG graph (Bradnam *et al.* 2013). NG graphs were constructed for each *montium* species using the corresponding genome size estimates from SGA Preqc (Simpson 2014). When a series intersects the x-axis, it means the total scaffold length is shorter than the estimated genome size. Similarly, if the series never touches the x-axis, then the assembly is longer than the estimated genome size. Due to filtering, the shortest scaffold present in any assembly is 1 kb.

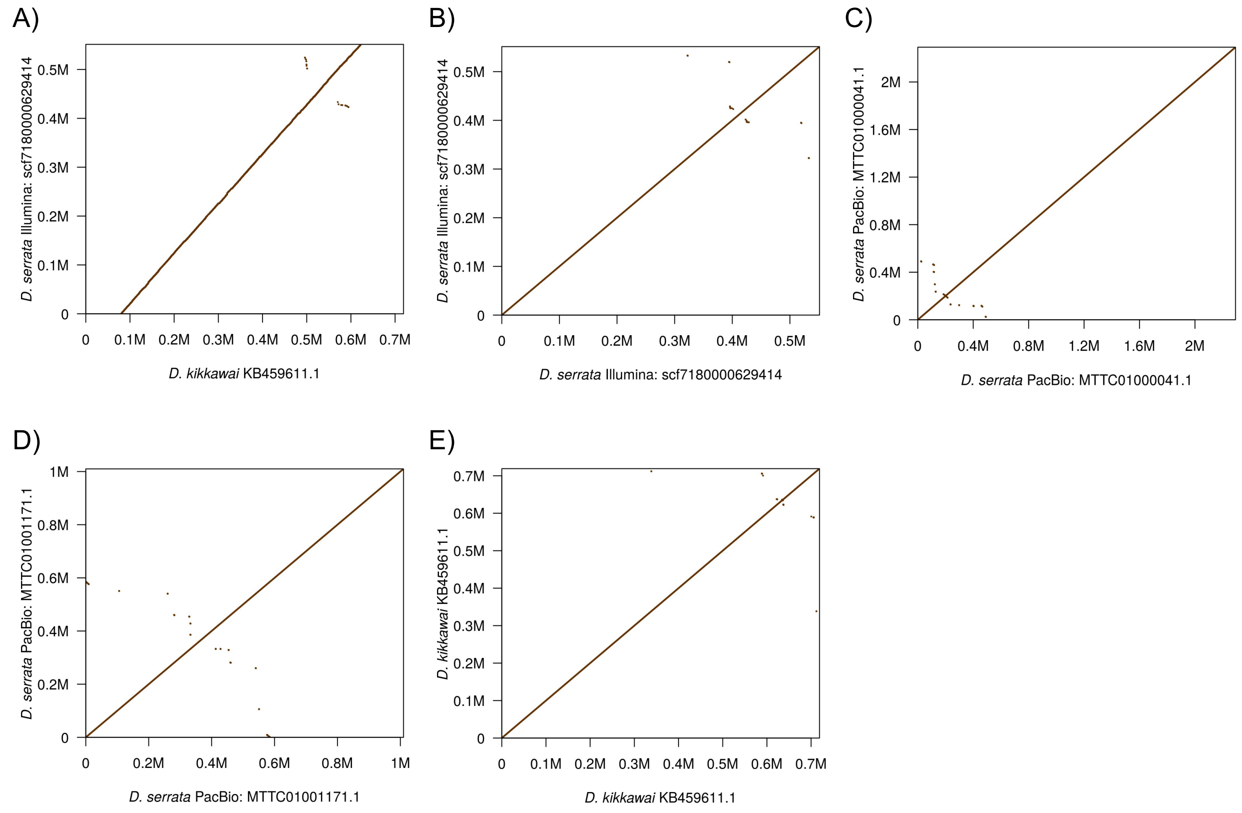


Figure S2. Additional dotplots.

A) The alignment of the fifth longest scaffold (scf7180000629414) from our Illumina *D. serrata* assembly (strain 14028-0681.02) to the orthologous scaffold from the previously published *D. kikkawai* assembly (Chen *et al.* 2014). The alignment is highly collinear, and our scaffold aligns end-to-end within the longer *D. kikkawai* scaffold. B) The alignment of scf7180000629414 to itself. C) and D) The alignment of contigs MTTC01000041.1 and MTTC01001171.1 from the previously published *D. serrata* assembly (strain Fors4) (Allen *et al.* 2017) to themselves. Portions of these contigs aligned to scf7180000629414. E) The alignment of scaffold KB459611.1 from the *D. kikkawai* assembly (Chen *et al.* 2014) to itself. This is the same *D. kikkawai* scaffold from Part A). All pairwise alignments were generated by LASTZ (Harris 2007).

References

Allen, S. L., E. K. Delaney, A. Kopp, and S. F. Chenoweth, 2017 Single-Molecule Sequencing of the Drosophila serrata Genome. G3 7: 781–788.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner *et al.*, 2013 Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2: 10.

Chen, Z. X., D. Sturgill, J. Qu, H. Jiang, S. Park *et al.*, 2014 Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. 24: 1209–1223.

Earl, D., K. Bradnam, J. St John, A. Darling, D. Lin *et al.*, 2011 Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. 21: 2224–2241.

Harris, R. S., 2007 Improved pairwise alignment of genomic DNA [Ph.D.]: The Pennsylvania State University.

Simpson, J. T., 2014 Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30: 1228–1235.